Design principles of promoter and enhancer activity in mammalian genomes

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Dominik Hartl

aus

Gallneukirchen, Österreich

Basel, 2017

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Dr. Dirk Schübeler und Prof. Dr. Patrick Tschopp

Basel den 19.09.2017

Prof. Dr. Martin Spiess (Dekan)

Acknowledgements

This thesis would not have been possible without the support and contribution of many people.

First, I would like to thank my PhD advisor, Dirk Schübeler, for the opportunity to work in his lab and for being a great mentor. I really enjoyed our discussions and I am grateful for scientific and non-scientific advice you gave me throughout the years. I also want to thank you for the trust in me to try new paths and for giving me so much freedom and flexibility.

I want to thank Leslie Hoerner and Christiane Wirbelauer for managing that everything runs smoothly in the lab. I especially want to thank Christiane for her empathy and great advices how to survive a PhD. I always enjoy talking to you and you made life in the lab much easier.

I especially would like to thank my collaborators: Arnaud Krebs for collaborations on the retina and CpG island project, mentoring, scientific discussions and a lot of fun. Tuncay Baubec, Lukas Burger, Ralph Grand and Christiane Wirbelauer for collaboration on the CpG island project. I also want to thank my collaborators from the Roska lab, Josephine Jüttner and Botond Roska. Special thanks to Josephine for her work effort that was essential for the retina project.

I would like to thank the former and present Lab members for the great scientific environment and working atmosphere. This thesis would not have been possible without lots of scientific discussions and help with methods that were new to me.

I would like to thank the members of my thesis committee, Jeffrey Chao and Sarah Teichmann, for taking the time to discuss and offer valuable input on my project and Patrick Tschopp to stand in as a Co-Referee.

For funding and providing me with the opportunity to visit courses and conferences as well as meet great people, I am grateful to the Boehringer Ingelheim Fonds.

Ich danke meiner Familie für ihre Unterstützung während meines PhDs selbst in Zeiten, in denen andere Dinge übermächtig schienen. Besonderer Dank gilt

Marietta, danke für all die Unterstützung und die wunderbare Zeit, die wir miteinander haben. Du bist die Beste.

Table of Contents

Acknowledgements i

List of abbreviations v

1 Summary 1

2 Introduction 3

2.1 Transcriptional regulation 3

2.2 *Cis*-regulatory elements 4

2.2.1 Promoters 4

2.2.2 Enhancers 5

2.3 DNA binding proteins 7

2.4 Chromatin and transcriptional regulation 8

2.4.1 Histone modifications 8

2.4.2 DNA Methylation 9

2.5 CpG islands 11

2.5.1 CpG islands are specific to vertebrates 11

2.5.2 Theories why CGIs have a higher CpG density than the rest of the genome 12

2.5.3 CpG islands overlap with promoters 13

2.5.4 CpG islands and chromatin 15

2.5.5 ZF-CxxC domain containing proteins bind unmethylated CpGs specifically 16

2.6 Transcription factors and chromatin 17

2.7 Identification of *cis*-regulatory elements 19

2.8 Quantification of *cis*-regulatory element activity 20

2.9 Cell type specificity 21

2.10 The retina, an example for high cell type specificity within one tissue 22

2.10.1 The anatomy of the retina and cell type specificity 22

2.10.1 Gene therapy 24

2.11 Scope of this thesis 26

3 Results 28

3.1 Design principles of CpG Island promoter activity 28

Prepared manuscript 28

3.2 Cis-regulatory landscape of four cell types of the retina 74

Published manuscript 74

4 General discussion and Conclusions 101

4.1 Design principles of CpG island promoter activity 101

4.2 *Cis*-regulatory landscape of four cell types of the retina 103

5 References 105

6 Curriculum vitae 119

List of abbreviations

A	adenine
AAV	Adeno-associated virus
ATAC	Assay for transposase accessible chromatin
BC	barcode, unique DNA sequence for identification
bp	basepairs
C	cytosine
CAGE	cap analysis by gene expression
CGI	CpG island
ChIP	Chromatin Immuno Precipitation
CpG	cytosine nucleotide followed by a guanine nucleotide
CRE	cis-regulatory element
DHS	Dnase I hypersensitive site
DNA	deoxyribonucleic acid
Dnmt	DNA methyltransferase
ES cell	embryonic stem cell
G	guanine
Gabpa	GA Binding Protein Transcription Factor Alpha Subunit
H3KXme3	histone 3 lysine X (e.g. 4, 27) trimethylation
HC	horizontal cells
LMR	low methylated region
NRF1	nuclear respiratory factor 1
Nrf1	Nuclear respiratory factor 1
OE	observed to expected, CpG density
PRA	parallel reporter assay
Pwp2	Periodic tryptophan protein 2 homolog
qPCR	quantitative polymerase chain reaction
RMCE	recombinase-mediated cassette exchange
RNA	ribonucleic acid
SAC	starburst amacrine cells
Snx3	Sorting Nexin 3
Sp1	Specificity protein 1
Sp3	Specificity protein 3
T	thymine
TF	transcription factor
TKO	triple knock out, knock out of Dnmt1, Dnmt3a and Dnmt3b
TSS	transcriptional start site
UMR	unmethylated region
WT	wildtype

1 Summary

Correct gene expression patterns are central for cellular function and the development of organisms. This is controlled by regulatory elements such as enhancers and promoters. In this thesis, I present work from two projects with the goal to identify design principles of promoter and enhancer activity in mammalian genomes.

In the first part of the thesis, I focused on CpG island promoters. This promoter type represents the majority of mammalian promoters and is characterised by a high density of the CpG dinucleotide. However, to what extent and how this characteristic dinucleotide contributes to promoter activity is still unclear and is one central question of this project. By monitoring binding of transcription factors (TFs) assumed to play a role in CpG island activity and quantifying the activity of promoter mutants and artificial promoters, we gained insight into the role of CpGs in transcriptional activity. The generated data suggests that high CpG density is not sufficient for transcriptional activity, yet necessary when combined with more complex TF binding motifs. We could further show that DNA methylation decreases activity of promoter mutants with low CpG density. Our experiments led us to hypothesise that high CpG density is required to generate a chromatin environment permissive for transcriptional activity.

In the second part of the thesis, I focused on cell type and tissue specific regulatory elements. To illustrate an experimental workflow to identify and test regulatory elements for transcriptional activity in specific cell types, we used the mouse retina, a very specialised tissue comprised of ~50 cell types. To identify regulatory elements, we combined transcriptome and epigenome profiling to map the regulatory landscape of four distinct cell types isolated from mouse retinas (rods, cones, horizontal and starburst amacrine cells). This data also revealed sequence determinants and candidate TFs that control cellular specialisation. We tested previously identified regulatory regions using a parallelised reporter assay for their ability to autonomously control transcriptional activity in the four cell types. We were able to generate a catalogue of *cis*-regulatory regions active in retinal cell types and further demonstrate their utility as a potential resource for cellular tagging and manipulation.

Taken together, the work presented here advances our knowledge about location and regulation of regulatory regions that function in specialised cell types and also provides insight into the regulation of CpG island promoters that tend to be ubiquitously expressed.

2 Introduction

2.1 Transcriptional regulation

The blueprint of organisms is encoded within long deoxyribonucleic acid (DNA) molecules comprised of only four different subunits. These subunits, termed nucleotides, each consist of deoxyribose, a phosphate group, and one of the four bases; cytosine (C), guanine (G), adenine (A), and thymine (T). The shape and function of all cells in an organism are encoded in stretches of these four nucleotides termed regulatory regions and genes.

The human haploid genome is estimated to be more than 3 gigabases in size (Venter *et al*, 2001) while the genome of the bacterium *Escherichia coli* is around 4.6 megabases (Blattner, 1997), a difference of nearly 700-fold. Despite this large difference in genome size, humans only have about seven times the number of genes of *Escherichia coli* (Venter *et al*, 2001; Blattner, 1997). This indicates that the complexity of an organism is not simply determined by the number of genes. Additionally, gene size does not increase by more than 100-fold from *E. coli* to human, suggesting that the human genome contains more non-protein coding bases (Venter *et al*, 2001; Blattner, 1997). Indeed, while ~88% of the *E. coli* genome is coding for proteins (Blattner, 1997) only ~3% of the human genome codes for proteins (ENCODE, 2012). The vast majority of the human genome represents non-protein coding regulatory regions and relicts of evolution (Palazzo & Gregory, 2014; ENCODE, 2012). Such a genome composition in multicellular eukaryotes, like human or mouse, requires additional layers of regulation compared to unicellular eukaryotes or prokaryotes. The different shapes and functions of cells in complex organisms require correct expression of genes and tuning of gene expression levels according to specific requirements. This is controlled by transcriptional regulatory regions in the DNA sequence. These regions can be located at the start of a gene or more distally (Maston *et al*, 2006). The sequence of these regulatory regions is interpreted by DNA binding proteins and the temporal integration of regulatory events can be performed by chromatin structure.

2.2 *Cis*-regulatory elements

Gene expression has to be controlled on several levels with the primary layer of regulation being the DNA sequence itself, due to the fact that it does not only encode gene products but also determines their expression patterns in the whole organism. *Cis*-regulatory elements are sequence stretches in the genome with the ability to control spatiotemporal gene expression levels (Maston *et al*, 2006). These elements can be broadly divided into those that lie proximal to the transcriptional start sites (TSS) of genes, called promoters (Grosschedl & Birnstiel, 1980a, 1980b), or more distal from the TSS, called enhancers (Banerji *et al*, 1981, 1983; Müller & Schaffner, 1990).

2.2.1 Promoters

Promoters are *cis*-regulatory elements located directly at and around the TSS. Besides enabling the initiation of transcription by RNA polymerase II, promoters also regulate gene expression patterns. The first discovered eukaryotic promoter was the one controlling the histone H2A gene in *Xenopus* oocytes nearly 40 years ago (Grosschedl & Birnstiel, 1980a, 1980b). Since then, most of the work focused on identifying sequence elements within the region surrounding the TSS, called the core promoter. This region allows RNA polymerase II to initiate transcription and extends about 40 base pairs (bp) upstream of the TSS. It consists of different core promoter elements that are bound by TFs that establish a pre-initiation complex (Haberle & Lenhard, 2016). The pre-initiation complex positions RNA polymerase II and denatures DNA in order for transcription to start. The most frequently occurring core promoter elements are the TATA-box and the Initiator element. The TATA-box lies approximately 30 bp upstream of the TSS and is bound by TFIID, which recruits the pre-initiation complex. The Initiator element overlaps with the TSS and directs transcriptional initiation (Figure 1). However, there are no universal promoter elements and although many promoters lack these elements, RNA polymerase II is still able to productively initiate transcription (Haberle & Lenhard, 2016).

Besides the core promoter, other sequence features control promoter activity. These are sequences that can be bound by DNA binding proteins such as TFs that interpret the regulatory sequence and are able to directly or indirectly control activity of the promoter. DNA binding proteins recruit co-activators and co-repressors, the sum of these regulatory inputs then results in a controlled transcriptional output that forms the basis of cellular function (Figure 1).

The first attempts to understand the logic of sequence elements controlling transcriptional activity in mammals started already more than 30 years ago (Myers *et al*, 1986). Despite this, we are still unable to predict transcriptional activity just based on the DNA sequence. But there are a number of sequence features that are predictive of regulatory function.

One feature that aids in the prediction of the regulatory activity of a DNA sequence in vertebrates is the density of the dinucleotide CpG (Ioshikhes & Zhang, 2000). CpG rich sequences tend to overlap with promoters predominantly controlling genes broadly expressed across different cell types and tissues, so called housekeeping genes. However, not all promoters are CpG rich. The density of CpGs within all promoters is distributed in a bimodal fashion, separating them in CpG poor and CpG rich promoters (Mohn & Schübeler, 2009) (also see chapter 2.5 CpG islands). Part of this thesis will focus on CpG rich promoters and the role of the dinucleotide CpG in transcriptional activity of this promoter type (see 3.1 Design principles of CpG island promoter activity).

2.2.2 Enhancers

Enhancers are distal regulatory elements that can enhance promoter activity independent of their distance (Müller & Schaffner, 1990). Together with promoters, they control transcriptional activity. Reports suggested that some transcription also takes place at enhancers (enhancer RNA). However, these transcripts are only lowly abundant and unstable (Kim *et al*, 2010; Wang *et al*, 2011; Hah *et al*, 2013).

The first identified enhancer was a 72 bp long sequence that originated from the SV40 virus genome. Placing this enhancer in front of a rabbit hemoglobin β1 gene on a plasmid resulted in 200 times higher activity in HeLa cells than without this sequence (Banerji *et al*, 1981). Two years later the same lab also identified the first animal enhancer (Banerji *et al*, 1983).

Similar to promoters, enhancers are bound by DNA binding proteins such as TFs. These proteins can recruit additional factors with activating or repressive function and together, the complex of enhancer and proteins interacts with the promoter to control gene expression levels (Figure 1). The combined regulatory inputs of the promoter and enhancer(s) tune gene expression depending on cellular function (Shlyueva *et al*, 2014). Looping allows for interaction between enhancers and promoters even when they are several kilo- to megabases away from each other (Amano *et al*, 2009).

Enhancer function is especially central for the control of cell-type specific expression, but also for genes active across several cell-types. The modularity of enhancers allows the cell to utilise different enhancers in distinct cell types for the same promoter or several enhancers can act in concert to establish required expression levels of a gene in a specific cell type (Xu & Smale, 2012; Smith & Shilatifard, 2014; Smallwood & Ren, 2013; Calo & Wysocka, 2013; Arnone & Davidson, 1997).

Part of this thesis will focus on identification of active enhancers in different cell types and functional testing of their autonomous activity (See 3.2 Cis regulatory landscape of the retina).

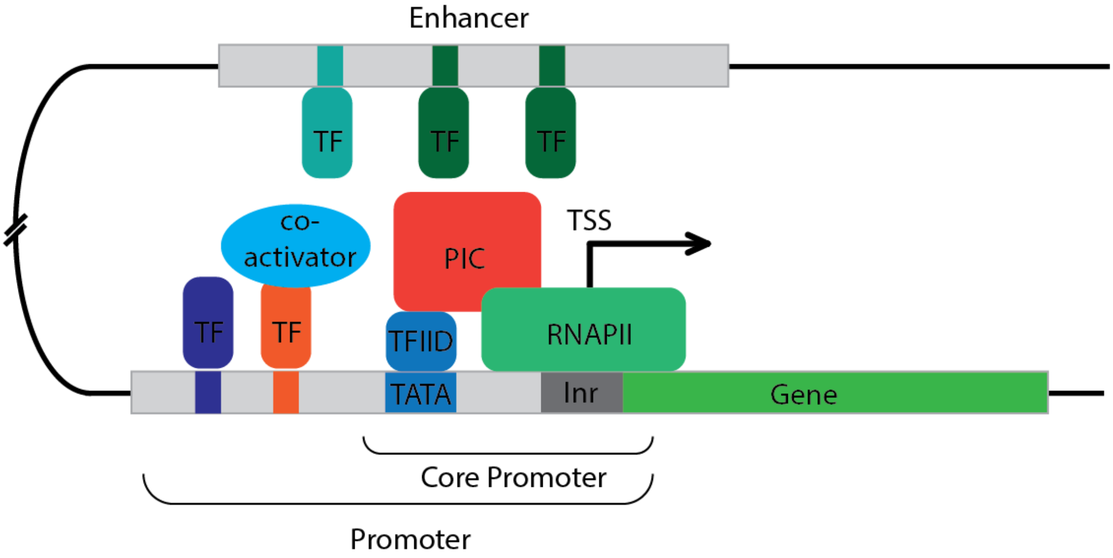


Figure 1: Promoters and enhancers control gene expression

Enhancers and promoters are bound by TFs that together regulate gene activity. The promoter contains the core promoter including TATA-box (TATA) and the initiator element (Inr). TFIID binds to the TATA-box and enables recruitment of the *pre-initiation complex (PIC) that positions RNA polymerase II (RNAPII) and denatures DNA in order for transcription to start. Transcription factors (TF) can recruit co-activators or co-repressors.*

2.3 DNA binding proteins

DNA sequence is interpreted by proteins in order to establish correct transcriptional levels. The first protein interacting with DNA and controlling gene expression was identified in prokaryotes. The authors named this protein 'regulator', while now they are generally called transcription factors (Jacob & Monod, 1961). Extensive research on these types of proteins uncovered many more TFs.

TFs can directly lead to increased transcriptional activity by interaction with the transcriptional machinery by promoting initiation, elongation or re-initiation of transcription (Maston *et al*, 2006). For example, the TF Sp1 has two transactivation domains (Courey & Tjian, 1988; Oka *et al*, 2004) that can directly interact with TBP (TATA-Binding Protein) (Emili *et al*, 1994) and TAF4 (TATA-Box Binding Protein Associated Factor 4) of the transcriptional initiation machinery (Gill *et al*, 1994) and thereby promote initiation. Alternatively, transcription can be indirectly influenced by the recruitment of cofactors that, for example, allow binding of other TFs (de la Serna *et al*, 2005) (See also chapter 2.6 Transcription factors and chromatin).

TFs typically recognize 6-8bp long stretches of DNA called TF motifs (Kadonaga, 2004). Such motifs occur every 4,000-70,000bp in the genome by chance. However, only a small proportion of occurrences of a motif sequence in the genome are bound with the majority remaining unbound. This opens the interesting question on what determines TF binding besides motif sequence. One possible explanation is that TF binding requires direct or indirect interactions between TFs. This suggests that if motifs of a number of factors co-occur, they are more efficiently bound. Often this requires that the TF motifs are placed at a certain distance from each other to sterically allow interaction of the factors (Reiter *et al*, 2017). In line with this, many TFs form homo- or heterodimers, and therefore, their motifs are often comprised of TF motif pairs (Jolma *et al*, 2013, 2015) (See also chapter 2.6 Transcription factors and chromatin).

2.4 Chromatin and transcriptional regulation

In eukaryotic cells, DNA is associated with proteins, forming a complex called chromatin. The most abundant proteins in the nucleus are histones that together with DNA form the nucleosome. A stretch of 147bp of DNA is wrapped around the histone octamer consisting of two copies each of histones H2A, H2B, H3 and H4 (Kornberg & Thomas, 1974; Richmond & Davey, 2003). Initially, it was thought that nucleosomes only play a role in chromosome compaction to fit eukaryotic genomes into the nucleus. Now we know that different chromatin states also serve a regulatory function with active regulatory regions residing in open chromatin (euchromatin) and inactive regions located within tighter packed closed chromatin (heterochromatin) (Voss & Hager, 2013). It was shown that active promoters are located within euchromatin and have low nucleosome occupancy and a nucleosome free region at the TSS (Schones *et al*, 2008). Such differences in chromatin structure can affect TF binding to DNA, this is discussed in chapter 2.6.

2.4.1 Histone modifications

In addition to differences in the positioning and occupancy of nucleosome across the genome, their histone components can be posttranslationally modified. Such histone modifications occur mainly on the N-terminal tails of histone H3 and H4, and correlate with active or inactive regulatory states (Kouzarides, 2007). For example, inactive heterochromatin is marked by the methylation of lysine 9 or lysine 27 on histone H3 (H3K9me3 and H3K27me3) and H2AK119 ubiquitylation. By contrast, active euchromatin is marked by acetylated lysines on histones H3 and H4 (Kouzarides, 2007).

Within euchromatin, different *cis*-regulatory elements contain characteristic histone modifications. Active promoters are marked by H3K4me3 and H3K9ac around the TSS and H3K36me3 in the gene body, while active enhancers are modified by H3K4me1 and H3K27ac (Barth & Imhof, 2010; Shlyueva *et al*, 2014).

Posttranslational histone modifications probably do not affect transcription directly but via intermediate steps since they can be bound and interpreted by proteins. For example, H3K9me3 has been shown to be bound by heterochromatin protein 1 (HP1), which mediates transcriptional repression (Loyola *et al*, 2001). Many other proteins have been identified that are able to recognise different histone modifications and thereby potentially influence transcriptional regulation (Yun *et al*, 2011). However, for histone modifications in both eu- and heterochromatin it is still unclear how gene activation or silencing is established and if histone marks are a cause or consequence of changes in TF binding and transcription.

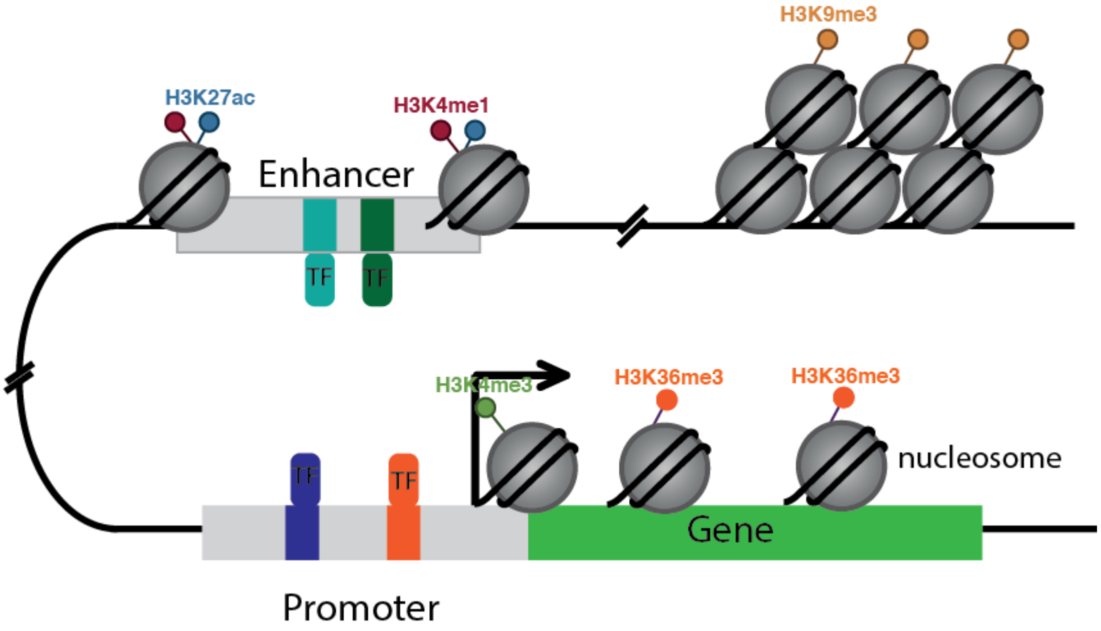


Figure 2: Histone modifications at active enhancers and promoters and outside of regulatory regions. *Inactive chromatin has high nucleosome occupancy and is marked by H3K9me3. Active enhancers are marked by H3K4me1 and H3K27ac. TSS of promoters is marked by H3K4me3 while gene bodies display H3K36me2/3.*

2.4.2 DNA Methylation

In addition to nucleosome occupancy and posttranslational histone modifications, DNA itself can be modified and thereby contribute to chromatin structure and regulation. Most of the Cs in the dinucleotide CpG are methylated at the fifth position of the pyrimidine ring in vertebrate genomes (Lister *et al*, 2009). In vertebrates, this CpG methylation is catalysed by Dnmt3a and Dnmt3b, and maintained through cell division by Dnmt1 (Hermann *et al*, 2004). Removal of DNA methylation can be brought about either passively by cell division (Chen *et al*, 2003) or enzymatically by members of the TET family (Tahiliani *et al*, 2009).

DNA methylation is linked to gene repression (Cedar, 1988). Two possible mechanisms have been suggested how DNA methylation can lead to repression: (1) by preventing the binding of TFs that require CpGs in their motif to be unmethylated (Watt & Molloy, 1988; Iguchiariga & Schaffner, 1989; Prendergast & Ziff, 1991; Campanero *et al*, 2000; Domcke *et al*, 2015), or (2) through attracting proteins that bind methylated CpGs specifically (Meehan *et al*, 1989; Hendrich & Bird, 1998) and consequently block binding of other factors or recruit repressors.

A group of proteins that specifically bind methylated CpGs are the MBD (Methyl-CpG-binding domain) proteins (Ohki *et al*, 2001; Ho *et al*, 2008; Baubec *et al*, 2013; Hendrich & Bird, 1998). These proteins are able to recruit cofactors that mediate chromatin repression (Meehan *et al*, 1989; Hendrich & Bird, 1998; Hendrich & Tweedie, 2003). For example, Mbd2 is part of the Mi2/NuRD histone deacetylase repressor complex (Zhang *et al*, 1999) or MeCp2 that also associates with histone deacetylase complexes (Nan *et al*, 1997).

Although the majority of CpGs within vertebrate genomes are methylated, there are specific regions that have decreased or no DNA methylation. For example, active enhancers and promoters tend to have low methylation levels (Stadler *et al*, 2011). By contrast, the methylation state of inactive promoters depends on sequence composition (Schübeler, 2015). Promoters with a high density of CpGs are unmethylated even if they are inactive, these promoters are called CpG islands (CGIs) (Bird *et al*, 1985) (Figure 3).

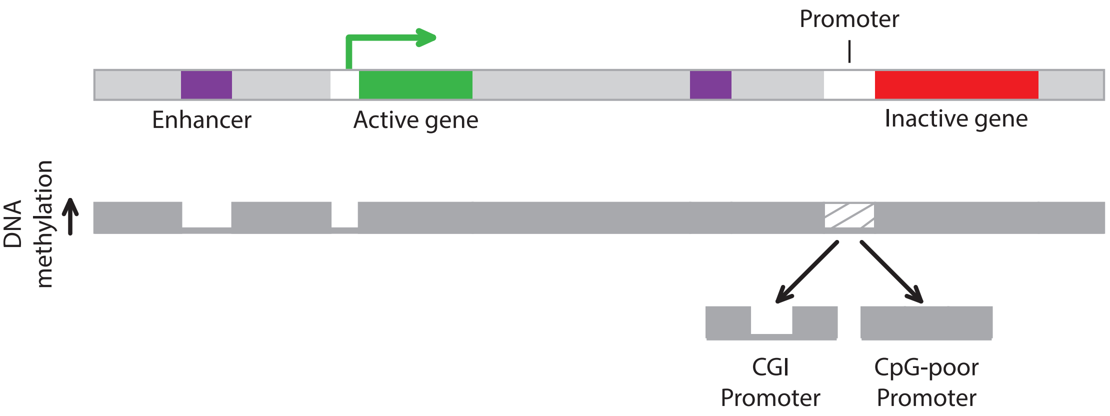


Figure 3: Distribution of DNA methylation in vertebrate genomes.

The majority of CpGs in vertebrate genomes are methylated, while active regulatory regions and inactive CpG rich promoters are only lowly or unmethylated. (adapted from (Schübeler, 2015)*)*

2.5 CpG islands

While the majority of CpGs in mammalian genomes are methylated, unmethylated CpGs are concentrated in specific regions called CGIs (Bird *et al*, 1985). CGIs overlap with ~60% of human and mouse promoters, resulting in a bimodal distribution of CpG density in promoters (Mohn & Schübeler, 2009).

Within vertebrates, CGIs have been defined as at least 200bp long regions with a G+C content of at least 50% and an observed-to-expected ratio (OE) of at least 0.6 where OE is the number of CpGs / (number of Cs x number of Gs) x length of the region in nucleotides (Gardiner-Garden & Frommer, 1987). However, if and how the higher CpG density at CGIs compared to the rest of the genome contributes to transcriptional activity of CGI promoters has not yet been comprehensively assessed and is a central question of this thesis.

2.5.1 CpG islands are specific to vertebrates

CGIs have been mainly studied in mammals but are present in all vertebrates that have extensive CpG methylation (Han *et al*, 2008). The branching of invertebrates and vertebrates coincides with the appearance of DNA methylation. *Ciona intestinalis*, an organism that is close to the invertebrate-vertebrate boundary, exhibits a mosaically methylated genome. Genes that are located within methylated domains of the *Ciona* genome have been shown to be sometimes associated with short CGI-like, unmethylated regions at the TSS (Suzuki *et al*, 2007).

2.5.2 Theories why CGIs have a higher CpG density than the rest of the genome

Genomes of organisms with DNA methylation in the germ line are generally depleted in CpGs (Bird, 1980; Jones, 2012). This phenomenon is thought to be caused by different mutation rates of methylated versus unmethylated CpGs. C to T conversion accounts for most of the spontaneous mutations within DNA (Shen *et al*, 1994). Unmethylated Cs can deaminate to Uracil, which is an improper base in DNA, it is efficiently recognized by the DNA mismatch repair machinery and replaced by a C (Barnes & Lindahl, 2004). By contrast, methylated Cs within the CpGs are deaminated to Ts, which is a proper base that is incorrectly paired with G after the mutation event. Although this mismatch is thought to be repaired by glycosylases capable of replacing the T, such as MBD4 and TDG, this is still less efficient than repair of Uracils (Millar, 2002; Hendrich *et al*, 1999; Neddermann & Jiricny, 1993). Therefore, following a round of replication, a methylated C is more likely to mutate to a T than an unmethylated C and this results in the depletion of CpGs throughout the genome. CpG islands are regions that are unmethylated in the germ line, which accounts for their decreased loss of CpGs during evolution (Bird, 1980; Cohen *et al*, 2011). This theory suggests that CpGs can be seen as a footprint of evolution due to different mutation rates of chemically distinct forms of CpGs without precluding any functional role. Alternatively, CpG density could play a functional role, leading to selective pressure that could contribute to CpG maintenance throughout evolution (Bird, 2011; Deaton & Bird, 2011). In fact, CpG density is so far the best predictor of promoter activity from the DNA sequence alone (Ioshikhes & Zhang, 2000). A functional role of CpGs for transcriptional activity is not mutually exclusive with a model where CpGs are a footprint of evolution. There would be no purifying selection on CpG density if the density of CpGs for functionality is lower than the equilibrium of deamination rates versus CpG gain by spontaneous mutations.

Taken together CpG islands are likely a product of different mutation rates of methylated and unmethylated CpGs, however this does not preclude a functional role for CpGs.

2.5.3 CpG islands overlap with promoters

About 60% of mouse and human promoters in the genome are CpG islands, leading to a bimodal distribution of promoter CpG densities in the genome, those that are CpG rich and those that are CpG poor (Figure 4) (Mohn & Schübeler, 2009). While CpG-poor promoters are generally associated with narrow expression patterns, CGI promoters are thought to be active across many cell types controlling ubiquitously expressed housekeeping genes (Larsen *et al*, 1992; Butler & Kadonaga, 2002). However, ~30% of CGI genes are tissue specific and include important developmental regulators such as the hox genes (Mohn & Schübeler, 2009).

Several lines of evidence suggest that CGI promoters are regulated differently from CpG poor promoters: (1) At the DNA level, the skewed representation of CpGs, Gs and Cs in CGIs implies that a different set of factors is involved in interpreting these sequences. (2) CGI promoters can initiate transcription across a rather broad region, while CpG poor promoters typically have very precise TSSs, as evident from CAGE (cap analysis by gene expression) datasets (Carninci *et al*, 2006). This is thought to be, in part, attributed to the fact that CGIs generally lack a TATA-box that enables focused initiation (Sandelin *et al*, 2007). To date, most of the biochemical work on transcriptional initiation has focused on CpG poor promoters, simply because of experimental convenience and the fact that the first model promoters were of viral origin and CpG poor (Zhu *et al*, 2008; Saxonov *et al*, 2006; Antequera & Bird, 1993; Ioshikhes & Zhang, 2000; Bajic *et al*, 2006; Shen *et al*, 2012; Butler & Kadonaga, 2002; Benoist & Chambon, 1981). As a consequence, we lack knowledge about functional promoter elements controlling transcriptional activity within CGI promoters.

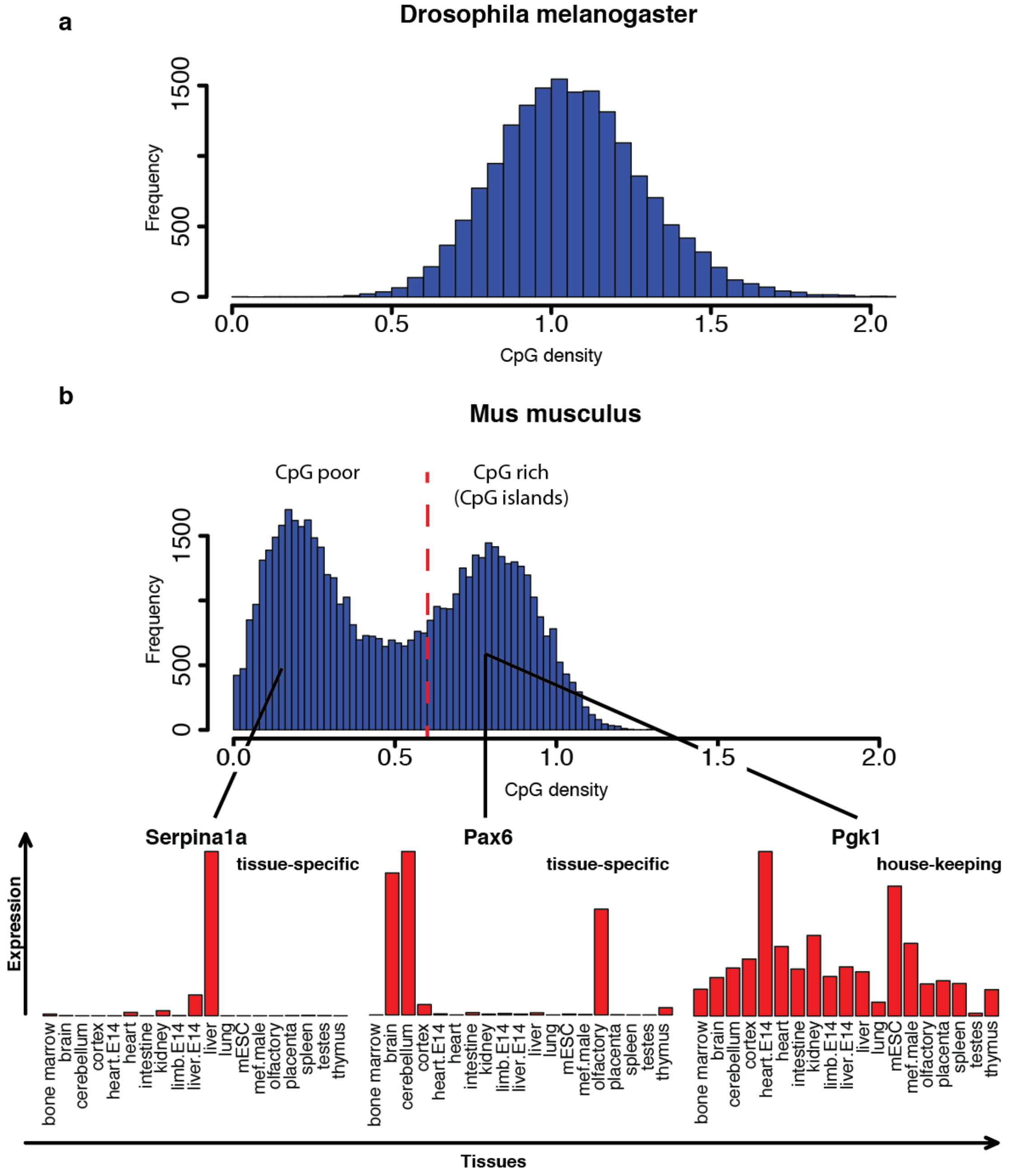


Figure 4: The CpG density of vertebrate promoters has a bimodal distribution.

Histogram of CpG densities of all promoters in the Drosophila (a) and mouse genome (b).

(a) CpG densities of promoters in the invertebrate genome of Drosophila melanogaster is unimodally distributed.

(b) CpG densities of promoters in the vertebrate genome of Mus musculus is bimodally distributed. Barplots below show expression levels for three genes with different CpG densities across 19 different tissues in mouse.

2.5.4 CpG islands and chromatin

Besides their unique sequence composition, CpG islands also have a characteristic chromatin structure, with the distinctive hallmark being that they are mostly unmethylated in normal cell types. It has been shown that high CpG density alone is sufficient for a sequence to remain unmethylated (Lienert *et al*, 2011; Krebs *et al*, 2014). Additionally binding of specific TFs has an even higher potential to prevent DNA sequence from getting methylated especially in combination with high CpG density (Lienert *et al*, 2011; Krebs *et al*, 2014). This is in agreement with a putative role of the CGI binder Sp1 in keeping CGIs unmethylated (Brandeis *et al*, 1994; Macleod *et al*, 1994).

CpG islands are marked by the histone modification H3K4me3 independent of transcriptional activity (Weber *et al*, 2007; Guenther *et al*, 2007; Mikkelsen *et al*, 2007). H3K4me3 can interact with the NuRF chromatin remodeling complex suggesting a role in keeping the promoter accessible (Li *et al*, 2006; Wysocka *et al*, 2006). In agreement with that, CGIs have been shown to be depleted of nucleosomes independent of transcriptional activity (Fenouil *et al*, 2012) (Figure 5).

CpG island promoters show reduced levels of H3K36me2 compared to CpG poor promoters. Removal of this modification is mediated by the histone demethylase activity of the ZF-CxxC domain containing protein KDM2A (Tsukada *et al*, 2006). Inactive CpG-island promoters are marked by H3K27me3 (Mohn & Schübeler, 2009) (Figure 5).

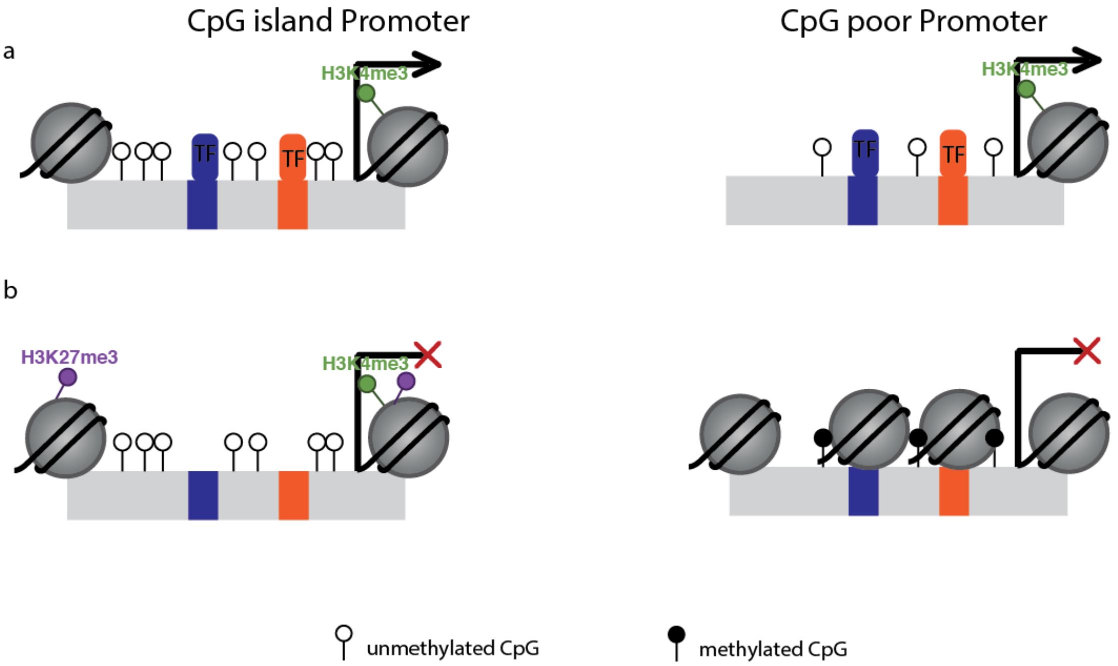


Figure 5: Chromatin at CGI promoters versus CpG poor promoters

Chromatin at (a) active promoters and (b) inactive promoters. In contrast to CpG poor promoters, CGI promoters stay open and DNA methylation free independent of transcriptional activity.

2.5.5 ZF-CxxC domain containing proteins bind unmethylated CpGs specifically

In addition to classical TFs that bind 6bp or longer motifs, there are also DNA binding proteins that only bind very short motifs such as MBD proteins, which bind methylated CpGs (discussed in 2.4.2), or ZF-CxxC domain containing proteins that specifically bind unmethylated CpGs (Voo *et al*, 2000; Long *et al*, 2013).

The bipartite modification and distribution pattern of CpGs leads to a ~50-fold higher concentration of unmethylated CpGs at CpG islands than elsewhere (Bird, 2011). The shortness of the motif and the strong asymmetry in the density of unmethylated CpGs suggest that ZF-CxxC domain containing proteins function in a concentration dependent manner allowing them to have high specificity for CpG islands despite the simplicity of the motif compared to classic TFs (Bird, 2011).

ZF-CxxC domain containing proteins have been shown to alter the chromatin environment at their binding site. For example, KDM2A and KDM2B have been described as H3K36 demethylases with preference for H3K36me2. They are thought to prevent H3K36me2 spreading into the promoter, which would interfere with transcriptional initiation (Tsukada *et al*, 2006; Blackledge *et al*, 2010). Another ZF-CxxC domain containing protein linked to an active chromatin state is the H3K4 methylase Cfp1. H3K4me3 occurs around the TSS of promoters and can be bound by the chromatin remodeler Chd1 (Clouaire *et al*, 2012; Flanagan *et al*, 2005). Additionally, ZF-CxxC domain containing proteins were suggested to play a role in protecting CGIs from methylation (Thomson *et al*, 2010; Long *et al*, 2013; Boulard *et al*, 2015).

Not all proteins containing ZF-CxxC domains are linked to active chromatin. The DNA methylation maintenance enzyme, Dnmt1, also has a ZF-CxxC domain. Structural studies showed that in this case the ZF-CxxC domain ensures that unmethylated CpGs stay unmethylated through the cell cycle. If Dnmt1 binds to unmethylated CpGs, this domain blocks access of the catalytic site to the CpG dinucleotide (Song *et al*, 2012).

Taken together, ZF-CxxC domain containing proteins could interpret CpG density to directly or indirectly influence chromatin environment.

2.6 Transcription factors and chromatin

TFs bind only a subset of their potential binding sites in higher eukaryotic genomes (Biggin, 2011). One possible explanation why not all motifs are bound is that not all regions in the genome are equally accessible for TFs due to differences is chromatin structure. High nucleosome occupancy could, for example, prevent TF binding (John *et al*, 2011; Svaren & Hörz, 1997). Indeed, it has been shown that accessibility does correlate with occupancy of many TF classes (Biggin, 2011).

The difference in accessibility of TF binding sites offers the opportunity for more sophisticated regulatory mechanisms. For example, TFs might have to cooperate in order to penetrate closed chromatin, with only the combined DNA affinity of two or more TFs being high enough to bind low accessible regions (Figure 6a)

(Miller & Widom, 2003). In another mechanism, one TF can be required to bind and modify the chromatin environment so that another TF can bind. It is thought that TFs that are able to bind regardless of chromatin accessibility recruit chromatin remodelling complexes to establish accessible chromatin allowing binding of other TFs (Figure 6b) (Voss *et al*, 2011). Additionally, binding of TFs to DNA can be prohibited by DNA methylation (Watt & Molloy, 1988; Iguchiariga & Schaffner, 1989; Prendergast & Ziff, 1991; Campanero *et al*, 2000; Domcke *et al*, 2015). For example, Nrf1 cannot bind its motif if it is methylated. If another factor such as CTCF binds a methylated region containing an Nrf1 motif, this leads to demethylation of the surrounding region and allows Nrf1 to bind (Figure 6c) (Domcke *et al*, 2015). Such mechanisms do not necessarily require direct interaction between TFs and, therefore, could be utilised by the cell to temporally integrate regulatory events.

Rather than interacting with other TFs and their motifs, the DNA context itself in which the motif is placed could serve as means to control binding. For example, one could imagine that high CpG density excludes DNA methylation allowing DNA methylation sensitive TFs to bind (Lienert *et al*, 2011; Krebs *et al*, 2014). This would make sense especially for genes that are active across many cell types and, therefore, accessibility has to be ensured for TFs to bind. Part of this thesis focuses on the impact of CpG density on transcriptional activity, TF binding, and the link to DNA methylation.

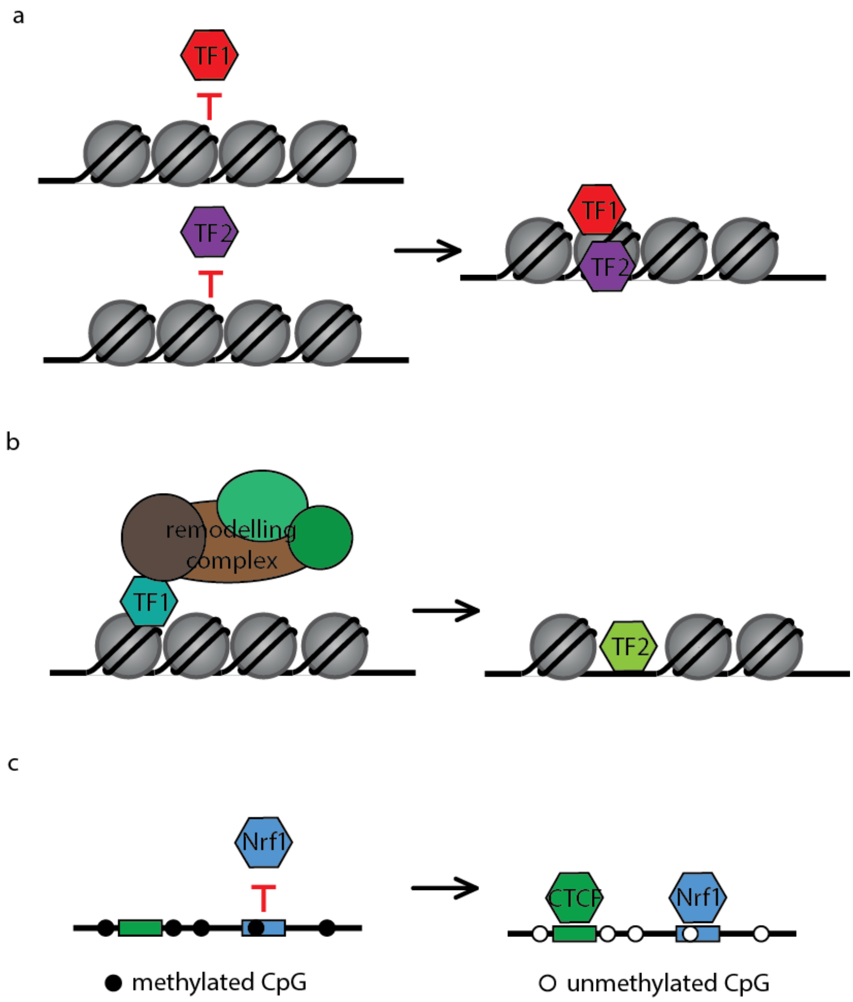


Figure 6: Models for cooperative access to TF motifs on chromatin.

(a) Cooperative binding of two TFs without chromatin remodeling.

(b) TF1 can bind regardless of chromatin state and recruits chromatin remodeling complexes making the DNA accessible for binding of TF2.

(c) CTCF binding leads to localised demethylation of DNA allowing binding of DNA methylation sensitive TFs, like Nrf1.

2.7 Identification of *cis*-regulatory elements

In order to understand the principles of transcriptional regulation, it is essential to identify *cis*-regulatory elements. Several methods have been developed to identify these elements. Promoters can be detected by mapping the TSS of genes across the genome. However, this only informs on where the transcript starts but not about where a promoter region starts and ends. Another, more general method to identify *cis*-regulatory regions is to map chromatin accessibility. One genome-wide method that takes advantage of the difference in chromatin accessibility at open and closed regions is DNase I hypersensitivity site sequencing (DHS-seq). DHS-seq is based on the fact that increased chromatin accessibility correlates with the increased probability that DNA will be cleaved by the endonuclease DNaseI. These cut sites are then located by high throughput sequencing and, therefore, provide a genome wide map of chromatin accessibility (Crawford *et al*, 2006).

A more recently developed method to identify accessible chromatin regions is ATAC-seq (Assay for transposase-accessible chromatin). ATAC-seq is based on a transposase that integrates preferentially at accessible regions. The transposon contains PCR amplification primers that are used to amplify genome wide integration sites. These sites are again detected by sequencing, giving a genome wide profile of chromatin accessibility (Buenrostro *et al*, 2015).

Cis-regulatory regions are marked by characteristic histone modifications such as acetylation of lysines at histones H3 and H4 or H3K4me3 at promoters. These histone marks can be utilised to identify *cis*-regulatory elements (Heintzman *et al*, 2007, 2009; Calo & Wysocka, 2013).

Another chromatin feature marking *cis*-regulatory elements is reduced DNA methylation at CpGs compared to inactive regions of the genome. As for histone marks, this characteristic footprint can be used to identify *cis*-regulatory elements (Stadler *et al*, 2011; Hodges *et al*, 2011; Ziller *et al*, 2013). Furthermore, DNA methylation is detected by bisulfite sequencing, which gives a quantitative measurement of DNA methylation. Another advantage of this technique is that only naked DNA is required as starting material while all the other described approaches require intact nuclei or chromatin, making it suitable also for samples that are difficult to handle.

2.8 Quantification of *cis*-regulatory element activity

Prediction of *cis*-regulatory elements by the above-mentioned methods does not inform on functional relevance or the ability of identified regions to autonomously drive transcription. Therefore, in order to elucidate regulatory principles, transcriptional activity of *cis*-regulatory elements outside of their genomic sequence context has to be functionally tested. To test the ability for *cis*-regulatory elements to autonomously activate transcription they are placed adjacent to a reporter gene whose gene product can be quantified, such as GFP. Promoter activity can be directly tested for their ability to autonomously initiate transcription while enhancers have to be placed in proximity to a minimal promoter, allowing RNA polymerase II to initiate transcription.

Quantification of the reporter gene product can be done on the RNA level using quantitative polymerase chain reaction (qPCR) or *in situ* hybridization followed by imaging. Such assays can be parallelised using next-generation sequencing techniques. To link the transcripts to the *cis*-regulatory element they originate from, unique sequences ('Barcodes') are included in the transcribed sequence (Patwardhan *et al*, 2009, 2012; Shen *et al*, 2015; Mogno *et al*, 2013; Melnikov *et al*, 2012; Kwasnieski *et al*, 2012; White *et al*, 2013). Alternatively, enhancers can be placed downstream of a minimal promoter to directly transcribe the assayed sequence (Arnold *et al*, 2014).

Expression of reporter genes can also be quantified by its enzymatic activity (*e.g.* luciferase or β-galactosidase as reporter genes), fluorescence (*e.g.* GFP) or with specific antibodies. Many of these approaches are compatible with determining activity in whole organisms using imaging based assays that visualise abundance and location of reporter gene products (Shlyueva *et al*, 2014). The use of fluorescent proteins under the transcriptional control of the *cis*-regulatory element can be used for parallelisation. Cells can be sorted based on intensity of the fluorescent protein and DNA from the sorted populations sequenced to identify the *cis*-regulatory elements in each bin of activity. This method is less quantitative compared to barcodes integrated in the RNA but allows to also monitor heterogeneity of activity in the population (Sharon *et al*, 2012; Levo *et al*, 2017).

2.9 Cell type specificity

Cellular identity is brought about by activation and repression of genes leading to characteristic cell type specific gene expression patterns. This is thought to rely on the interplay between TFs and *cis*-regulatory elements. Enhancers display high variability in activity across different cell types, suggesting that they play a central role in cell type specific regulation of gene expression (Xu & Smale, 2012; Smith & Shilatifard, 2014; Smallwood & Ren, 2013; Calo & Wysocka, 2013). To learn more about how cell type specificity is brought about, enhancers have been systematically mapped in a plethora of tissues and cell lines based on their chromatin states. The data from these studies provided a large catalogue of putative regulatory regions (Neph *et al*, 2012; Thurman *et al*, 2012; Ernst *et al*, 2011). One tissue that has an extraordinary diversity of different cell types is the retina, making it an interesting model to study cell type specificity.

2.10 The retina, an example for high cell type specificity within one tissue

The eye is an elaborate sensory organ that allows vision. Perception of light is enabled by the retina, a complex neuronal tissue. It is comprised of more than fifty functionally specialised cell types, each of them contributing to the generation of an image.

2.10.1 The anatomy of the retina and cell type specificity

The cells in the retina are organised in layers. The photoreceptor layer and outer nuclear layer consists of light sensitive cells called photoreceptors that capture light. There are two types of photoreceptors; rods, which grant black and white vision in low light condition, and cones, which allow color vision under bright light conditions. Their outer segments are embedded in the retinal pigment epithelium. Photoreceptors are connected to interneurons like horizontal cells (HCs) and amacrine cells in the outer plexiform layer to integrate and process their signals in the inner nuclear layer (Swaroop *et al*, 2010). Horizontal cells (HCs) are a low abundant cell type. These cells are thought to adjust the systems' response to overall illumination level and enhance contrast between adjacent light and dark regions (Masland, 2001).

Within the inner plexiform layer, connections are made to ganglion and amacrine cells that reside in the ganglion cell layer (Swaroop *et al*, 2010). Part of the signal is processed in this layer. For example, a special type of amacrine cells, the starburst amacrine cells (SACs), are able to discriminate directionality of a stimuluses' movement making them essential for image motion stabilisation (Yoshida *et al*, 2001). Ganglion cells relay the signal via the optic nerve in the optic nerve fiber layer to visual brain centers (Figure 7) (Masland, 2001; Swaroop *et al*, 2010).

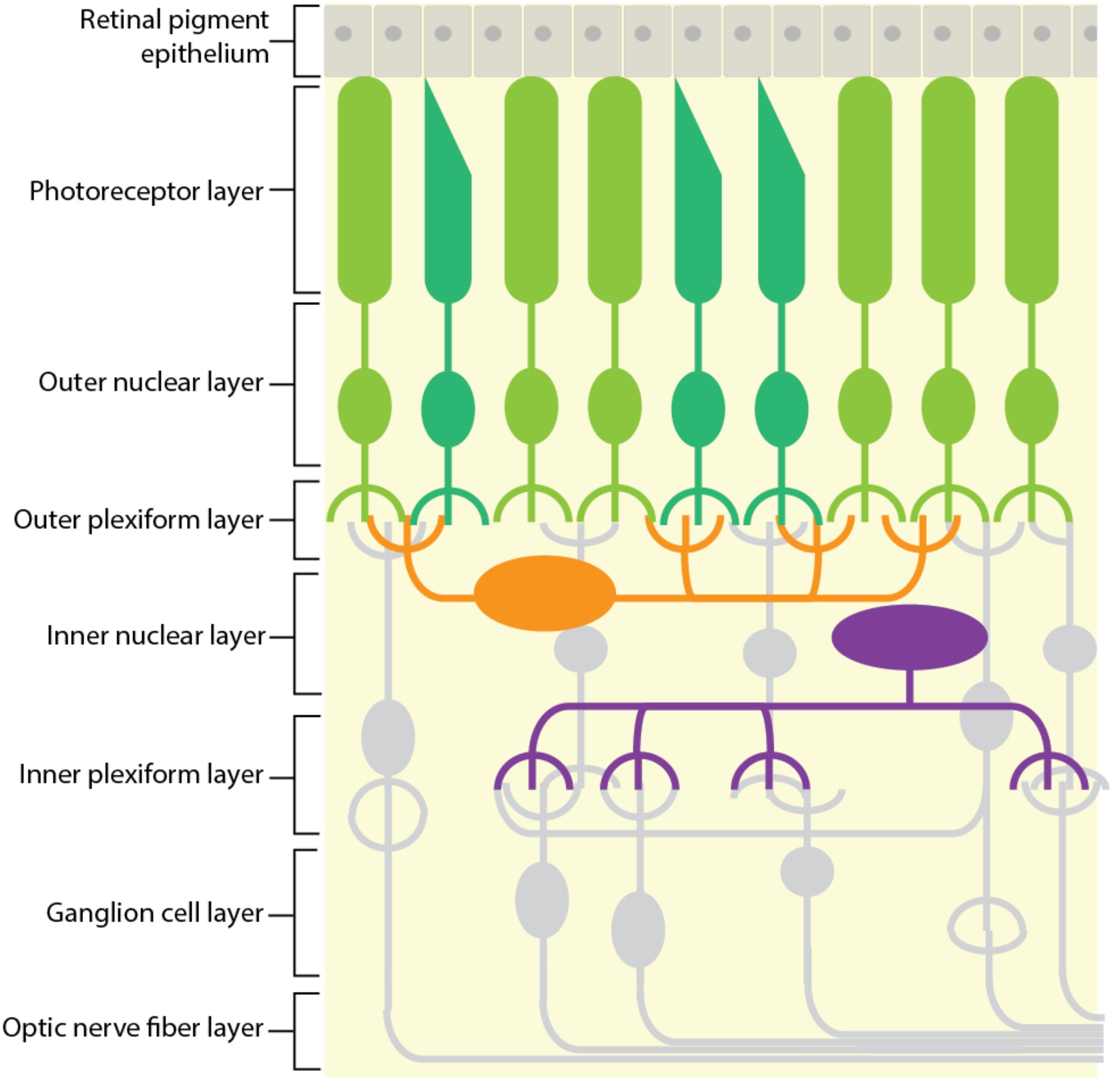


Figure 7: Anatomy of the Retina.

The cells within the retina are organised in layers.

The development and maintenance of different cell types with very specific functions within the retina is based on correct spatiotemporal expression of genes. This has to be tightly controlled by regulatory elements. While the gene expression patterns of many retinal cell types have been studied in health and disease (Siegert *et al*, 2009, 2012), what regulatory elements bring about these expression patterns is still unclear. Information on important regulatory elements is crucial to identify the molecular players interpreting the instructive code to better understand cell type formation and maintenance.

The identification of autonomously active regulatory regions and key TFs is not only important to understand cell type specificity in healthy cells, it is also a crucial step on the way to design gene therapies for retinal diseases. Part of the work in this thesis focuses on the identification and functional validation of regulatory regions in retinal cell types.

2.10.1 Gene therapy

Incorrect gene expression or faulty gene products in retinal cell types can lead to retinopathies that can result in visual impairment and even blindness (Sahel & Roska, 2013). One approach that could aid in improving or curing visual impairment or blindness is gene therapy. The unique morphological characteristics of the eye and the fact that it is immune privileged, makes it especially suited for this type of therapy (Bainbridge *et al*, 2006; Roosing *et al*, 2014). Gene therapy relies on the delivery of a transgene controlled by a promoter-enhancer construct that ensures expression of the transgene in the right cell types. The eye is especially suited for viral delivery of transgenes since it is a small, closed compartment that allows high viral concentrations with relatively low amounts of virus. Additionally, viruses can be delivered to different ocular structures due to the eyes compartmentalisation (Sahel & Roska, 2013).

An example for the potential of gene therapy is RPE65 gene replacement in leber congenital amaurosis, a retina dystrophie leading to visual impairment (Sahel & Roska, 2013). One form of this disease is caused by mutations of the RPE65 gene that is specific for the retinal pigment epithelium. Adeno-associated virus mediated delivery of the functional RPE65 gene, controlled by a CMV promoter, into the eye of dogs that are affected by RPE65 mutations led to restoration of vision (Acland *et al*, 2001). Later, clinical trials in humans showed that this therapy also leads to improved vision in patients (Bainbridge *et al*, 2008;

Hauswirth *et al*, 2008; Maguire *et al*, 2008). This encouraging example shows the potential of gene therapy.

One risk of gene therapy can be adverse effects due to expression of transgenes outside of target cells or wrong expression levels. This could be reduced by specific expression of the transgene at physiological levels in the target cell type. The results of part of this thesis demonstrate a strategy how to make a step towards identification of autonomously active cell type specific enhancers with different activity levels.

2.11 Scope of this thesis

Even though fundamental to biology, we still have a limited understanding of how DNA sequence controls transcriptional activity of enhancers and promoters. The majority of promoters in our genome are rich in the CpG dinucleotide and their high CpG density has been linked via correlation to transcriptional activity (Weber *et al*, 2007; Guenther *et al*, 2007; Thomson *et al*, 2010; Deaton & Bird, 2011; Fenouil *et al*, 2012).

In the first part of the work presented here we asked if CpGs contribute to the transcriptional activity of CpG island promoters. We explored if CpGs play a regulatory role only when they are located in TF motifs or if their overall density in regulatory regions is more important. We further asked how their function relates to DNA methylation.

We addressed these questions by monitoring binding of TFs assumed to play a role in CGI activity. Additionally, we generated a large number of promoter mutants, including artificial promoter sequences, and quantified their transcriptional activity in wild-type and DNA methylation-free murine embryonic stem cells. Together these experiments gave insight into how very short motifs such as the CpG dinucleotide impact transcriptional activity.

To gain further insight into how regulatory regions contribute to cell type specific gene expression patterns, we investigated cell type specific control of gene expression in the mouse retina in the second part of this thesis. Identifying *cis*-regulatory elements is essential in order to understand the transcriptional regulatory principles that control cell-type specification. Important questions that remain include: Which regions in the genome are involved in gene activity within specific cell types? Are the identified regulatory regions autonomously active and how can this be tested in high-throughput in specific cell types? Which TFs play a role in the transcriptional activity of identified regulatory elements?

Towards addressing these questions, we established an experimental framework that allows identification of regulatory elements within specific, and even rare, cell types of the mouse retina. To test autonomous activity of these elements in specific cell types we developed a high-throughput reporter assay. Additionally, we used the generated data to identify TFs that play a role in the activity of regulatory regions in different cell types and tested them by quantifying activity of regions with mutated motifs.

3 Results

3.1 Design principles of CpG Island promoter activity

Prepared manuscript

Design principles of CpG Island promoter activity

Dominik Hartl[1,2], Arnaud R. Krebs[1], Lukas Burger[1], Tuncay Baubec[3], Christiane Wirbelauer[1],Ralph Grand[1], Dirk Schübeler[1,2]

1: Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH 4058 Basel, Switzerland
2: University of Basel, Faculty of Sciences, Petersplatz 1, CH 4003 Basel, Switzerland
3: Department of Molecular Mechanisms of Disease, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

Correspondence should be addressed to: dirk@fmi.ch

Abstract

CpG islands represent the majority of promoters in mammalian genomes. They are CpG rich and mostly devoid of DNA methylation. To which extend and how these characterizing CpGs contribute to promoter activity and regulation remains open. To gain insights into this matter, we monitored binding of transcription factors assumed to play a role for CpG island activity and tested a large number of promoter mutants including artificial promoter sequences for their transcriptional activity with and without insertion into the genome. This revealed that CpG density and motif occurrence is a good predictor for transcription factor binding. Rigorous functional testing of promoter mutants showed that high CpG density is not sufficient for transcriptional activity, yet necessary when combined with more complex transcription factor motifs. Our comprehensive study also reveals that DNA methylation results in a further decrease in transcriptional activity of promoter mutants with low CpG density. This leads to a model where a high CpG density is required to generate a chromatin environment permissive for transcriptional activity.

Introduction

Gene regulation establishes correct spatio-temporal expression patterns essential for cellular function. Correct gene expression is controlled at multiple levels, the first being DNA sequence. Regulatory regions are interpreted by DNA binding proteins such as transcription factors (TFs). Additionally, changes in chromatin structure enable temporal integration of regulatory events through dynamic processes including cell division and organism development.

One essential chromatin component in mammals is DNA methylation of Cytosines (Cs) that reside in a Cytosine-phosphate-Guanine (CpG) context (Lister *et al*, 2009; Stadler *et al*, 2011). The majority of CpGs in mammalian genomes are methylated while unmethylated CpGs are concentrated in specific regions called CpG islands (CGIs) (Bird *et al*, 1985). CGI criteria have been defined to be at least 200bp in length with a G+C content of at least 50% and an observed to expected (OE) ratio of at least 0.6, where OE is number of CpGs / (number of Cs x number of Gs) x length of the region in nucleotides (Gardiner-Garden & Frommer, 1987). In human and mouse, CGIs overlap with ~60% of the promoters, resulting in a bimodal distribution of promoter CpG density (Figure 1A) (Mohn & Schübeler, 2009). Since many CGI promoters are ubiquitously active across many cell types (Larsen *et al*, 1992), most initiation events of RNA Polymerase II occurs within CGI promoters.

There are two mutually not exclusive hypotheses for the enrichment of CpGs at CGI promoters compared to the rest of the genome: (1) CpGs represent a footprint of evolution since the mutation rate differs between methylated and unmethylated CpGs (Bird, 1980). Unmethylated Cs deaminated to uracils (Barnes & Lindahl, 2004) which are efficiently repaired while methylated Cs deaminate to Ts which is a proper genomic base and inefficiently repaired leading to a higher mutation rate. This is supported by the fact that most of the CpG islands do indeed have a lower CpG frequency than expected by chance based on G and C content (OE<1) and that most divergence between close mammalian species is observed at Cs residing within CpGs methylated in the germline (Weber et al., 2007).

In an alternative scenario (2) CpGs can act as a signaling module: CpGs have been thought to contribute to gene regulation via several mechanisms (Bird, 2011). It has been shown that CpG density can protect a DNA sequence from DNA methylation (Lienert *et al*, 2011; Krebs *et al*, 2014; Wachter *et al*, 2014; Long *et al*, 2016). Possible mechanisms could be direct binding of ZF-CxxC domain containing proteins that have been proposed to inhibit or counteract methyl transferase activity (Ooi *et al*, 2007; Cedar & Bergman, 2009). Additionally, also TFs like Sp1 have been linked to keeping CpGs unmethylated (Brandeis *et al*, 1994; Macleod *et al*, 1994).

To distinguish between these two scenarios, the role of CpGs has to be tested functionally. Careful mutation of CpGs within CpG island promoters should inform on whether CpGs themselves contribute to transcriptional activity or not.

Transcriptional activity is mediated by motif specific TFs. However, the majority of TF motif occurrences within the genome are not bound by corresponding TFs (Biggin, 2011). CpG density is a good predictor for promoters (Ioshikhes & Zhang, 2000), since transcriptional activity is mediated by TFs, this could indicate that also CpGs play a role for TF binding. Since CpGs can be part of TF motifs it is hard to separate the contribution of the two to transcriptional activity from each other making functional testing necessary. CpG density can be functionally tested by mutating sequence components of CGI promoters and quantifying transcriptional output in a reporter assay. In order for such experiments to be conclusive a high number of mutations has to be assayed. Using high throughput sequencing transcriptional reporter assays have recently been parallelized (Patwardhan *et al*, 2009, 2012; Shen *et al*, 2015; Mogno *et al*, 2013; Melnikov *et al*, 2012; Kwasnieski *et al*, 2012; White *et al*, 2013). Yet in higher eukaryotes such assays have not been performed on genomic DNA with integration of the constructs in the same genomic site in every cell. Since chromatin environment is thought to contribute to transcriptional activity of CGI promoters genomic integration is essential to obtain conclusive results. Additionally, integration at the same genomic locus allows quantification of transcriptional activity independent of positional effects.

To comprehensively address CpG contribution to transcriptional activity, we monitored binding of transcription factors assumed to play a role for CpG island activity and analyzed the relationship of binding data and sequence features. To functionally test the role of CpGs we quantified transcriptional activity of a large number of promoter mutants including artificial promoter sequences in murine embryonic stem cells as a model. To monitor the relationship of DNA methylation and promoter mutant activity we assayed the mutant libraries in wildtype and DNA methylation deficient cells.

Results

TFs binding motifs in CGIs

Transcriptional activity is mediated in principle by TFs that bind to specific motifs on DNA. However, predicting TF binding from their preferred motif generally proofs to be difficult in higher eukaryotes. Only a minor fraction of occurring motifs tends to be bound at any given cell state (Biggin, 2011) because local chromatin environment is thought to be an additional determinant of occupancy (Biggin, 2011). This absence of correct prediction ultimately requires to map binding sites in vivo. To gain insights into this problem at CpG islands, we mapped a set of four TFs that are broadly expressed across many different cell types and tissues making them candidates to control housekeeping gene activity (Supplementary Figure 1 a-d). Of these, Sp1 and Sp3 have rather low complexity motifs (Figure 1c and d). The low complexity implies that the motif occurs frequently by chance. Indeed, if we generate random sequences with the length of CpG islands and their average dinucleotide composition, half of these “CGI like” sequences contain Sp1 or Sp3 motifs. In contrast, Gabpa and Nrf1 motifs display higher complexity (Figure 1e and f) and consequently their motifs only occur in 15% and 10% of random sequences (matches to highest scoring 7bp). This randomization illustrates an intrinsic problem in that G+C rich motifs of low complexity occur at high frequency by chance in any CGI sequence. In turn this makes them frequent hits in any motif prediction. Notably, Sp1 and Sp3 have already been directly implicated in controlling CGI activity based on single promoter analysis (Brandeis *et al*, 1994; Macleod *et al*, 1994). To test how this actually resembles and predicts binding, we performed ChIP-Seq for Sp1, Sp3, Gabpa and Nrf1 in mouse embryonic stem cells using the 'Rambio' approach (Baubec *et al*, 2013). This system of controlled expression of a protein of interest combined with a strong affinity tag generated reproducible high quality data for all four factors (Figure 1b and Supplementary Figure 1e and f). In case of NRF1, we had previously generated ChIP with an antibody enabling us to directly compare tagged versus endogenous protein, which revealed a highly consistent binding pattern (Supplementary Figure 1g). In case of all three proteins, a comparison of Bio-ChIP signal to motif strength reveals that bound sites contain the motif but that nevertheless the motif itself is only a poor predictor of actual binding due to many motif occurrences that are unbound. In total only ~5% of high scoring Sp1 and Sp3 sites are bound while ~25% of high scoring Gabpa and Nrf1 motifs are bound. Comparison of motif strength with TF binding reveals that for all four factors only high scoring motifs are robustly bound. This suggests that these factors are indeed highly specific to their motif (Figure 1g-j). Importantly, this is also true for Sp1 and Sp3 that have low complexity motifs that in variations might also occur by chance.

CpGs are enriched at the majority of promoters compared to the rest of the genome and can even be used to predict promoters (Ioshikhes & Zhang, 2000). We were wondering if this short dinculeotide has predictive power for TF binding. For such a short sequence the motif itself cannot be used but rather the local frequency at which it occurs. Consequently, we focused on CpG density that we calculated by normalizing to the expected frequency based on G and C content in the sequence. We then contrasted CpG density with TF binding. For all four factors the relative number of bound sites starts to rise at normalized CG density (Observed over expected=OE) of ~0.6 and then gradually increases (Supplementary Figure 1i-l). We note that an OE of 0.6 is also the most commonly used threshold for the definition of CGIs (Gardiner-Garden & Frommer, 1987).

CpG density of OE >0.6 and no motif information performs slightly better in predicting binding than a high scoring motif alone. For Sp1 and Sp3 ~20% and for Gabpa ~40% of windows with high CpG densities are bound (Figure 1k-m, Supplementary Figure 1i-l). How can it be explained that high frequency of a low information dinucleotide predicts binding better than individual occurrence of more complex binding sites? One possibility is that either CpG density itself directly or indirectly increases affinity for Sp1, Sp3 and Gabpa. Alternatively, CpG density could correlate with another genomic feature such as open chromatin which defines the accessible part of the genome. Indeed, when comparing both CpG density does correlate with accessibility (Supplementary Figure 1h). Accessibility performs better as a predictor arguing that all binding events of above factors occur in open chromatin (Supplementary Figure 1m-p). While this correlation does not allow to infer causality, it is tempting to speculate that CpGs play a role since at CpG densities bigger than ~0.6 most regions are indeed accessible (Supplementary Figure 1h).

Interestingly CpG density performs much worse in predicting Nrf1 binding than the motif of this factor itself (Figure 1n, Supplementary Figure 1l). This could be a direct reflection of motif complexity. If all binding events occur in open chromatin and at a correct motif, the variable with the least occurrences will perform better in the predictor. In case of a low complexity motif this might be open chromatin but not in case of a high complexity motif.

To test the predictability of CpG density and high scoring motif together we combined the two. Indeed, windows with high CpG density that contain the motif are most likely to be bound for all factors. ~20% are bound by Sp1 and Sp3 while Nrf1 and Gabpa with their higher complex motifs bind ~80% and ~90% of the windows respectively (Figure 1k-n). One possible interpretation of this result is that high CpG density facilitates binding for all four factors. CpG rich regions often co-occur with CpG island promoters that tend to be active. Therefore, an alternative explanation would be that increased binding is simply based on non-functional co-occurrence, rather than on a functional relationship of increased binding and CpG density. This would be also in agreement with the fact that the fraction of bound sites starts to increase from a CpG density of 0.6, which is the lower limit of the CpG island definition (Supplementary Figure 1i-l) (Gardiner-Garden & Frommer, 1987). Indeed, within active promoters the differences in bound sites of promoters with the motif versus promoters with both, motif and high CpG density, are lower for all factors (Figure 1o-r) compared to the data in windows tiling chromosome 19 (Figure 1k-n). While this could suggest that the contribution of CpG density does not play a major role here, we also cannot exclude that high CpG density is required for TF binding and promoter function.

Taken together we cannot conclude from this data whether CpG density functionally contributes to binding of the tested TFs. CpG rich regions tend to be more accessible and are more likely to be transcriptionally active than CpG poor regions. To control for such confounders in genomics analysis, one would usually compare sequences of similar base composition but different regulatory potential.

This however is not possible for CGIs since all of these are regulatory regions that are under functional selection. This makes it difficult to elucidate if CpGs themselves aid TF binding or if this is just a coincidence of their evolutionary origin.

Factors like Sp1 that have a G and C rich, low complexity motif are even more prone to this bias which explains why they are so frequently enriched upon motif enrichment on genomic events that occur on CGIs such as transcription but also Polycomb repression.

The absence of control sequences that lack regulatory activity excludes to elucidate the role of CpGs for TF binding and transcriptional activity by correlative studies. Rather this requires rigorous experimental testing of functionality in transcription upon sequence variation.

Accessing reporter assay at a defined chromosomal site

Dissecting what DNA sequence components define the transcriptional activity of CGIs requires an experimental approach that allows systematic iteration to test many different promoter sequences in parallel and on chromosomal DNA. This in turn requires parallel analysis of sequence libraries and the ability to link RNA molecules to a specific regulatory region that is not part of the transcript.

Towards this goal we designed a reporter cassette that allows cloning of promoter sequences as a pool. In order to assign transcripts to their specific promoter we added barcodes (BCs) to the transcribed region. Such strategy has already been performed successfully in transient reporter assays (Patwardhan *et al*, 2009, 2012; Shen *et al*, 2015; Mogno *et al*, 2013; Melnikov *et al*, 2012; Kwasnieski *et al*, 2012; White *et al*, 2013) (Figure 2a). To account for the contribution of chromatin to transcriptional regulation, we integrate the promoter-barcode constructs into a defined chromosomal context using recombinase mediated cassette exchange (RMCE) (Lienert *et al*, 2011; Arnold *et al*, 2013; Krebs *et al*, 2014; Jermann *et al*, 2014). Specifically, we integrated the libraries into the β-globin locus in ESCs that was shown to be epigenetically and transcriptionally inactive in most cells besides the erythroid lineage of the blood (Fromm & Bulge, 2009). This allows quantification of transcriptional activity on chromatin independent of positional effects. BC frequencies can be quantified by isolating RNA and DNA, amplification of BCs followed by next generation sequencing of the BCs in the two fractions. Frequencies of individual barcodes in the RNA fraction are then normalized to the representation of every promoter-barcode construct in the cell population (Figure 2a).

Averaging the signal of all barcodes corresponding to one promoter mutant (at least 3, median across all BCs) allows us to accurately and reproducibly quantify activity of individual promoter mutants (Supplementary Figure 2a-b, Supplementary Figure 3a,d-f, Supplementary Figure 4 and 5).

We utilized this assay to measure up to ~3100 Promoter-BC constructs within a single experiment. In total we tested more than 10000 Promoter-BC constructs representing ~ 270 unique sequences. To our knowledge this is by far the highest number of sequences tested for their transcriptional potential at a single genomic site in a higher eukaryote.

High CpG density alone is insufficient for CGI activity

CpG density performs remarkably well in predicting binding of Sp1, Sp3 and Gabpa. If this is a direct consequence of CpG density is unclear. It has been shown that high CpG density coincides with transcriptionally permissive chromatin, such as trimethylation at lysine 4 of histone H3 (H3K4me3) and lack of DNA methylation also at artificial sequences (Wachter *et al*, 2014; Lienert *et al*, 2011; Krebs *et al*, 2014).

To test if high CpG density alone is sufficient for transcriptional activity we quantified transcriptional activity of DNA sequences containing high CpG density but that are not under functional selection for regulation and thus should only contain TF binding sites by chance. More specifically, we chose sequences from the prokaryotic *E. coli* genome which were amplified and inserted in front of a minimal promoter and inserted them into the mouse genome. When quantifying activity of these constructs we observed no or very little transcriptional activity (Figure 2b). This suggests that high CpG density is not sufficient for transcriptional activity when assayed on chromatin. To ask if this reflects repression by chromatin we tested the same library transiently without integration into the genome. However, also on a plasmid only low to no transcriptional activity could be observed (Figure 2C).

Taken together, this data suggest that CGIs are not transcriptionally active based on their CpG density alone. However, even if CpGs are not sufficient for transcriptional activity, they might still be a necessary feature for functional CGIs. Alternatively, they could be only a footprint of evolution. Methylated CpGs show faster mutation rates than the unmethylated form. Since CGIs are unmethylated in the germline this could lead to a decreased loss of CpGs. This is supported by the fact that most of the CpG islands do indeed have a lower CpG frequency than expected by chance based on G and C content (OE<1). Therefore, lack of DNA methylation would pose a plausible explanation for higher CpG density at CGIs compared to the rest of the genome.

Independent of this explanation, CpGs could also serve as a signaling module. CpG rich promoters tend to be transcriptionally more active than CpG poor promoters and as shown above, CpGs positively correlate with TF binding. To directly test for CpG contribution, CpGs within active CGI promoters have to be carefully mutated to decrease CpG density followed by monitoring of their transcriptional activity.

Identification of CGI promoters that are autonomously active

Having established that CpG density alone does not confer transcriptional activity, we wanted to next test endogenous CGIs for their ability to drive reporter activity. We chose promoters that are broadly active (based on data from (Shen *et al*, 2012)) and that are bound by the TFs Sp1, Sp3, Gabpa and Nrf1. We used DNase hypersensitive regions to define borders of the promoters (data from (Domcke *et al*, 2015)). Based on these sequences we generated libraries of promoters and tested their transcriptional activity in the reporter assay. About half of the tested promoters displayed transcriptional activity when inserted in the same locus on genomic DNA (Supplementary Figure 3b and c). This provided us with sufficient candidate promoters for further studies.

CpG density contributes to CGI activity

Having identified functional CGIs that drive promoter activity, we wanted to test if high CpG density is a required feature for transcriptional activity of CGIs. In order to do so we systematically mutated CpGs within CGI promoter sequences and quantified the effect on transcriptional activity. More specifically we mutated

CpGs within two housekeeping gene promoters (Snx3 and Pwp2) that are autonomously active in our reporter assay (Supplementary Figure 3b and c). Since mutations of CpGs within TF motifs would lead to effects unrelated to CpG density, we used previously published TF binding data (see methods) and our set of mapped TFs to avoid changing CpGs within TF motifs. We mutated CpGs in four (Pwp2) or five (Snx3) tiling windows and generated all possible combinations of WT and mutant windows. The resulting promoter mutant libraries were then tested for transcriptional activity in the reporter assay after chromosomal insertion (Figure 3a).

This revealed that presence of CpGs positively correlates with transcriptional activity for both promoters (Figure 3b, c). For both tested promoters activity decreases rapidly between the normal CpG density and roughly ~0.6 (Figure 3b, c). Further decrease does not cause further reduction but activity plateaus at a low level for the Snx3 promoter likely suggesting that the promoter is off (Figure 3b). Strikingly, this also coincides with the lower limit of CGI CpG density (Figure 1a). In case of the Pwp2 promoter we did not recover constructs with very low CpG densities (Figure 3c).

Theis data argues that CpG density, while not alone sufficient for activity nevertheless contributes to CGI activity. In order to test if this is a general feature we mutated ten additional CGI promoters but with a lower number of permutations. We chose CGI promoters that span a range of CpG densities (0.71 and 1.06) and transcriptional activity in the reporter assay (24-fold) (Figure 3d and Supplementary Figure 3g). As before, we mutated CpGs outside of identified or predicted TF motifs. Five different CpG densities were generated for each promoter using random combinations of CpGs.

This more comprehensive set of tested promoters shows the same response to loss of CpGs. The more CpGs were mutated the lower the transcriptional activity of the promoter (Figure 3c). Even with this lower number of tested CpG densities per promoter it becomes again apparent that at a CpG density of ~0.6 most promoters show strongly reduced activities (Figure 3c).

While this argues for a general contribution of CpGs to CGI transcriptional activity, we cannot formally exclude that the performed mutations of CpGs affected unknown TF motifs and thereby causing a decrease in transcriptional activity.

While this might seem unlikely given how mutated CpGs were chosen, we nevertheless wanted to explore this scenario using a different permutation approach.

Dissecting CpGs from TF motifs

We reasoned that mutating CpGs that are critical components of complex TF motifs should have direct effects on promoter output. To identify such CpGs we generated mutants of short regions throughout the entire promoter and monitored their transcriptional activity. More specifically, we mutated 10 bp windows covering the entire Pwp2 promoter and in each case replaced the 10 bp with a random CpG free sequence.

Measuring the resulting 42 mutants after genomic insertion revealed indeed highly variable effects on promoter function. About half of the mutated 10bp windows do not have an effect on transcriptional activity (Figure 4b). When we focus on a region with predicted and high scoring TF motifs we see that many but not all of these have an effect on transcriptional activity (Figure 4a, b). This agrees with mutations of individual motifs that correspond to TFs that bind the Pwp2 promoter (Figure 4d). Comparison of ChIP-seq signal at the endogenous Pwp2 promoter and transcriptional activity of mutants shows that ChIP-seq, as expected, lacks the spatial resolution to correctly predict if a specific motif is bound and in turn contributes to activity of the promoter.

Additionally, we observe a decrease in activity when mutating regions that are downstream of the highest signal for initiation as measured by CAGE in the endogenous promoter -(cap analysis by gene expression) (Forrest *et al*, 2014) (Figure 4b,c). This could indicate that these regions are important for initiation, leading to decreased transcriptional activity when mutated.

Taken together we were able to characterize regulatory function of the Pwp2 promoter at a 10bp resolution. Knowledge about regions important for transcriptional activity allows us to generate mutants in a more educated manner without unintentionally disrupting CpGs critical for transcriptional activity since they are part of a complex motif.

Mutation of CpGs within regions that themselves do not contribute to transcriptional activity

To assess the contribution of CpGs outside of TF motifs, we subsequently mutated only those CpGs located within regions of the Pwp2 promoter that showed no or only minor effects on transcriptional activity. We randomly mutated these CpGs to obtain mutants with different levels of CpG density (Figure 5a).

Following insertion of the resulting 11 tested sequences we measured their transcription output. The resulting data reveal that mutating CpGs outside of TF motifs decreases activity. More specifically, CpG densities lower than ~0.7 showed decreased activity compared to the WT (Figure 5b). This indicates that CpGs outside of important regions in the Pwp2 promoter nevertheless contribute to transcriptional activity. Interestingly, we do observe an initial increase in activity at CpG densities between 1 and 0.8 (Figure 5b). This increase occurred across different mutated sets of CpGs, therefore we do think this is a function of CpG density rather than a mutation specific effect. At this point we cannot determine if this is a promoter specific effect. However, among the 12 tested promoters, this was the only one that displayed a clear increase upon a specific CpG mutation.

Taken together, deletion of CpGs that are not part of TF motifs also causes reduced activity. We next wondered if adding CpGs into an artificial promoter would in turn increase activity.

Effect of increasing CpG density of an artificial promoter

To further test if CpG density leads to increased activity we increased CpG at an artificial promoter. We replaced all regions of the Pwp2 promoter that showed no or only minor effects on transcriptional activity with the CpG free sequence used for the mutants described above (Figure 5c). This allowed us to replace ~60% of the sequence and to decrease CpG density from ~1 to ~0.6. We then added CpGs into this random sequence at the same position as in the WT promoter. This lead to a gradual increase in activity, depending on CpG density. We reached up to ~53% of wildtype Pwp2 activity and an activity of around 25% of WT Pwp2 when adding the same number of CpGs as in the WT (Figure 5d).

These data argue again that CpG density itself does contribute to CGI activity.

DNA methylation decreases activity of mutants with low CpG density

Several mechanisms how CpGs could contribute to promoter activity have been suggested. High CpG density has been shown to antagonize DNA methylation, therefore one explanation for decrease in activity when mutating CpGs could be the resulting increased DNA methylation (Lienert *et al*, 2011; Krebs *et al*, 2014). This effect is more likely to play a role for mutants in the lower range of CpG density as mutants with higher CpG density are expected to be still unmethylated (Lienert *et al*, 2011; Krebs *et al*, 2014). Such de novo methylation might lead to decreased binding of methylation sensitive TFs such as Nrf1 (Domcke *et al*, 2015) or recruit MBD proteins (Baubec *et al*, 2013).

In order to test directly if de novo DNA methylation accounts for reduced transcriptional activity of CpG depleted mutants we repeated these experiments in cells that lack DNA methylation. More specifically, we integrated CpG mutants into Dnmt1, Dnmt3a and Dnmt3b triple-knockout (TKO) ESCs (Domcke *et al*, 2015) but at the same genomic site. We similarly tested their transcriptional activity and compared it to that of the methylation competent cells. This revealed that presence of DNA methyltransferases has no significant effect on transcriptional activity for those mutated promoters that show high CpG density (>0.6), which seems in agreement with our expectations that sequences with high CpG density stay unmethylated in any case.

Sequences with lower CpG densities, however, behave differently as they tend to show higher activity in cells without DNA methylation than in wildtype cells (WT) (Figure 6a). Importantly, the difference in activity between WT cells and cells lacking DNA methylation increases with decreasing CpG density (Figure 6a, b). Consistent with this result, activity starts to deviate between WT and TKO significantly at a CpG density of about 0.6 which nicely aligns with the CGI definition. One potential interpretation of this result is that DNA methylation might partially contribute to the decrease in transcription upon CpG mutation. The difference is CpG density dependent. This supports a model where decreased CpG density leads to increased DNA methylation. This in turn could prevent binding of TFs sensitive to DNA methylation to the promoter, leading to decreased transcriptional activity. Removal of the DNA methylation machinery therefore leads to an increase in binding and therefore increased transcriptional activity at lower CpG densities (Figure 6c).

In order to test whether decreased activity of promoter mutants correlates with DNA methylation we plan to monitor DNA methylation levels (work in progress).

Discussion

This study discloses a functional role of CpG density for the transcriptional output of CpG island promoters. By combining genome wide profiling of transcription factors with high throughput genomic insertion of promoter mutants we show that a high CpG density is not sufficient yet necessary for full activity of CGIs.

In order to obtain comprehensive data, development of an improved parallel reporter assay proved essential. Previous studies described substantial differences in transcriptional activity of constructs depending on chromosomal or episomal context (Inoue *et al*, 2017) or genomic location (Akhtar *et al*, 2013). In this study, highly reproducible and sensitive measurements were enabled by assaying constructs after insertion in the same genomic locus. As a result, only one construct was tested per cell but multiple measurements for each fragment within the cell population. This sensitivity allowed proper quantification of subtle changes in transcriptional activity. To our knowledge such throughput and sensitivity has not been achieved before in a higher eukaryote. Importantly, this assay can also be utilized to explore other sequence features of promoters and can directly be adapted for enhancers.

The data in this study reveals that CpG density functionally contributes to transcriptional activity of promoters in combination with more complex TF motifs. This finding agrees with correlative evidence linking high CpG density to active chromatin and transcriptional activity (Weber *et al*, 2007; Guenther *et al*, 2007; Thomson *et al*, 2010; Deaton & Bird, 2011; Fenouil *et al*, 2012; van Arendsbergen *et al*, 2016) as well as functional assays showing that CpG dense sequences are free of DNA methylation (Lienert *et al*, 2011; Krebs *et al*, 2014). In contrast to our results, a recent study suggested based on four artificial constructs that activity of CGIs depends on high G and C content rather than high CpG density (Wachter *et al*, 2014). Wachter et al monitored activity based on chromatin status while transcripts were not quantified. We cannot formally exclude a contribution of G and C content to transcriptional activity as we focused on CpGs. However, given the large number of mutants, high resolution of mutated CpGs and coinciding minimal changes of G and C content within CpG mutants strongly argues that it is primarily CpG density that contributes to transcriptional activity.

CpGs could support transcriptional activity indirectly by increasing DNA accessibility and thereby facilitating TF binding. Such increase in openness likely depends on local concentration of CpGs. In agreement with this hypothesis, accessibility of genomic regions correlates with CpG density genome wide. This is consistent with published data showing that CpG rich artificial sequences display marks of open chromatin (Wachter *et al*, 2014; Lienert *et al*, 2011; Krebs *et al*, 2014) and that the TFs tested in our study preferentially bind their motif when located in CpG rich regions. This relationship raises the question if accessibility decreases upon CpG depletion in promoter mutants. To answer this question, accessibility could be monitored at mutant constructs using NOMe-seq (Kelly *et al*, 2012; Nabilsi *et al*, 2014; Krebs *et al*, 2017).

One possible explanation how CpGs mediate accessible chromatin could be via direct recruitment of CpG binders such as ZF-CxxC domain containing proteins. These bind only unmethylated CpGs and have been correlated to an accessible chromatin environment before (Blackledge *et al*, 2010; Clouaire *et al*, 2012, 2014; Boulard *et al*, 2015). One candidate for translating CpG density into open chromatin is Cfp1 which is part of the H3K4 methyltransferase complexes Setd1A and Setd1B (Clouaire *et al*, 2012). H3K4me3 occurs around the TSS of promoters and can be bound by the chromatin remodeler Chd1 (Clouaire *et al*, 2012; Flanagan *et al*, 2005). Consequently, ZF-CxxC domain containing proteins could interpret CpG density and indirectly lead to chromatin remodeling, allowing TF binding and thereby resulting in transcriptional activity. Such putative role of ZF-CxxC domain containing proteins could possibly be tested by protein reduction in cells containing promoter mutant libraries and quantification of transcriptional activity. Changes in H3K4me3 can be monitored performing ChIP on individual constructs with and without knock-downs.

Importantly, the effects of mutating CpGs were rather uniform regardless if positioned distal or within the site of transcriptional initiation. This was similarly the case in the artificial promotor constructs. This result argues that CpGs have no particular local function. Nevertheless, overall CpG density in the promoter could indirectly affect initiation. It was shown that CpG island promoters are depleted in H3K36me2, a chromatin mark that can interfere with transcriptional initiation (Carrozza *et al*, 2005; Li *et al*, 2009; Strahl *et al*, 2002; Youdell *et al*, 2008). Removal of H3K36me2 is catalyzed by the ZF-CxxC domain containing protein KDM2A (Blackledge *et al*, 2010). Therefore, KDM2A binding as a result of high CpG density could protect the promoter from H3K36me2 thus preventing interference with transcriptional initiation (Blackledge *et al*, 2010). To test presence of H3K36me2 at promoter mutants with low CpG density ChIP can be performed on individual constructs.

In non-transformed cells DNA methylation does not occur at CpG rich DNA sequences but only at low CpG densities (Lienert *et al*, 2011; Krebs *et al*, 2014). We showed here that removal of DNA methyltransferases leads to increased transcriptional activity of mutants with low CpG densities. This data indicates that DNA methylation causes part of the decreased transcriptional activity upon depletion of CpGs. The change at low CpG densities exclusively agrees with previously published data showing that CpG dense regions do not display DNA methylation (Krebs *et al*, 2014). In order to test whether decreased transcriptional activity of promoter mutants correlates with DNA methylation we plan to monitor DNA methylation levels (work in progress). It is tempting to speculate that high CpG density is required at CGIs to prevent DNA methylation (Lienert *et al*, 2011; Krebs *et al*, 2014) which allows binding of methylation sensitive TFs (Watt & Molloy, 1988; Iguchiariga & Schaffner, 1989; Prendergast & Ziff, 1991; Campanero *et al*, 2000; Domcke *et al*, 2015) (Figure 6c). A protective function of high CpG density against DNA methylation is a potential explanation why high CpG density together with motif occurrence is such a good predictor for TF binding. Protection from DNA methylation at CGIs could be mediated by ZF-CxxC domain containing proteins like KDM2B. Genetic deletion of this protein has indeed been shown to result in rather slow but cumulating DNA methylation at inactive CGIs in stem cells (Boulard *et al*, 2015).

How does our observation of a functional link between CpG density and transcriptional activity relate to existing models of the evolutionary origin of CpG islands? Previous theoretical analysis indicated that the high CpG content in CGIs can be explained by a neutral effect of slow deamination associated with the lack of methylation revealing no evidence for purifying selection on CpG densities

(Cohen *et al*, 2011). The study by Cohen et al are fully compatible with our data that argues that overall density rather than individual positions is relevant. Morever, the observed and evolutionary maintained CpG density at CGIs is sufficiently high to mediate transcriptional activity without the need for classical natural selection. For this model to be true the equilibrium between spontaneous deamination and gain of CpGs within CGIs has to be higher than the threshold of CpG density for transcriptional activity, which we estimate to be ~0.6. Within unmethylated genomes like the invertebrate *Drosophila melanogaster* CpG density is ~1 which most likely also represents the equilibrium. We assume that the balance point between spontaneous deamination and gain of CpGs is lower within CGIs since they reside in methylated genomes. Although CGIs generally do not display DNA methylation, stochastic methylation events in the germline could lead to a lower equilibrium than in unmethylated genomes. The exact number remains to be investigated but we expect the equilibrium to be higher than 0.6 as otherwise CGIs with lower CpG densities would be under selective pressure.

Taken together the data in this study underlines the importance and complexity of sequence context for transcriptional activity.

The here reported functional link between CpG density and transcriptional activity at CGI promoters exposes a function of dinucleotide frequencies. Given the different structure of CpG poor promoters and enhancers it is tempting to speculate that other low complexity motifs might function at these elements as an additional means of regulation. Our study should serve as a starting point as it provides the experimental framework for rigorous testing of putative regulatory roles of dinucleotides in promoters and enhancers.

Materials and Methods

Cell culture

Mouse ES cells were cultured on cell culture plates coated with 0.2% gelatine in DMEM medium with 15% fetal calf serum, 2mM L-glutamine, 1 x non-essential amino acids, LIF and 0.001% β-mercaptoethanol. Cells were incubated at 37 °C with 7% CO2. TC-1 cells with a RMCE site in the beta-globin locus were used for integration of expression libraries for the transcriptional reporter assay (Lienert et al, 2011).Dnmt1, Dnmt3a and Dnmt3b was deleted in these cells using CRISPR-Cas9 gene editing as previously described (Domcke *et al*, 2015).

Mouse HA36 ES cells (mixed 129-C57Bl/6 strain) were used for Bio-ChIP (Baubec *et al*, 2013).

Reporter Assay

Generation of barcoded reporter vector

A cassette containing loxP site, multiple cloning site, poly-A signal and another loxP site was synthetized and cloned into a plasmid backbone containing ampicillin resistance (Lienert *et al*, 2011). Barcodes were generated by annealing CGCCGAANNNNWNNNNWNNNNNAGCTCGG and TCGACCGAGCTNNNNNWNNNNWNNNNTTCGGCGCATG. Vector was cut using SphI and SalI and ligated with the annealed barcodes with T4 ligase. Ligation was precipitated and 100 ng were transformed into MegaX DH10B™T1ᴿ Electrocomp™ Cells (Thermo Fisher). 1:10 000 dilution was distributed on a LB agar plate containing 50mg/L ampicillin to estimate transformation efficiency. The rest was incubated in 50ml LB containing 50mg/L ampicillin shaking at 300 rpm at 37 °C over night. Plasmids were isolated using Quiagen Plasmid Midi Kit.

Library cloning and RMCE

Promoter libraries were cloned into the expression vector using ClaI and NheI aiming for at least ten times more colonies than unique promoters. To link barcodes and promoters the Promoter-BC fragment was amplified with Primer DH.P39 and one of the Indexing Primers containing the Illumina flow cell annealing sequences using Phusion Hot Start II polymerase (Thermo Scientific). PCR products were purified using AmPure XP beads (Beckman Coulter, #A63880) and directly sequenced using MiSeq 500 or 600 cycle Kits. The vector was cut with SphI and PacI or NheI and a sequence containing a CpG free eGFP and the annealing sequence for Primer DH.P6 was cloned in optional the insert contained a 31bp minimal promoter in front of eGFP.

RMCE was performed as previously described (Krebs *et al*, 2014).

RNA / DNA isolation and preparation for next-generation sequencing

RNA was isolated from cell lines containing the expression libraries with Quiagen RNeasy® Mini Kit with on-column DNase digestion and reverse transcribed using Takara PrimeScript RT Reagent Kit (#RR047A). For DNA isolation cell pellet was resuspended in Bradleys-Buffer, 6 µl RNaseA (10mg/ml) was added and samples were incubated for 1h at 37 °C. Subsequently 30 µl protease K was added and samples were incubated at 50 °C over night. Then DNA was extracted using Phenol and Chloroform.

DNA and cDNA barcodes were amplified with KAPA HIFI Hotstart using Primer DH.P6 and indexing primer (Table 1). PCR products were purified using AmPure XP beads (Beckman Coulter, #A63880) and sequenced using 50 cycle Kit on HiSeq 2500.

Generation of Biotin-Tagged TF cell lines

Biotin-tagged TF cell lines were generated as previously described (Baubec *et al*, 2013). Bio-Gabpa was expressed using a Cag as well as a CMV promoter while Bio-Sp1 and Bio-Sp3 were expressed using Tet-inducible promoters induced with 1mg/L Doxorubicin for 24h.

ChIP

Bio-ChIP was performed as previously described (Baubec *et al*, 2013).

Reporter Assay Data Analysis

Barcode to Promoter assignment

Fastq files were trimmed to the promoter sequence and aligned using bowtie. If design of mutants did not allow alignment reads were matched to the reference sequences using "stringdistmatrix" function in R. We allowed 100 errors throughout the read and a minimum distance to the next closest reference sequence of n-1 where n is the minimum distance to the next closest distance within the reference sequences. Sequences from Read2 between “CGTTTAAACTGTCGACCGAGCT” and 'TTCGGCGCATG” were extracted as barcodes and the reverse complement was generated. Barcodes and aligned reads were matched by read ID. Barcodes that were associated with one unique sequence or with a sequence that represents >90% of all reads of a barcode were used for the analysis, the rest was discarded.

Quantification of transcriptional activity

Analysis was performed on triplicates. Barcode sequences were extracted from 50bp reads by taking only reads starting with the expected backbone sequence: “TCCTGCTGGAGTTCGTGACCTGCATGCGCCGAA”. From these reads the sequence at position 34-48 was extracted. The frequency of each barcode sequence was calculated to get counts for each sample. Counts of barcodes were normalized to library size. Enrichment of barcodes in the RNA sample was calculated over their representation in genomic DNA. Only barcodes that were sufficiently represented on genomic DNA were used for further analysis (depending on sequencing depth 10-50 reads). In case a barcode was sufficiently represented on genomic DNA but not sequenced in the RNA fraction we assumed that this was caused by low expression levels and assigned 0 counts to the RNA barcode. The median activity of all barcodes per CRE was calculated. Only CREs that were covered in at least 2 out of 3 replicates with at least three barcodes were used for downstream analysis.

ChIP Data Analysis

Reads were aligned using bowtie (Langmead *et al*, 2009) to mm9 and peaks were called using Peakzilla with default parameters (Bardet *et al*, 2013).
Position weight matrices of motifs were generated based on called peaks in bio-ChIP-seq data using HOMER (Heinz *et al*, 2010).

Published data sets

The following ChIP-seq datasets were downloaded from GEO:

Nanog, Mycn, Oct4, Smad1, Sox2, STAT3, TCFCP2l1, Zfx (GSE11431) (Chen *et al*, 2008), Rex1 (GSE36417) (Gontan *et al*, 2012), Tbx3 (GSE19219) (Han *et al*, 2010), Tcf3 (GSE11724) (Marson *et al*, 2008), YY1 (GSE31786) (Vella *et al*, 2012), Zic2 (GSE61188) (Luo *et al*, 2015), CTCF (GSE30206/GSM747534) (Stadler et al, 2011), NFYA (GSE25533/GSM632038) (Tiwari et al, 2011), NRF1 (GSE67867/GSM1891641) (Domcke *et al*, 2015), REST (GSE27148/GSM671093) (Arnold *et al*, 2013)

DNase hypersensitivity dataset was received from GEO under the accession number (GSE67867) (Domcke *et al*, 2015)

The Cage dataset was downloaded from FANTOM Consortium homepage (http://fantom.gsc.riken.jp) (Forrest *et al*, 2014)

References

Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LFA, van Lohuizen M & van Steensel B (2013) Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell* **154:** 914–927

van Arendsbergen J, Fitzpatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ & van Steensel B (2016) Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.*

Arnold P, Schöler A, Pachkov M, Balwierz P, Jørgensen H, Stadler MB, van Nimwegen E & Schübeler D (2013) Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res.***:** 60–73

Bardet AF, Steinmann J, Bafna S, Knoblich JA, Zeitlinger J & Stark A (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29:** 2705–2713

Barnes DE & Lindahl T (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.* **38:** 445–76

Baubec T, Ivánek R, Lienert F & Schübeler D (2013) Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. *Cell* **153:** 480–92

Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21:** 611–26

Bird A (2011) The dinucleotide CG as a genomic signalling module. *J. Mol. Biol.* **409:** 47–53

Bird A, Taggart M, Frommer M, Miller OJ & Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40:** 91–99

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8:** 1499–1504

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ & Klose RJ (2010) CpG islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell* **38:** 179–90

Boulard M, Edwards JR & Bestor TH (2015) FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat. Genet.* **47:** 1–9

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A & Cedar H (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435–8

Campanero MR, Armstrong MI & Flemington EK (2000) CpG methylation as a mechanism for the regulation of E2F activity. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 6481–6486

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia WJ, Anderson S, Yates J, Washburn MP & Workman JL (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123:** 581–592

Cedar H & Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10:** 295–304

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-H, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-K, et al (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–17

Clouaire T, Webb S & Bird A (2014) Cfp1 is required for gene expression dependent H3K4me3 and H3K9 acetylation in embryonic stem cells. *Genome Biol.* **15:** 451

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee JH, Skalnik

D & Bird A (2012) Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev.* **26:** 1714–1728

Cohen NM, Kenigsberg E & Tanay A (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145:** 773–86

Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev.* **25:** 1010–22

Domcke S, Bardet AF, Ginno PA, Hartl D, Burger L & Schübeler D (2015) Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528:** 575–579

Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I & Andrau J-C (2012) CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* **22:** 2399–408

Flanagan JF, Mi L-Z, Chruszcz M, Cymborowski M, Clines KL, Kim Y, Minor W, Rastinejad F & Khorasanizadeh S (2005) Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* **438:** 1181–5

Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy I V., Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, et al (2014) A promoter-level mammalian expression atlas. *Nature* **507:** 462–470

Fromm G & Bulge M (2009) A spectrum of gene regulatory phenomena at mammalian β-globin gene loci. *Biochem. Cell Biol.* **87:** 781–790

Gardiner-Garden M & Frommer M (1987) CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–282

Gontan C, Achame EM, Demmers J, Barakat TS, Rentmeester E, van IJcken W, Grootegoed JA & Gribnau J (2012) RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation. *Nature* **485:** 386–390

Guenther MG, Levine SS, Boyer LA, Jaenisch R & Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88

Han J, Yuan P, Yang H, Zhang J, Soh BS, Li P, Lim SL, Cao S, Tay J, Orlov YL, Lufkin T, Ng H-H, Tam W-L & Lim B (2010) Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature* **463:** 1096–1100

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H & Glass CK (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38:** 576–589

Iguchiariga SMM & Schaffner W (1989) CpG Methylation of the Camp-Responsive Enhancer Promoter Sequence TGACGTCA Abolishes Specific Factor Binding As Well As Transcriptional Activation. *Genes Dev.* **3:** 612–619

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N & Shendure J (2017) A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.***:** 38–52

Ioshikhes IP & Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **26:** 61–3

Jermann P, Hoerner L, Burger L & Schubeler D (2014) Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. *Proc. Natl. Acad. Sci.* **111:** E3415–E3421

Kelly TK, Liu Y, Lay FD, Liang G, Berman BP & Jones PA (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22:** 2497–506

Krebs A, Dessus-Babus S, Burger L & Schübeler D (2014) High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife* **3:** 1–18

Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L & Schübeler D (2017) Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell***:** 1–12

Kwasnieski JC, Mogno I, Myers CA, Corbo JC & Cohen BA (2012) Complex effects of nucleotide variants in a mammalian cis -regulatory element.

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

Larsen F, Gundersen G, Lopez R & Prydz H (1992) CpG islands as gene markers in the human genome. *Genomics* **13:** 1095–1107

Li B, Jackson J, Simon MD, Fleharty B, Gogol M, Seidel C, Workman JL & Shilatifard A (2009) Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s Histone deacetylase complex and to repress spurious transcription. *J. Biol. Chem.* **284:** 7970–7976

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F & Schübeler D (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **43:** 1091–7

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar a H, Thomson J a, Ren B & Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315–22

Long HK, King HW, Patient RK, Odom DT & Klose RJ (2016) Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res.* **44:** 6693–6706

Luo Z, Gao X, Lin C, Smith ER, Marshall SA, Swanson SK, Florens L, Washburn MP & Shilatifard A (2015) Zic2 is an enhancer-binding factor required for embryonic stem cell specification. *Mol. Cell* **57:** 685–694

Macleod D, Charlton J, Mullins J & Bird AP (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8:** 2282–2292

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp P a, Bartel DP, Jaenisch R & Young R a (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134:** 521–33

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, Kellis M, Lander ES & Mikkelsen TS (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30:** 271–7

Mogno I, Kwasnieski JC & Cohen BA (2013) Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.*

Mohn F & Schübeler D (2009) Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.* **25:** 129–36

Nabilsi NH, Deleyrolle LP, Darst RP, Riva A, Reynolds BA & Kladde MP (2014) Multiplex mapping of chromatin accessibility and DNA methylation within targeted single molecules identifies epigenetic heterogeneity in neural stem cells and glioblastoma. *Genome Res.* **24:** 329–339

Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S-P, Allis CD, Cheng X & Bestor TH (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448:** 714–717

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, Ahituv N, Pennacchio L a & Shendure J (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30:** 265–70

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D & Shendure J (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27:** 1173–5

Prendergast GC & Ziff EB (1991) Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science* **251:** 186–9

Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG & Corbo JC (2015) Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.***:** 1–18

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov V V & Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* **488:** 116–20

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK & Schübeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490–5

Strahl BD, Grant PA, Briggs SD, Sun Z-W, Bone JR, Caldwell JA, Mollah S, Cook RG, Shabanowitz J, Hunt DF & Allis CD (2002) Set2 Is a Nucleosomal Histone H3-Selective Methyltransferase That Mediates Transcriptional Repression. *Mol. Cell. Biol.* **22:** 1298–1306

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R & Bird A (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082–6

Tiwari VK, Stadler MB, Wirbelauer C, Paro R, Schübeler D & Beisel C (2011) A chromatin-modifying function of JNK during stem cell differentiation. *Nat. Genet.* **44:** 94–100

Vella P, Barozzi I, Cuomo A, Bonaldi T & Pasini D (2012) Yin Yang 1 extends the Myc-related transcription factors network in embryonic stem cells. *Nucleic Acids Res.* **40:** 3403–3418

Wachter E, Quante T, Merusi C, Arczewska A, Stewart F, Webb S & Bird A (2014) Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife* **3:** 1–16

Watt F & Molloy PL (1988) Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev.* **2:** 1136–1143

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M & Schübeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39:** 457–66

White MA, Myers CA, Corbo JC & Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis - regulatory function of ChIP-seq peaks.

Youdell ML, Kizer KO, Kisseleva-Romanova E, Fuchs SM, Duro E, Strahl BD & Mellor J (2008) Roles for Ctk1 and Spt6 in regulating the different methylation states of histone H3 lysine 36. *Mol. Cell. Biol.* **28:** 4915–4926

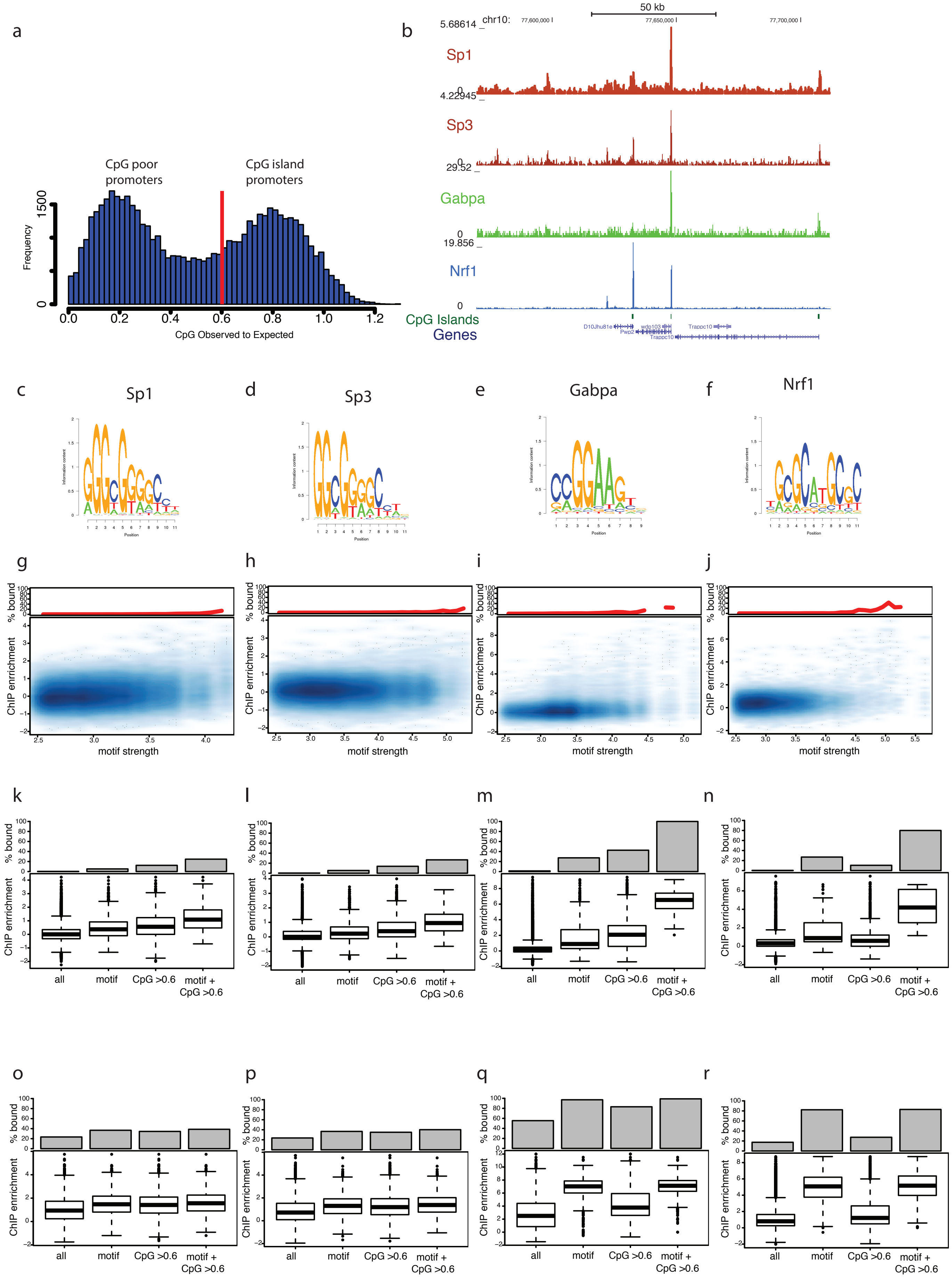

a
CpG poor promoters
CpG island promoters
Frequency
1500
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
CpG Observed to Expected
b
50 kb
chr10:
77,600,000
77,650,000
77,700,000
5.68614
Sp1
0
4.22945
Sp3
0
29.52
Gabpa
0
19.856
Nrf1
0
CpG Islands
Genes
D10Jhu81e
Pwp2
wdp103
Trappc10
Trappc10
c
Sp1
d
Sp3
e
Gabpa
f
Nrf1
Information content
Position
g
h
i
j
% bound
ChIP enrichment
motif strength
k
l
m
n
o
p
q
r
all
motif
CpG >0.6
motif + CpG >0.6

Figure 1: TF binding can be predicted by CpG density and TF motif occurence

(a) Histogram of CpG densities of all mouse promoters. CpG density is distributed in a bimodal fashion. 400bp upstream to 200bp downstream of TSS were defined as the promoter region. CpG density was calculated as Observed to Expected ratios (OE = number of CpGs / (number of Cs x number of Gs) x length of the region in nucleotides). The red line indicates the threshold in OE for CpG islands (Gardiner-Garden & Frommer, 1987).

(b) Browser screenshot of Sp1, Sp3, Gabpa and Nrf1 Rambio ChIP-seqs.

(c)-(f) Position weight matrices of Sp1 (c), Sp3 (d), Gabpa (e) and Nrf1 (f).

(g)-(j) Tested TFs are highly specific to their motifs. Scatterplots of motif strength versus ChIP enrichment (fold enrichment over input) in tiling 600bp windows across chromosome 19 (lower panel) and percent of windows that overlap with ChIP-seq peaks of corresponding TFs (upper panel) for each of the four TFs.

(k)-(n) High CpG density and TF motif occurrence together perform best for TF binding prediction. Barplots showing percent of windows that overlap with ChIP-seq peaks in the different categories (upper panel) and boxplots of ChIP enrichment (fold enrichment over input) versus sequence features in 600bp windows across chromosome 19 (lower panel) for each of the four TFs. 'All' displays data from all 600bp windows on chromosome 19, 'motif' depicts data from windows containing a motif that is > 90% of the maximum motif score, 'CpG > 0.6' shows data from windows having an OE of 0.6 or bigger and motif + CpG>0.6 shows data from windows containing a high scoring motif and high CpG density.

(o)-(r) Same as in (k)-(n) but within active promoters (TSS -400bp to +200bp).

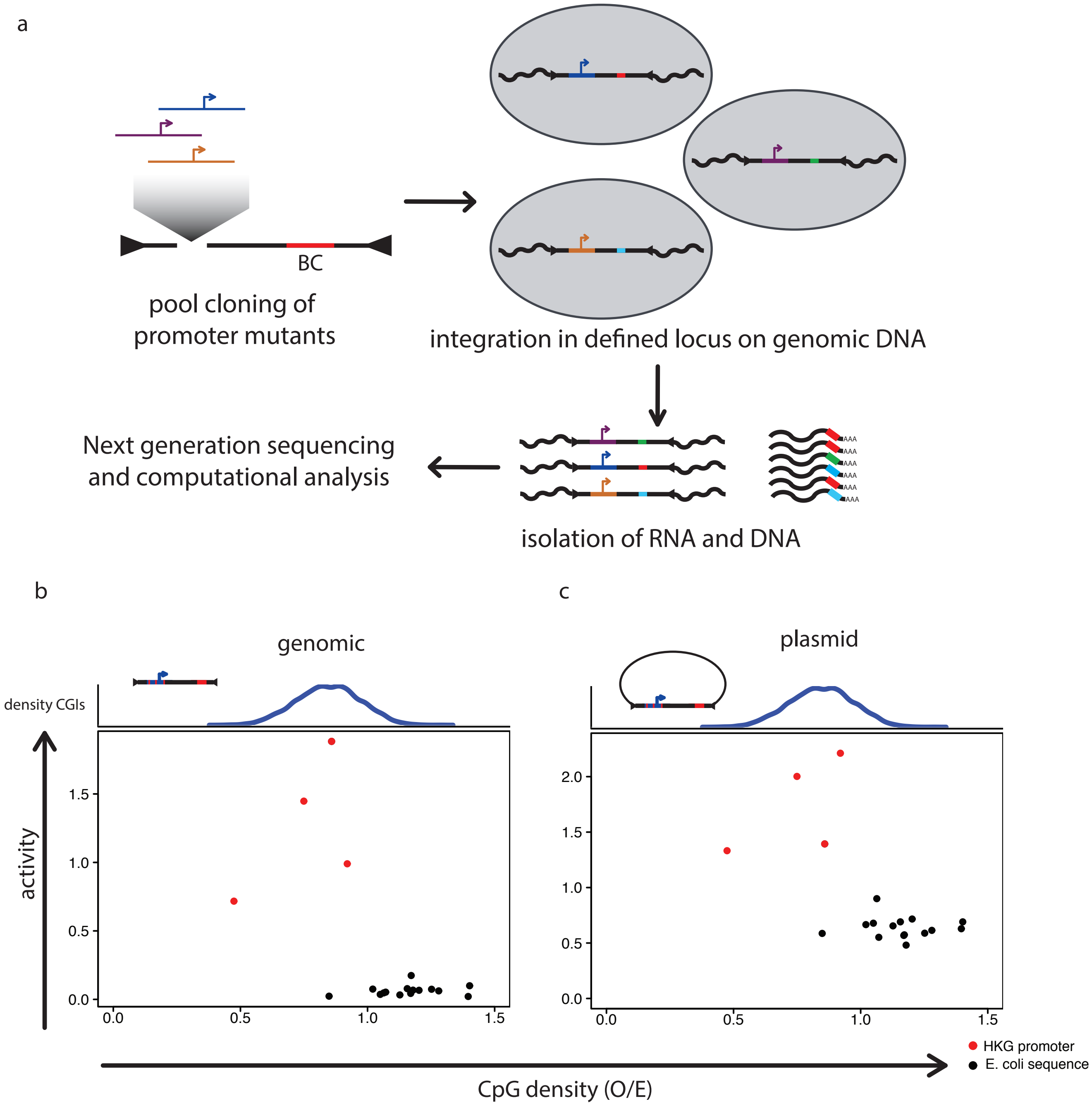


Figure 2: High CpG density alone is not sufficient for transcriptional activity

(a) Schematic representation of the procedure used to perform parallel reporter assays in a defined genomic locus. Promoter mutants are batch-cloned in front of GFP as a spacer sequence and an unique barcode. The expression cassette is flanked by loxP sites that allow integration into the β-globin locus of the used embryonic stem cell line replacing a selection cassette. After selection for cells containing the reporter construct DNA and RNA is isolated and the latter reverse transcribed. Barcodes are PCR amplified and sequenced using next-generation sequencing. Normalization of RNA barcode frequency to DNA barcode frequency results in transcriptional activity of the construct.

(b) A high CpG density alone is not sufficient for transcriptional activity. Scatterplot of CpG density versus transcriptional activity of sequences from the E. coli genome (black dots) and active housekeeping genes (red dots) on genomic DNA. The line above the scatter plot depicts CpG density distribution within promoters overlapping with CGIs.

(c) Same as in (b) but on plasmidic DNA.

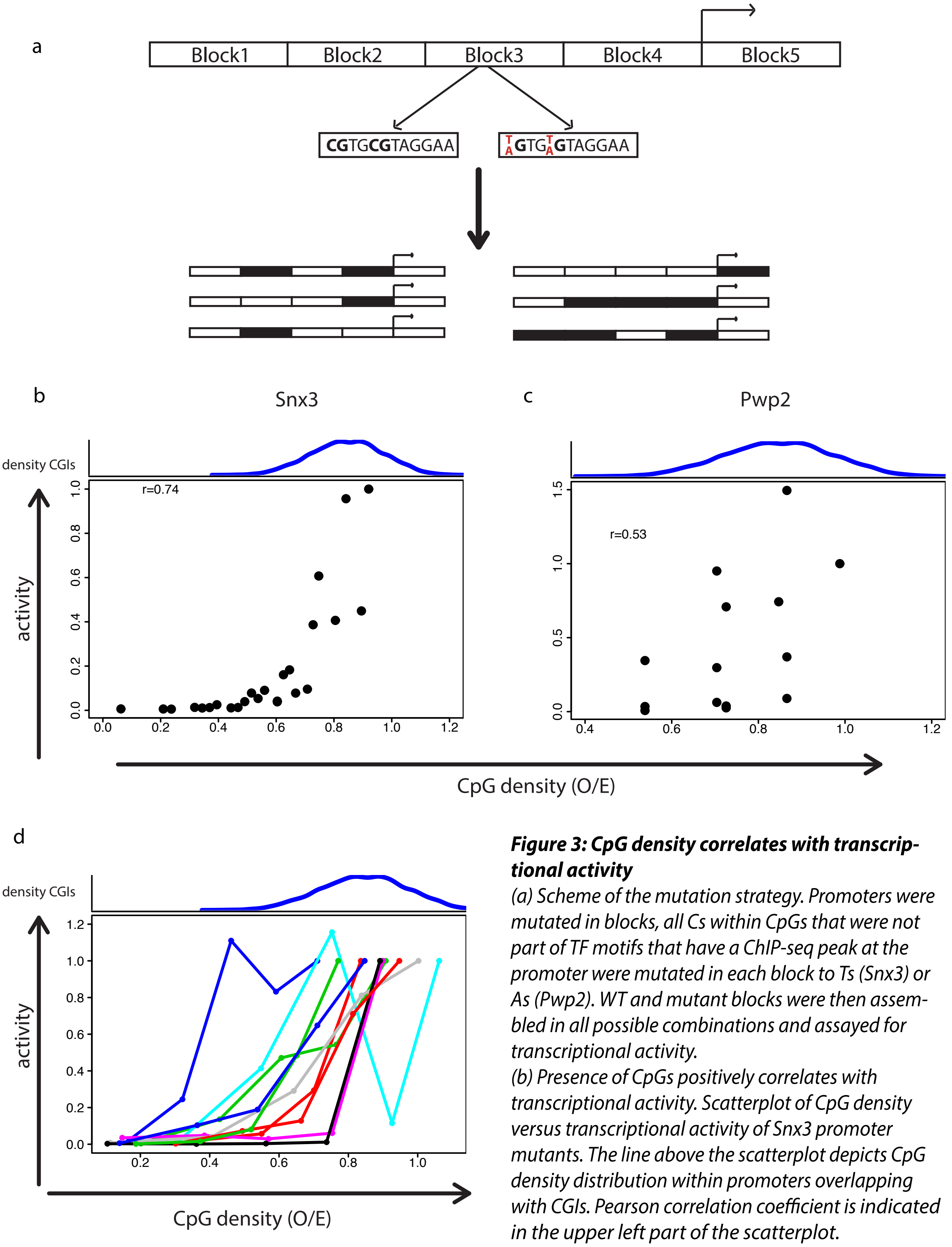


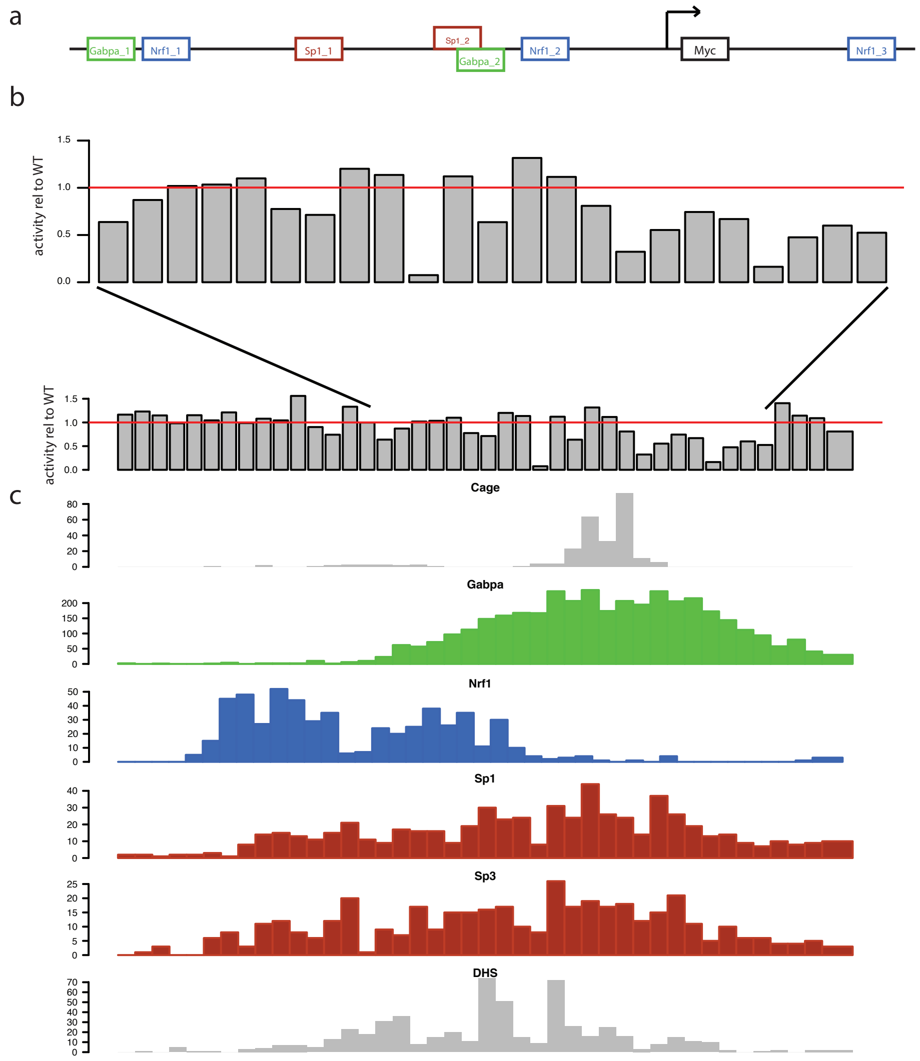
Figure 3: CpG density correlates with transcriptional activity

(a) Scheme of the mutation strategy. Promoters were mutated in blocks, all Cs within CpGs that were not part of TF motifs that have a ChIP-seq peak at the promoter were mutated in each block to Ts (Snx3) or As (Pwp2). WT and mutant blocks were then assembled in all possible combinations and assayed for transcriptional activity.

(b) Presence of CpGs positively correlates with transcriptional activity. Scatterplot of CpG density versus transcriptional activity of Snx3 promoter mutants. The line above the scatterplot depicts CpG density distribution within promoters overlapping with CGIs. Pearson correlation coefficient is indicated in the upper left part of the scatterplot.

(c) Same as in (b) for the Pwp2 promoter.

(d) Correlation of CpG density with transcriptional activity is a general feature. Scatterplot showing CpG density versus transcriptional activity in the reporter assay for 10 promoters. Mutants were generated by random mutation of Cs within CpGs to As if they were not part of TF motifs that have a ChIP-seq peak at the promoter. Different numbers of CpGs were mutated to generate five different CpG densities per promoter.



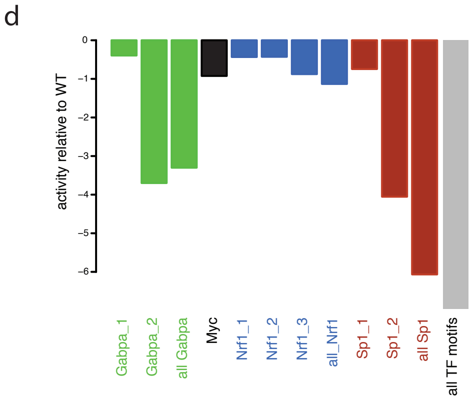


Figure 4: Characterization of the Pwp2 promoter

(a) Schematic view of a region of the Pwp2 promoter that contains TF motifs and the TSS.

(b) Mutation of 10 bp windows revealed highly variable effects on promoter function. Barplots showing transcriptional activity relative to the WT construct of promoters with mutated windows, in a zoom-in of the part of the promoter where TFs bind and the whole promoter respectively. Tiling 10 bp windows were mutated to a random CpG free sequence to assess contribution of each window to transcriptional activity.

(c) Genomic features overlap with regions important for transcriptional activity. Barplots displaying reads per 10 bp window for mRNA 5'ends (Cage), Gabpa, Nrf1, Sp1, Sp3 ChIP-seq and DNAseI hypersensitivity mapping at the endogenous Pwp2 promoter.

(d) Mutation of specific TF motifs leads to decreased transcriptional activity. Barplots showing log2 activity relative to WT of constructs with single TF motif mutations or mutations of all TF motifs of one TF.

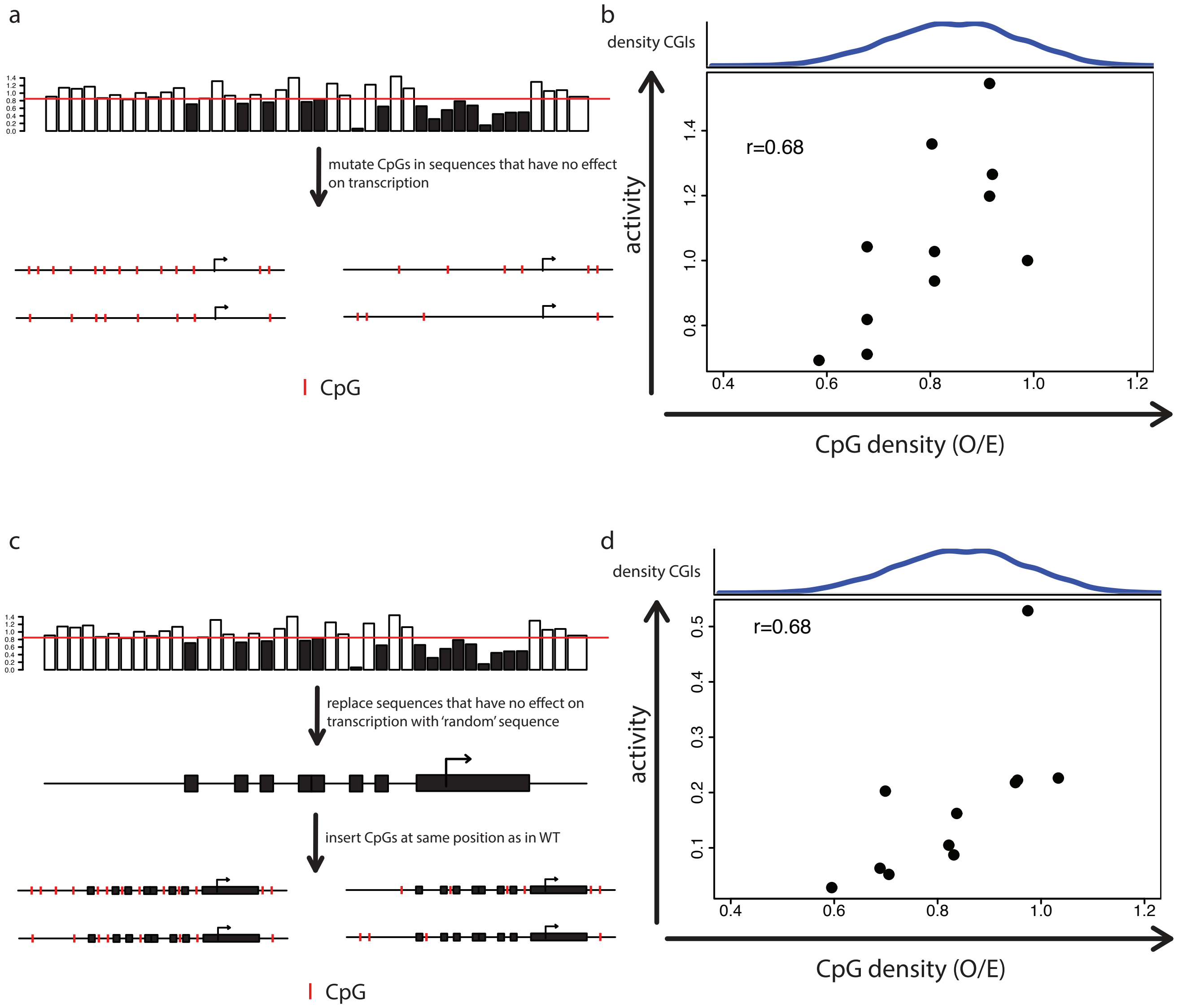


Figure 5: CpGs outside of TF motifs contribute to transcriptional activity of CGIs

(a) Scheme of the mutation strategy for CpG mutants. Cs in CpGs were mutated to As within 10bp windows that showed > 85% activity of WT activity when mutated. Different amounts of CpGs were mutated in different combinations

(b) CpGs outside of important regions in the Pwp2 promoter contribute to transcriptional activity. Scatterplot of CpG density versus transcriptional activity relative to WT Pwp2 for promoter mutants. CpG density correlates with transcriptional activity. Pearson correlation coefficient is indicated in the upper left part of the scatterplot.

(c) Scheme of the mutation strategy to generate an artificial promoter and strategy for adding CpGs. Sequences within windows that had > 85% activity of WT when mutated were replaced with a random CpG free sequence to generate an artificial chimeric promoter. Then different numbers of CpGs were introduced into the replaced sequence at the same location as in WT Pwp2.

(d) CpGs density itself does contribute to CGI activity. Scatterplot of CpG density versus transcriptional activity relative to the activity of WT Pwp2 for artificial promoters. CpG density correlates with transcriptional activity. Pearson correlation coefficient is indicated in the upper left part of the scatterplot.

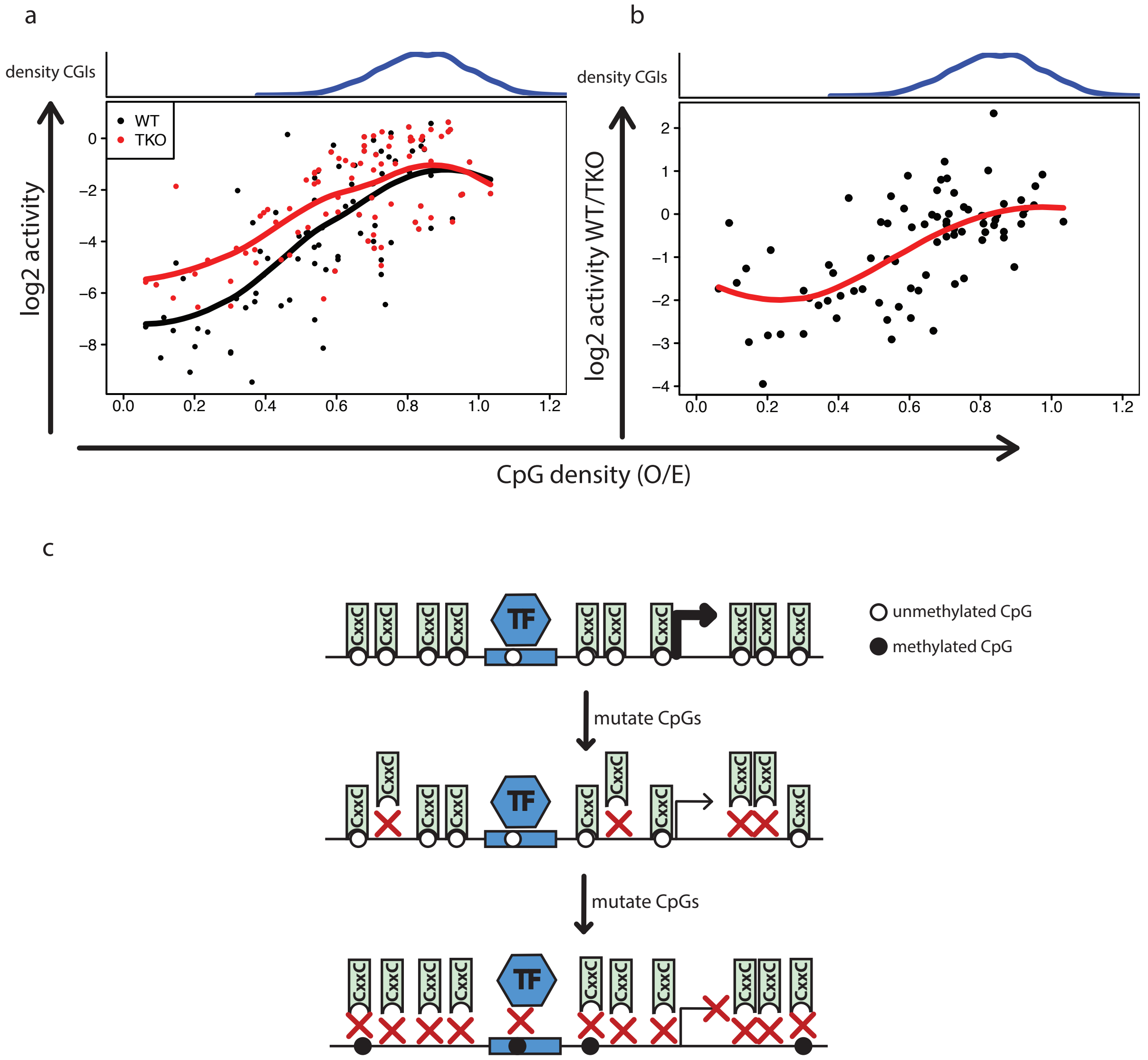
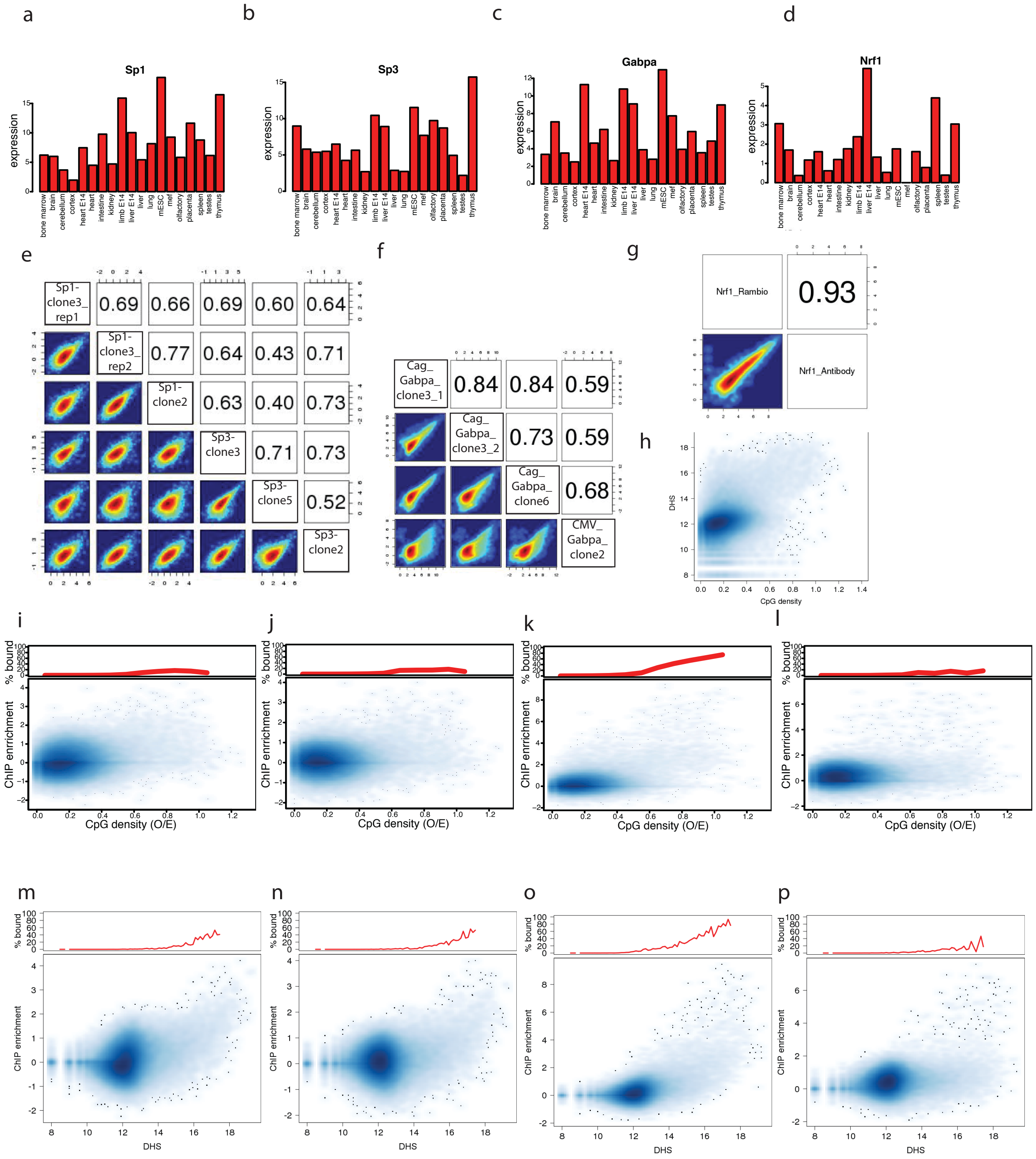


Figure 6: DNA methylation decreases transcriptional activity at mutants with low CpG density

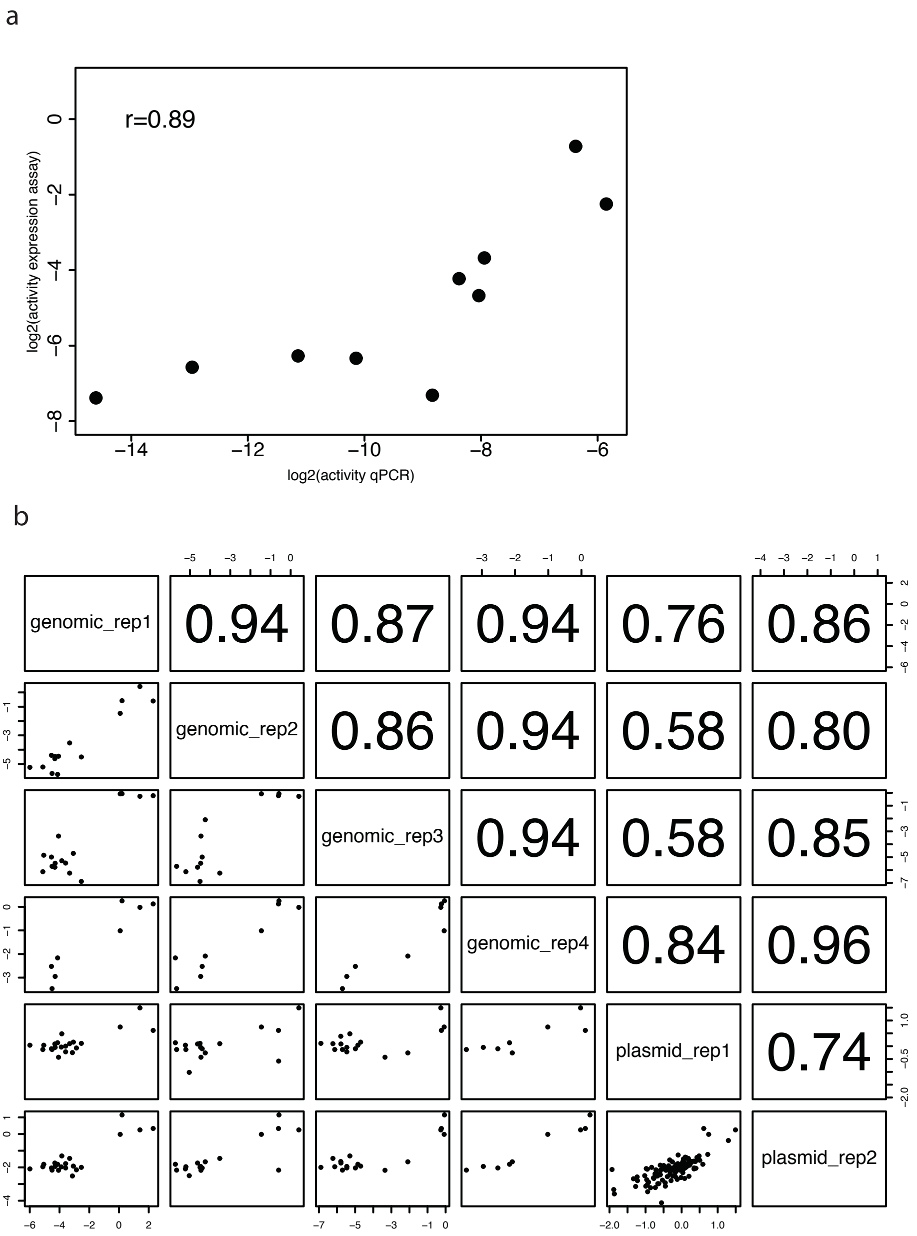
(a) Constructs are more active in TKO than in WT at lower CpG densities. Scatterplot showing CpG density of promoter mutants versus log2 activity relative to corresponding WTs. Libraries were assayed in WT (black) and methylation free mouse embryonic stem cells (TKO) (red).

(b) TKO activity deviates stronger from WT activity at lower CpG densities. Scatterplots showing CpG density versus log2 activity of WT/TKO.

(c)Model how decrease in activity upon CpG mutation could be explained. Mutation of CpGs leads to decreased activity, this could be due to decreased concentration of ZF-CxxC domain containing proteins at the promoter. Further decrease in CpG density leads to methylation which could inhibit TF binding resulting in very low to no transcriptional activity.



Supplementary Figure 1:
(a)-(d) Tested TFs are broadly expressed. Barplots of expression levels (reads per kilobase per million) of Sp1 (a), Sp3 (b), Gabpa (c) and Nrf1 (d) across 19 different tissues. All four factors are broadly expressed.
(e)-(f) Pairplots showing correlations of ChIP-seq samples.
(g) Pairplot showing correlation Biotin-tagged Nrf1 ChIP versus antibody Nrf1 ChIP. Antibody and Biotin ChIP-seq data is highly similar.
(h) CpG density correlates with accessibility. Scatterplot showing CpG density versus accessibility as measured by DNAse hypersensitivity.
(i)-(l) High CpG density correlates with TF binding. Scatterplots of CpG density versus ChIP enrichment over input in 600bp windows across chromosome 19 (lower panel) and percent of windows within each bin that overlap with ChIP-seq peaks of corresponding TFs (upper panel) for each of the four TFs.
(m)-(p) Accessibility correlates with TF binding. Scatterplots of DHS signal versus ChIP enrichment over input in 600bp windows across chromosome 19 (lower panel) and percent of windows within each bin that overlap with ChIP-seq peaks of corresponding TFs (upper panel) for each of the four TFs.

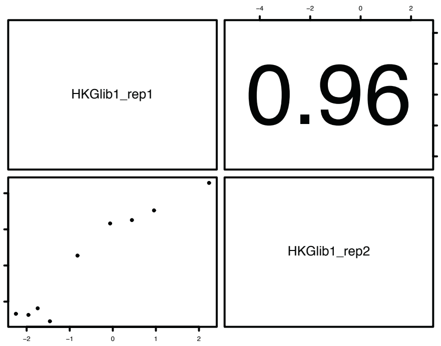


Supplementary Figure 2:

(a) Scatterplot displaying activity of promoter mutants measured by qPCR in single clones on cDNA with primers targeting the GFP transgene versus activity in the reporter assay. Spearman correlation coefficient indicates a high agreement between the two types of measurements (0.89).

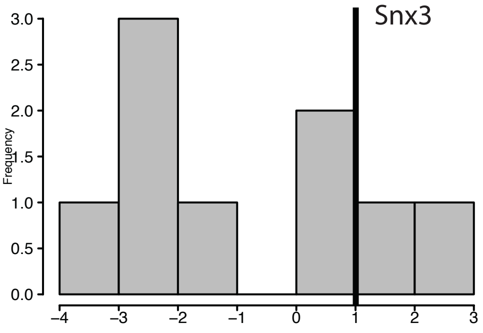
(b)Pairplot displaying correlation of replicates of libraries used in Figure 2.

a



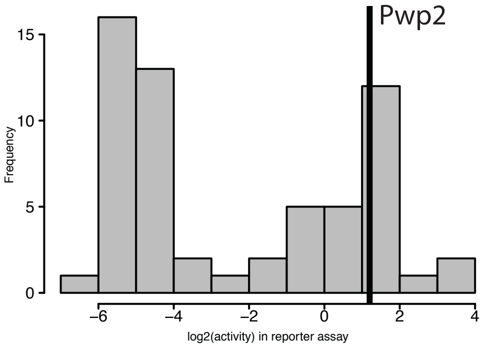
HKGlib1_rep1
0.96
HKGlib1_rep2

b



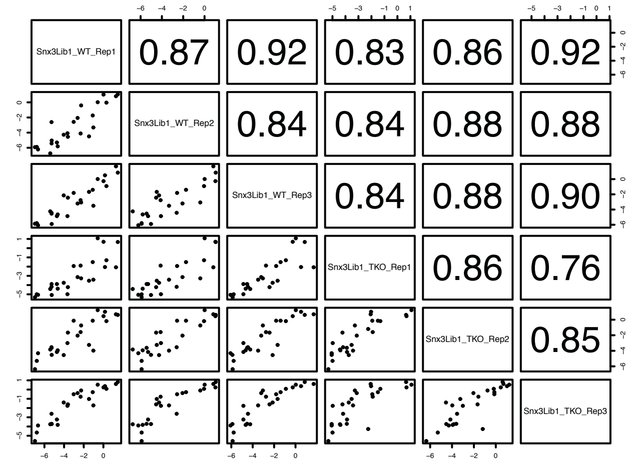
Snx3
Frequency
log2(activity) in reporter assay
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−4
−3
−2
−1
0
1
2
3

c



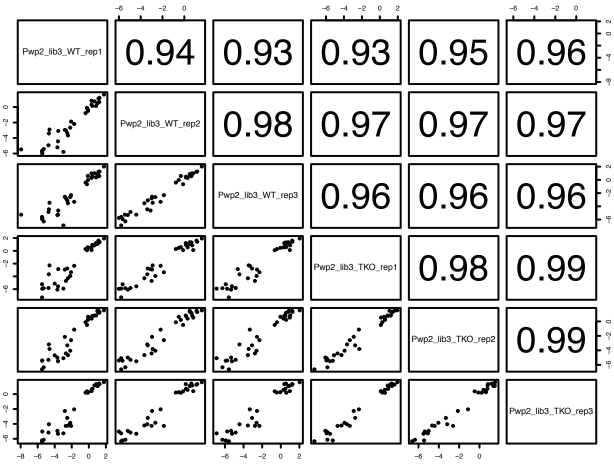
Pwp2
Frequency
log2(activity) in reporter assay
0
5
10
15
−6
−4
−2
0
2
4

d



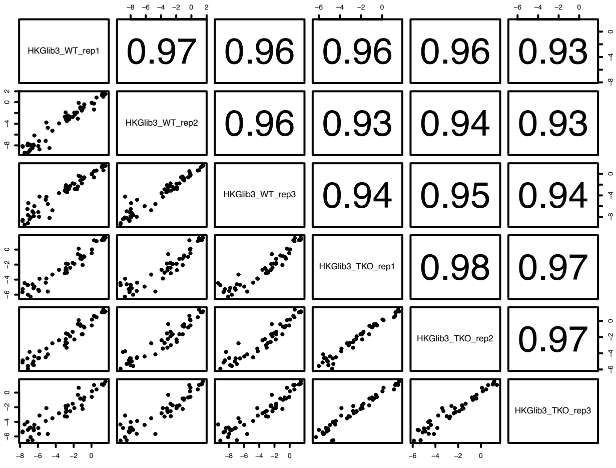
Snx3Lib1_WT_Rep1
0.87
0.92
0.83
0.86
0.92
Snx3Lib1_WT_Rep2
0.84
0.84
0.88
0.88
Snx3Lib1_WT_Rep3
0.84
0.88
0.90
Snx3Lib1_TKO_Rep1
0.86
0.76
Snx3Lib1_TKO_Rep2
0.85
Snx3Lib1_TKO_Rep3

e



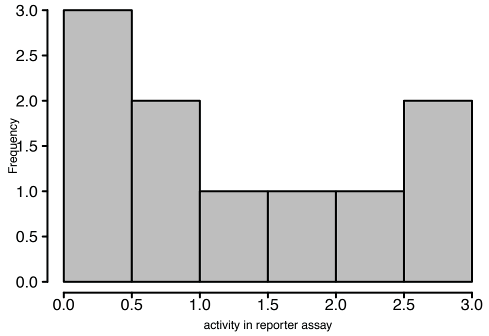
Pwp2_lib3_WT_rep1
0.94
0.93
0.93
0.95
0.96
Pwp2_lib3_WT_rep2
0.98
0.97
0.97
0.97
Pwp2_lib3_WT_rep3
0.96
0.96
0.96
Pwp2_lib3_TKO_rep1
0.98
0.99
Pwp2_lib3_TKO_rep2
0.99
Pwp2_lib3_TKO_rep3

f



HKGlib3_WT_rep1
0.97
0.96
0.96
0.96
0.93
HKGlib3_WT_rep2
0.96
0.93
0.94
0.93
HKGlib3_WT_rep3
0.94
0.95
0.94
HKGlib3_TKO_rep1
0.98
0.97
HKGlib3_TKO_rep2
0.97
HKGlib3_TKO_rep3

g



Frequency
activity in reporter assay
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Supplementary Figure 3:

(a) Pairplot displaying correlation of replicates of one of the libraries containing CGI promoters to identify candidates for mutation analysis.

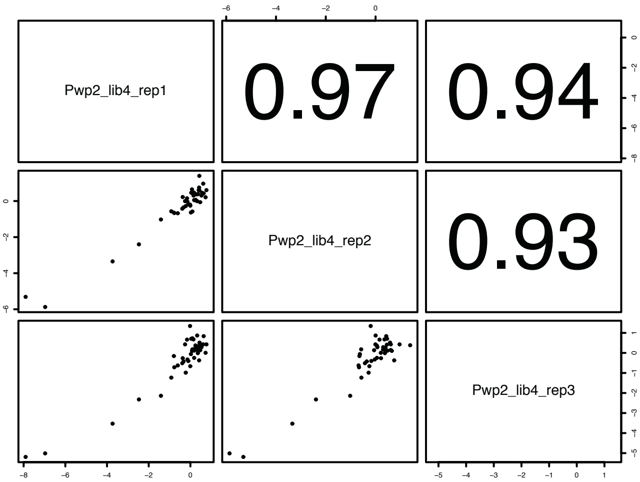
(b)-(c) Histograms of activity of two libraries containing CpG island promoters, activities of Snx3 (b) and Pwp2 (c) are indicated to compare to activities of all promoters in the library.

(d) Pairplot displaying correlations of replicates of Snx3 promoter CpG mutants.

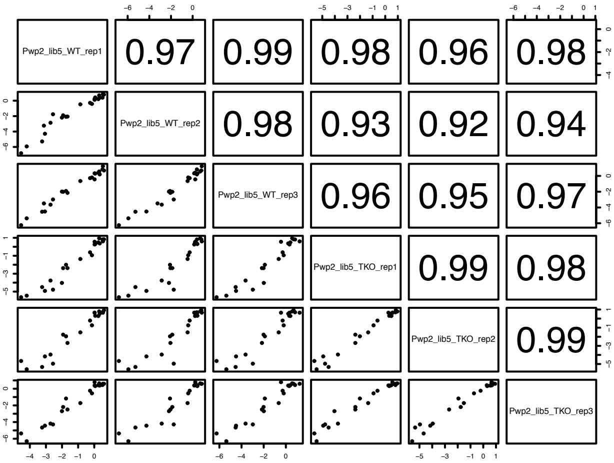
(e) Pairplot displaying correlations of replicates of Pwp2 promoter CpG and TF motif mutants.

(f) Pairplot displaying correlations of replicates of the library with CpG mutants of 10 different promoters.

(g)Histogram of WT activities of the 10 tested promoters.



Supplementary Figure 4:
Pairplot displaying correlations of replicates of library containing Pwp2 10bp window mutants.



Supplementary Figure 5:

Pairplot displaying correlations of replicates of library containing artificial promoters

Name	Sequence	
DH.P6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	BC amplification Primer
DH.P8	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 2 Index
DH.P26	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 6 Index
DH.P27	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 12 Index
DH.P30	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 4 Index
DH.P31	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 7 Index
DH.P32	CAAGCAGAAGACGGCATACGAGATGACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 5 Index
DH.P33	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 8 Index
DH.P34	CAAGCAGAAGACGGCATACGAGATGTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Indexing Primer NEB 9 Index
DH.P39	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCACTGGGAGAAGAGGAAGTCAAA	BC to Promoter amplification Primer

Table 1: Primers used for the high throughput reporter assay.

3.2 Cis-regulatory landscape of four cell types of the retina

Published manuscript

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx923

Cis-regulatory landscapes of four cell types of the retina

Dominik Hartl[1,2,†], Arnaud R. Krebs[1,*,†], Josephine Jüttner[1,†], Botond Roska[1,3,*] and Dirk Schübeler[1,2,*]

[1]Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH 4058 Basel, Switzerland, [2]University of Basel, Faculty of Sciences, Petersplatz 1, CH 4003 Basel, Switzerland and [3]University of Basel, Department of Ophthalmology, Mittlere Strasse 91, CH 4031 Basel, Switzerland

Received May 17, 2017; Revised July 28, 2017; Editorial Decision September 27, 2017; Accepted October 02, 2017

ABSTRACT

The retina is composed of ∼50 cell-types with specific functions for the process of vision. Identification of the *cis*-regulatory elements active in retinal cell-types is key to elucidate the networks controlling this diversity. Here, we combined transcriptome and epigenome profiling to map the regulatory landscape of four cell-types isolated from mouse retinas including rod and cone photoreceptors as well as rare inter-neuron populations such as horizontal and starburst amacrine cells. Integration of this information reveals sequence determinants and candidate transcription factors for controlling cellular specialization. Additionally, we refined parallel reporter assays to enable studying the transcriptional activity of large collection of sequences in individual cell-types isolated from a tissue. We provide proof of concept for this approach and its scalability by characterizing the transcriptional capacity of several hundred putative regulatory sequences within individual retinal cell-types. This generates a catalogue of *cis*-regulatory regions active in retinal cell types and we further demonstrate their utility as potential resource for cellular tagging and manipulation.

INTRODUCTION

The retina is a complex neural tissue within the eye. It is comprised of a large number of cell types with specialized functions that together enable visual perception (1,2). Retinal cell types include two types of image forming photoreceptors, rods and cones, horizontal cells as well as many types of bipolar, amacrine, ganglion and glial cells. Cone photoreceptors are active at higher light levels and mediate high-resolution color vision, while rod photoreceptors are active at low light condition. Signals sensed by photoreceptors are subsequently integrated and processed by interneurons such as, horizontal cells (HCs), bipolar cells and amacrine cells (i.e. starburst amacrine cells (SACs))(1) and finally converge to ganglion cells, the output neurons of the retina. Some retinal cell types can be uniquely identified by morphology and localization within the different retinal layers, while others require identification by genetic markers (3). Retinal function is affected in a multitude of genetic disorders, a number of which are cell type specific (4). These diseases lead to vision impairment or blindness (5–7).

In mammals, cellular identity is conferred by the activation and repression of specific gene expression programs. This process is principally controlled through the binding of transcription factors (TF) to distal *cis*-regulatory regions (CRE) named enhancers. Enhancer activity is highly variable across cell types in line with the concept that these elements are central in regulating cell-type specific gene expression (8–11). Systematic mapping of active enhancers based on their characteristic chromatin states has provided a large catalog of putative regulatory regions in a plethora of tissues and cell lines (12–14). In turn, cataloging CREs at the resolution of individual cell types is a prerequisite for understanding the transcriptional regulatory principles that controls cell-type specification within a tissue. Activation of regulatory regions entails binding of transcription factors and coinciding changes in chromatin. This includes increased accessibility (15), specific histone modifications (11,16) and locally reduced DNA methylation (17–20). Histone modification such as acetylation is a direct reflection of transcriptional co-activator activities. It appeared to be one of the best predictor for CRE activity (21) and has been successfully used to identify active CREs in various tissues (22,23). Chromatin accessibility indirectly reflects TF binding activity and thus represents another feature that has been successfully used to identify active CREs in the genome (13,21,24,25). Additionally, in most tested tissues

---

[*] To whom correspondence should be addressed. Tel: +41 61 69 78270; Fax: +41 61 69 73976; Email: arnaud.krebs@fmi.ch; arnaud.krebs@embl.de
Correspondence may also be addressed to Botond Roska. Email: botond.roska@fmi.ch
Correspondence may also be addressed to Dirk Schübeler. Email: dirk@fmi.ch
[†] These authors contributed equally to this work as first authors.

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

and cell types DNA methylation and accessibility appeared to be tightly anti-correlated (17–20), making low methylation regions (LMRs) a useful proxy to detect active CREs. A notable exception to this rule was found in rod photoreceptors, where a fraction of low methylated regions appeared to be located within a closed chromatin environment (26). This unusual feature adds to the notion of a unique chromatin organization in rods, that has been associated with its cellular function (27).

While useful to identify putative CREs, chromatin based predictions do not inform on functional relevance nor on the ability of putative sequences to efficiently drive transcription in an ectopic context. Indeed a large fraction of putative enhancers fail to drive detectable expression levels when tested in an ectopic context irrespective of the feature used for their identification (21,25,28–31). Enhancers vary largely in size, yet even within very large regions such as the beta globin LCR most activation is conferred by smaller sub-fragments (32), which have proven powerful tools for ectopic gene expression. Thus testing the activity of isolated DNA pieces using transcriptional reporter systems not only informs on the features of *cis*-regulatory regions able to drive transcriptional activity in different cell types but also identifies potent elements for transgenic gene expression. Using high-throughput sequencing, such assays have recently been parallelized (23,33–38), which enables to quantify the ectopic activity of thousands of DNA sequence variants in a single experiment (39). Yet *in vivo* such approach was only applied at the resolution of entire tissues (23,38,40), which ignores cellular heterogeneity and thus lacks the resolution required to understand transcriptional regulation at the level of single cell-types.

To characterize the *cis*-regulatory landscape of the retina, we generated expression profiles and genome-wide DNA methylation maps for four cell types isolated from mouse retinas: cones, rods, HCs and SACs. These datasets identify large collections of putative CREs in each cell type, revealing sequence determinants and transcription factors (TF) potentially involved in the control of identity of these cells. To enable efficient characterization of the activity of the identified CREs, we adapted the principles of parallel reporter assays to *in vivo* measures at the resolution of single cell-types. As a proof of concept, we measured the activity of hundreds of CREs and defined their activity profile in the four individual retinal cell types. Additionally, we generated libraries of sequence mutants to probe the functional contribution at sequence level of enriched TF motifs. This revealed the co-existence in *cis* of active and repressive signals at highly active photoreceptor CREs and demonstrates how rational CRE editing can be used to modulate transgene expression levels in a desired cell type.

MATERIALS AND METHODS

Animal handling

All animal experiments and procedures were approved by the Swiss Veterinary Office. Cell type-specific Cre recombinase driver lines: D4-cre (41) for cones, B2-cre (42) for rods, Gja10-cre (4) for HCs and ChAT-cre (Jackson, stock: #006410) for SACs; were in-house crossed to the floxed tdTomato reporter line Ai9 (JAX mice B6.Cg-Gt(ROSA)26Sortm9(CAGtdTomato) Hze/J, Jackson stock: #007909) to generate retinas with one cell type fluorescently labelled. The age of mice was between 50 days and 150 days, sexes were all female for RNA-seq and WGBS and chosen randomly for PRA. Adult wild-type mice (C57BL/6) purchased from Charles River were used for single enhancer testing experiments.

RNA-seq library preparation and sequencing

After retina dissection and dissociation, cells were FACS-sorted directly in lysis buffer of the RNA-easy mini kit (Quiagen) that was used for RNA extraction. RNA-seq libraries were prepared using the Norgen single cell RNA-seq preparation kit (51 800). Each of the three biological replicates were prepared using independent sorts on individual retinas. The samples were run on an Illumina HiSeq2500 generating 50 bp single-end reads.

WGBS library preparation and sequencing

DNA was extracted from cells sorted from single retinas. 50–100 ng of DNA was used as an input for bisulfite conversion (Zymo Gold Kit). The converted DNA was used to prepare whole genome bisulfite libraries using Illumina Truseq DNA methylation preparation kit (EGMK81312) following manufacturer recommendation. PCR product was purified using AMPureXP beads (Beckman Coulter—A63880) and controlled on Bioanalyser High sensitivity (Agilent 5067-4626). The samples were run on an Illumina HiSeq2500 generating 100 bp paired-end reads (rapid-run).

Library generation

Fragments were PCR amplified in 384-well format using Phusion Hot Start II polymerase (Thermo Scientific, #F-549S), pooled, gel purified and cloned blunt ended using an EcoRV site into a vector containing the expression cassette. The expression cassette consists of a multiple cloning site, and a random 15 bp barcode sequence (NNNNWNNNNWNNNNN) and a polyA signaling sequence (pA). In order to average out the contribution of barcode specific biases to the signal we aimed for at least ten different barcodes per unique fragment. To link CREs to barcodes the CRE-barcode sequences were amplified using Primer #2 (see Supplementary for sequences) and one of the Indexing primers (Primers #3–11) containing the Illumina flow cell annealing sequences. PCR products were purified using AmPure XP beads (Beckman Coulter, #A63880). PCR products were directly sequenced using MiSeq 500 or 600 cycle Kits. Next the vector was cut with SphI and PacI and a sequence containing a 31bp minimal promoter, CpG free eGFP and the annealing sequence for Primer #1 was cloned in (Supplementary Figure S3A). This construct was cut out of the cloning vector using NotI and inserted into the AAV vector.

AAV production

AAV production was performed as previously described (43). Briefly, HEK293T cells were transfected with a plasmid containing the transgene between the internal terminal

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

repeats of AAV2, the AAV-helper plasmid encoding Rep2 and Cap for serotype 8, and the pHGTI-Adeno1 plasmid harboring helper adenoviral genes (both kindly provided by C. Cepko, Harvard Medical School, Boston, MA, USA) using polyethylenimine (Polysciences, no. 23966). Vectors were purified by iodixanol gradient (Sigma, Optiprep). Genome titer (genome copies/ml) of AAV vectors were determined by real-time PCR using TaqMan primer/probe set corresponding to the WPRE (Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element) region of the vector and linearized plasmid standards. Titers were between 1×10^{14} and 5×10^{14} GC/ml for viral enhancer libraries and between 7×10^{11} and 6×10^{12} GC/ml for individual enhancer validation experiments.

Subretinal AAV delivery

Viral particles were injected as previously described (44). Briefly, animals were anesthetized using 3% isoflurane. A small incision was made with a sharp 30-gauge needle in the sclera near the lens. 2 μl of AAV suspension was injected through this incision into the subretinal space using a blunt 5μl Hamilton syringe held in a micromanipulator.

Dissociation of retina and fluorescence-activated cell sorting (FACS)

Biological triplicates of one genotype were always done in the morning at the same time and with maximum two hours delay between the first and the last sample. Retinas were isolated and dissociated to single cells by papain digestion as previously described (45). For rare cell types (HCs, SACs) retinas from both eyes were pooled to have enough material. Cells positive for tdTomato were sorted by FACS (BD FACS Aria III (Becton Dickinson) using a 100-μm nozzle with the bandpass filter for RFP HQ616/26. Cells were gated based on their forward- and sideward-scatter. Pulse-width was used to exclude doublets. Fluorescence positive cells were sorted at RT into a low binding tube (Eppendorf) containing 350 μl RLT extraction buffer (RNAeasy, Qiagen) for PRA, 250 μl lysis buffer (50 mM Tris, 10mM EDTA, 4% SDS, adding 10 μl Proteinase K) for WGBS and 300 μl RL buffer (Norgen) for RNA seq. Collected cells were immediately processed or stored at –80°C.

Sample preparation for PRA

After injection, RNA was isolated from sorted cells of the three biological replicates using independent sorts on individual retinas with Quiagen RNeasy® Mini Kit with on-column DNase digestion. After reverse transcription using Takara PrimeScript RT Reagent Kit (#RR047A) barcodes were amplified with KAPA HIFI Hotstart using Primer #1 and indexing primers (Primers #3–11). Since isolation of the AAV DNA from the cellular material turned out to be difficult we isolated DNA from the input AAV for normalization of barcode abundance. DNA from AAV was isolated and barcodes amplified as in the cDNA samples. PCR products were purified using AmPure XP beads (Beckman Coulter, #A63880) and sequenced using 50 cycle Kit on HiSeq 2500.

Adeno-associated virus (AAV) construct for individual enhancer validation

For individual enhancer testing, the sequence of interest was PCR amplified from genomic mouse DNA (129S5) and inserted in front of a minimal promoter (pDis1.2). Clone orientation was determined by sanger sequencing. The CRE-minimal promoter cassette was PCR amplified with MluI and BclI flanking restriction sites and inserted into pAAV2-EF1a-ChR2-EGFP via the same restriction enzymes.

Immunohistochemistry

For individual enhancer testing, mice were euthanized by CO_2 and rapid cervical dislocation 3 weeks post AAV injections. Retinas were dissected from the eyecup and fixed for 20–30 min in 4% paraformaldehyde (PFA) (wt/vol) in PBS and washed overnight in PBS. To aid penetration of the antibodies, retinas were frozen and thawed three times after cryoprotection with 30% (wt/vol) sucrose. The retina was incubated in blocking solution: 10% donkey serum (vol/vol), Millipore, 1% bovine serum albumin (wt/vol) and 0.5% Triton X-100 (vol/vol, in PBS, pH 7.4) for 1 h. Primary and secondary antibody applications were done in 3% normal donkey serum, 1% bovine serum albumin, 0.02% sodium acid (wt/vol) and 0.5% Triton X-100 in PBS. Primary antibodies were applied for 3–7 days. After washing the retina three times for at least 10 minutes in PBS, the retina was incubated in fluorescence-conjugated secondary antibodies and 10 μg/μl Hoechst 33342, trihydrochloride, trihydrate at a dilution of 1:200 for 2 h, followed by three washes in PBS, and mounting on slides with ProLong Gold antifade reagent (Molecular Probes). Retinas for vibratome section were embedded in 3% agarose (wt/vol) (SeaKem Le Agarose, Lonza) in PBS, and 150μm vertical sections were cut with a Leica VT1000S vibratome. Antibody staining procedure was the same as in whole mounts. The following primary antibodies were used: rat anti-GFP (1:500; Nacalai/Brunschwig), rabbit anti-mouse cone arrestin, mCAR (1:200; Millipore). For the secondary antibodies, we used in donkey serum raised antibodies from Invitrogen (Alexa Fluor 488, Alexa Fluor 555, Alexa Fluor 633).

Microscopy

Zeiss LSM700 laser scanning confocal microscope was used to acquire images of antibody-stained retinas with an EC Plan-Neofluar 40×/1.30 oil M27 and a Plan-Acro Achromat 10×/0.45 objectives at three excitation laser lines (405 nm for Hoechst, 488 nm for GFP, 555 nm for mCAR). Morphologies of cell types were assessed from 512 × 512 pixel images in a z-stack with 0.85 μm z-steps. Images were processed using Imaris (Bitplane).

Bioinformatics procedures

All analyses were performed using R-Bioconductor. *Ad hoc* R scripts are available upon request.

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

WGBS alignment and data extraction

Methylation data processing. Raw sequence files were pre-processed using Trimmomatic (46) to remove Illumina adaptor sequences, discard low quality reads and trim low quality bases. The trimmed reads were then aligned using QuasR (using Bowtie as an aligner) (47,48) against a bisulfite index of the Mus Musculus genome (BSgenome.Mmusculus.UCSC.mm9). CpG methylation call was performed using QuasR. Conversion rates were determined (and controlled to be >95%) by calling methylation of mitochondrial DNA and non-CG context Cs. Methylation was called genome wide for CpGs covered at least 8 times. Since most datasets arise from female mice, sex chromosomes were excluded from the methylation analysis. Genomic tracks were obtained by smoothing data using a sliding window over 10 CGs.

Identification of putative CREs using genome segmentation. Methylation data from each cell type was used to segment the genome using MethylSeeker (49) to identify regions containing at least four consecutive CGs below 50% methylation (False Discovery Rate $< 5\%$ in all samples). The total set of CREs ($n = 104\,322$) was defined by merging low methylated regions (LMRs) smaller than 2000 bp from the four retinal cell types. CREs were classified in the (2^4) 16 possible combinations based on the average CRE methylation in each cell type (considering methylation $< 60\%$ as positive).

Generation of CRE libraries

Putative enhancer regions were defined based on LMR definition. Putative specific sequences were nominated by comparing methylation in regions that are LMRs in one but not the other tested cell types. Since any application would require that functional elements are compact we limited fragment sizes of all libraries to maximum ∼700 bp (Supplementary Figure S3C). Tested regions were mostly distal (Supplementary Figure S3B). Primers were batch designed using a custom R function based on Primer3.

We generated four libraries. Library 1 was designed to test the system and contained regions hypomethylated in rods independent of methylation in other cell types. Library 2 was designed to find rod or cone specific elements, therefore we chose sequences displaying differential methylation in rods and cones. To extend our approach to more cell types we generated a library with sequences hypomethylated in HCs and/or SACs in library 3. All three libraries contained a set of verified sequences with different activities discovered in Library 1. To test contributions of TF motifs to enhancer activity we generated Library 4 in which different motifs of two enhancers were mutated. For the pilot library in rods (Library 1), we randomly selected a subset of the putative rod CREs (LMRs). We also included as negative controls a set of fully methylated regions in rods (>75%) without DHS signal in whole retina (24). The Cone/Rod library (Library 2) was designed based on differential methylation between cones and rods. We required cone specific regions to be at least 75% methylated in rods and less than 75% methylated in cones and vice versa. Additionally we included regions that are similarly methylated (>75%) or unmethylated (<75%) to serve as controls to compare measurements as well as the CREs from the validation experiments. The SAC/HC library (Library 3) was designed using LMRs that are hypomethylated in only one cell type and at least 80% methylated in the other three cell types. Additionally the library contained the same set of negative controls negative controls and regions unmethylated in HC and SAC for normalization (<50% in both cell types) as well as the CREs from the validation experiments.

For the mutant library. Motifs present within fragments were identified by scanning known TF position weight matrices (50,51; score > 9). Motifs matches were then randomized. If multiple motifs for one TF were present within a fragment, all motifs instances were mutated together. The CRE mutants were synthesized (gblocks-IDT technologies, for sequences see Supplementary Information) and cloned into the PRA vector.

RNA-seq alignment and data extraction

Raw sequence files were aligned using QuasR (using Bowtie as an aligner; 47,48) against the Mus Musculus genome (BSgenome.Mmusculus.UCSC.mm9). Reads in genes were collected using QuasR based on UCSC transcript annotation for mm9 and RPKM were calculated.

Motif enrichment analysis

For motif analysis on methylome based classification: Putative CREs classified based on their average DNA methylation (see above) were used as an input for motif enrichment analysis using HOMER (52). A set of known matrices was created by combining JASPAR (50) with more recent SELEX datasets (51). Motif enrichment was calculated separately for each factor in each set using a randomized set of sequences with similar base composition as a background. Motifs were filtered for factors expressed in the studied cell types ($\log_2$(RPKM) > –2 in at least one cell type). Enrichments for the most significantly motifs (enrichment over background >3-fold; P-value < 0.01) in the cell type specific categories were displayed.

For motif analysis PRA based classification: Motif enrichment analysis was adapted to account for the small size of the sequences set. The same set of motifs was screened over each subset of sequences and a set of control sequences. Differences in motif occurrence frequency between the foreground and background set was calculated. Enrichment in the motif frequency was plotted for the most significantly enriched motifs.

PRA analysis

Assignment of the barcode with the CRE. Sequences from Read1 starting with 'TCCACTGGGAGAAGAGGAAG TCAAA' were aligned from position 55 on to the Mus Musculus genome (BSgenome.Mmusculus.UCSC.mm9) using Bowtie.

Sequences from Read2 between 'CGTTTAAACTGTCG ACCGAGCT' and 'TTCGGCGCATG' were extracted as barcodes and the reverse complement was generated. Barcodes and aligned reads were matched using their read IDs.

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

Barcodes were associated with a CRE if >90% of the reads linked to the barcode corresponded to this CRE.

Calculation of CRE transcriptional activity. Analysis was performed on biological triplicates. Barcode sequences were extracted from 50 bp reads by taking only reads starting with the expected backbone sequence: 'TCCTGCTG GAGTTCGTGACCTGCATGCGCCGAA'. From these reads the sequence at position 34–48 was extracted. The frequency of each barcode sequence was calculated to get counts for each sample. Counts of barcodes were normalized to library size. Enrichment of barcodes in the RNA sample was calculated over their representation in the AAV input. Barcodes not sufficiently covered in the AAV sequencing were discarded (2–16 reads, depending on the sequencing depth). The median activity of all barcodes per CRE was calculated. Only CREs that were covered in at least two out of three biological replicates with at least three barcodes were used for downstream analysis.

RESULTS

Transcriptome and epigenome of four cell types of the retina

The *cis*-regulatory landscape associated with the cellular diversity of the retina has only been partially characterized through chromatin measures at the level of the entire tissue (24) and for two types of photoreceptors (26,53). Performing similar experiments is considerably more challenging for most cell-types of the retina since these can be significantly less abundant (2). For instance, while rods make up 65% of the mouse retina, horizontal cells only represent 0.5% of the retina (2) which corresponds to <10 000 cells per retina.

Unlike chromatin associated marks, robust DNA methylation measures can readily be achieved from a low number of cells (<5000 cells) and furthermore do not require the preservation of cellular or nuclear integrity during the isolation of cells from complex tissues. Therefore, we took advantage of the fact that in most cell types low levels of DNA methylation indicate accessibility of putative CRE regions (17–20,49) to expand our understanding of the complexity of the *cis*-regulatory landscape of the retina. We measured gene expression and DNA methylation genome-wide for four cell-types; cones and rods representing two types of photoreceptors; starburst amacrine cells and horizontal cells representing two types of interneurons.

We made use of our previously generated library of transgenic mice that express florescent proteins in specific cell types (3,4) (Supplementary Table S1). Labeled cells were FACS-sorted followed by whole genome bisulfite sequencing (WGBS) and matching sequencing of RNA (RNA-seq) (Figure 1A). The purity of cell isolates was confirmed by their expression of established marker genes (Figure 1B) and the robustness of the procedure is reflected in the high reproducibility of genome-wide expression patterns (R>0.96 - Supplementary Figure S1A). Due to the profusion of rods in the photoreceptor layer of the mouse retina (2,54), we observed a systematic contamination of our cone sorted isolates with ∼10% rods (as observed previously (4,26)). This leads to a systematic underestimation of the rod-specific signal, which is particularly marked at the level of gene expression (Figure 1B). It is however much less pronounced in the methylation profiles since unlike RNA molecules, DNA molecules are directly proportional to the number of contaminating cells (Supplementary Figure S2B).

Despite this technical limitation, we could unambiguously identify a unique transcriptional signature defining each of the tested cell types (Figure 1B). When clustering cells by proximity of their transcriptomes, we observed a grouping by cellular subtypes, with the highest similarity observed between HCs and SACs (Figure 1C, Supplementary Figure S1A and B). In an attempt to list potential regulators of this diversity, we extracted the transcription factors showing the most transcriptional divergence between cell types (Figure 1D). We observed many TFs having differential expression between photoreceptors and interneurons, including many of the previously described regulators of these linages (Figure 1D; 4,55,56). For instance expression of Otx2, Nrl or Rax as well as several members of the nuclear receptor family appear as a clear signature of photoreceptors. Similarly, interneurons are characterized by high expression of neuronal fate markers such as Pax6. Additionally, each cell type shows unique TF expression signatures. For example rods show elevated levels of Essrb, Nr2e3, Olig3 while cones show preferential expression of RxR gamma, NfKB, Elk3. Similarly, HCs uniquely express several members of the Onecut TF family while SACs display high levels of Pou6f2, Tbx2 or Ap-2 (tfap2b) (Figure 1D). In summary, we define here a catalog of differentially expressed TFs in the four studied cell types that may play a role in regulating their unique transcriptional signature.

Defining the *cis*-regulatory landscapes of four cell types of the retina

We then used our genome-wide maps of DNA methylation to define a catalog of CREs putatively active in the four studied cell types. For each cell type including rare population of interneurons, we could determine the methylation level for ∼70% of the 20.3 million autosomal CpGs (≥64% in all samples – coverage ≥ 8×). We note that our data from rods and cones agree very well with recently published methylation datasets that were generated using a different isolation strategy (26) (Figure 2- Supplementary Figure S2A). To identify regulatory regions across the tested retinal cell types, we applied our established segmentation approach (18,49). We identified between 80–100 thousand LMRs in each cell type as putative CREs of the retinal system.

Since methylation differences at CREs is an indirect indication of regulatory activity, we compared the methylation status of these regions between cell types (Figure 2A, B, Supplementary Figure S2C-H). We observe only subtle differences in the DNA methylation levels at these regions between cones and rods (Figure 2A, Supplementary Figure S2C). However, HCs and to an even higher extend SACs show substantial differences in their methylation levels at LMRs (Figure 2A, B, Supplementary Figure S2D–G) in agreement with the functional divergence between photoreceptors and interneurons. Interestingly, contrasting our data with previously published methylation datasets from purified neurons (20) reveals a closer proximity of SACs but

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

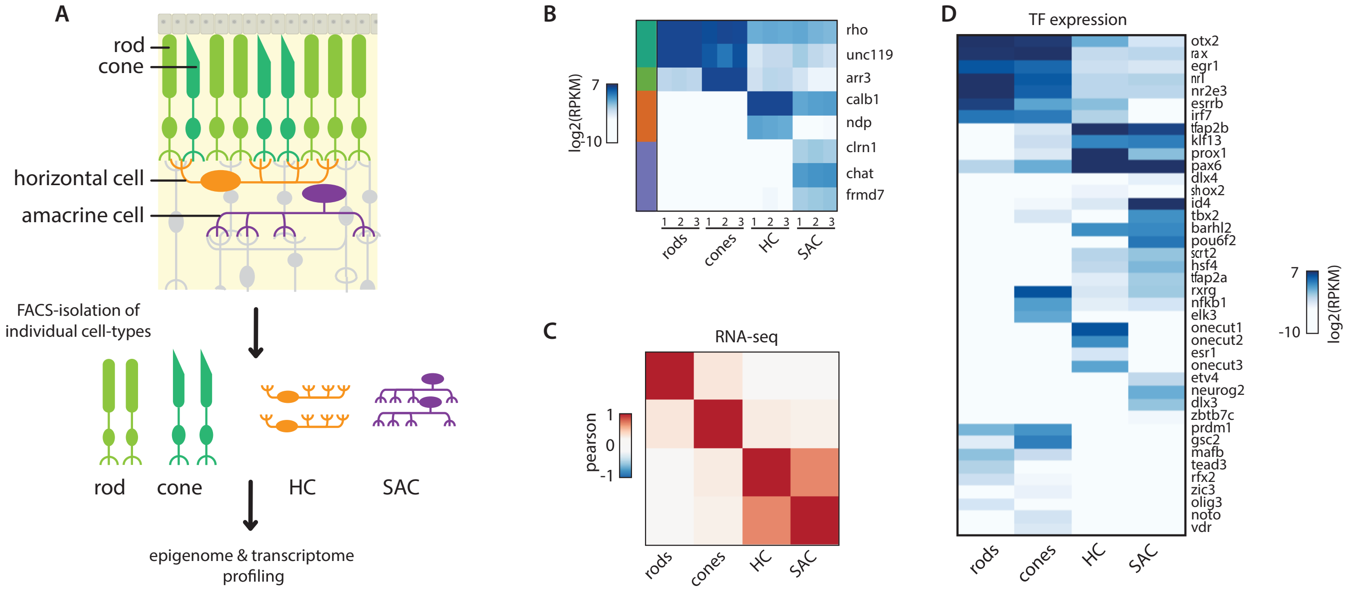


Figure 1. Transcriptome profiling of four cell types isolated from mouse retinas. (**A**) Scheme used to generate transcriptome and methylation datasets at the resolution of single cell types. Cell populations of rods, cones, horizontal cells and starburst amacrine cells are isolated by cell sorting from mouse transgenic lines carrying fluorescent markers that label these particular cell types of the retina. From cell isolates, whole genome methylation maps and expression datasets are generated. (**B**) Known cell specific expression markers reproducibly discriminate between cell isolates illustrating the reproducibility of the FACS procedure. Expression levels for markers of the studied retinal cell types. RPKM values for genic RNA-seq signal for samples issued from independent cell sorts. Side bar depicts the cell type associated with the marker (Siegert et al, 2009) (rods: dark green; cones: light green; HCs: orange; SACs: purple). Levels Rhodopsin in non-rod samples shows a particularly high degree of systematic contamination of cone samples with rods, as previously observed (Siegert et al, 2012; Mo et al, 2016). (**C**) The tested cell types show divergence in their expression profiles. Correlation heatmap comparing transcriptomes of the four studied cell types. Pearson correlation for genic RNA-seq signal merged across three biological replicates for each cell type. (**D**) Transcription factors showing differential expression between the tested cell types. Heatmap depicting the expression level of the most differentially expressed transcription factors between the tested cell types (top 10% variance). Shown are the RPKM values for genic RNA-seq merged across three biological replicates for each cell type. The heatmap was organized by hierarchical clustering.

not HCs with these neuronal types (Figure 2; Supplementary Figure S2A).

Reduced DNA methylation levels at regulatory regions is generally correlated with nucleosome depletion and increased chromatin accessibility. We therefore wondered if existing accessibility data derived from entire retinas (24,26) could support and strengthen our identification of cell type specific CREs. We first asked how the DNA methylation levels derived at the single cell-type level, would compare to existing ATAC-seq data derived from entire retinas or isolated photoreceptors (26). While we observed the expected anticorrelation between DNA methylation and accessibility for photoreceptors (Supplementary Figure S2I), we found no clear correlation for HCs or SACs. In line with the cellular composition of mouse retinas, this result suggests that accessibility data derived from entire retinas do not quantitatively reflect the chromatin status of rare cell-types such as interneurons. We nevertheless asked if interneuron specific LMRs would have low but detectable ATAC-seq signal in the whole retina datasets (Supplementary Figure S2J). This revealed that in contrast to photoreceptors, a large majority of interneuron LMRs (>95%) are not scored accessible in the dataset from whole retina (Supplementary Figure S2J). This suggests that whole retina is only poorly informative when studying CREs from rare cell types. Additionally, as previously reported (26), we noted that in rods, a significant fraction of LMRs are not accessible in a rod specific dataset (Supplementary Figure S2I, J). In order to simplify the identification of active regulatory regions, we excluded these regions for the downstream analyses.

Methylation information for individual cell types enabled us to classify these candidate regions based on their differential methylation (Figure 2B). For interneuron sub-types, this identifies several thousand cell type-specific regions (HC: 3520; SAC: 13520) (Figure 2B, exemplified in Figure 2D). Similarly, we identify several thousand regions specific for photoreceptors ($n = 4640$; Figure 2B, exemplified in Figure 2E). In addition, we observe significantly fewer elements specific for either rods ($n = 1090$) or cones ($n = 487$). This suggests that a large majority of regulatory regions are shared between these two developmentally closely related photoreceptors.

Identification of sequence determinants of cell-type specificity

Having defined sets of putative CREs in four cell types of the retina using cell type specific features of the methylome, we next asked if their sequences can educate on transcription factors involved in their regulation (18,19,49). Toward this goal, we performed motif enrichment analysis (Figure 2C) but restricted to sequence motifs linked to TFs expressed in at least one of the four retinal cell types that we studied in order to enhance the quality of our predictions.

This analysis revealed that each set of specific CREs is characterized by the enrichment of distinct sequence motifs

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

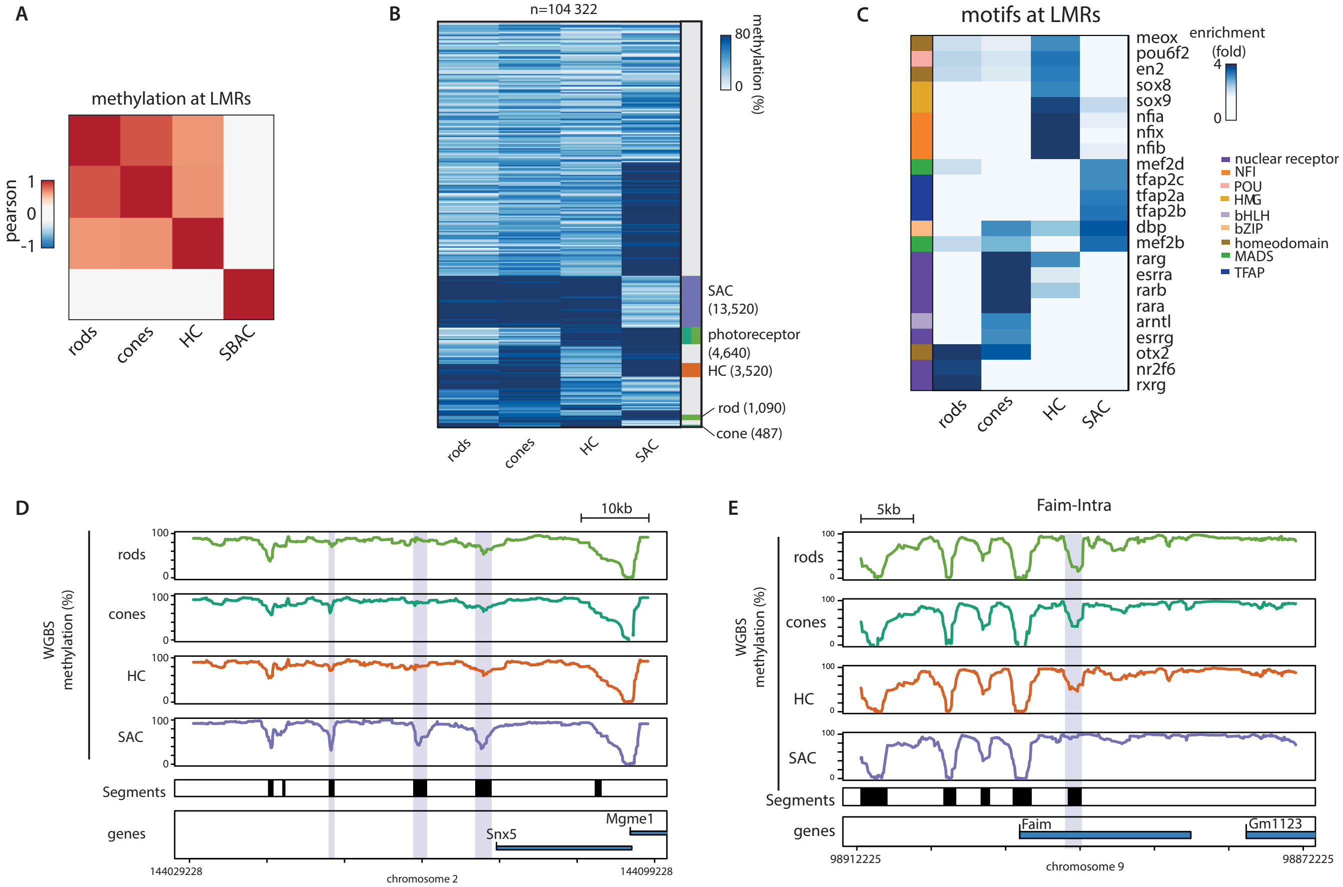


Figure 2. DNA methylation-based identification of putative CRE of the retinal system. (**A**) Correlation heatmap comparing methylation within low methylated regions detected in the retinal cell types. Pearson correlation for methylation of single CpGs located within the merged list of LMRs across cell types. Heatmap was subjected to hierarchical clustering. (**B**) Heatmap displaying average methylation of all putative CREs identified across the four studied retinal cell types. LMRs were grouped based on their binarized average methylation pattern (methylation < 60%). Sidebar indicates cell type specific clusters. Clusters were colored according to the specificity of the putative CRE (rods: dark green; cones: light green; HCs: orange; SACs: purple). (**C**) TF motifs enriched within subsets of putative CREs defined by their low methylated in individual cell types. Enrichment for known motifs was calculated for each set of cell-type specific LMRs and compared to a set of control sequences. Enrichments are displayed only for motifs significantly enriched in at least one cell-type specific subset, and for predicted transcription factors expressed in at least one of the cell types. Side-bar depicts the TF family associated with the enriched motif. (**D**, **E**) DNA methylation pattern of the four analyzed cell types of the retina at exemplified genomic regions. (D) Shown is the average WGBS signal around the Snx5 gene containing several SAC-specific low methylated regions (purple box) and the Faim gene containing a photoreceptor specific low methylated region (purple box). Black boxes denote regions displaying low methylation in at least one of the analyzed cell types, indicative of putative retinal CREs.

(Figure 2C). Photoreceptor CREs share enrichment for the motif of Otx2, which could also be bound by CRX. Both factors are critical for the development of the photoreceptor lineage (57–59). Interestingly, we find that distinct motifs for nuclear receptors (NRs) discriminate CREs in cones versus rods. These include canonical tandem repeat motifs implying NR dimerization, but also several monomeric motifs hinting at a role for orphan NR in this cellular specialization (Figure 2C). Our data further suggest a putative role for Sox- and Nfia-family factors in HCs identity, expanding on their previously described functions in multiple cellular differentiation processes within the central nervous system (60,61). Interestingly, TFAP-motifs discriminate SACs from other cell types. These motifs are recognized by AP-2 which has been recently linked to the amacrine cell fate (62).

Importantly, several of the motifs identified within these cell type specific CREs are potentially bound by TFs that are differentially expressed in our expression dataset (Figure 1D). In photoreceptors, these include Otx2 and several nuclear receptors (Figures 1D and 2C). In the case of SACs enrichment of motifs recognized AP-2 family of factors (Tfap2a–d), is in agreement with the high expression of Tfap2-a in these cell types. Altogether this dataset provides a comprehensive catalog of CREs putatively active in four cell types of the retina and identify sequence determinants potentially involved in their specificity.

Parallel reporter assay (PRA) in isolated cell types

Having identified a large set of putative CREs active in several cell types of the retinal system, we aimed to systematically test their ability to drive transcription autonomously in a reporter assay. We adapted the PRA-based strategy (23,34) to enable measurement of enhancer activity for thousands of constructs in distinct cell populations of the mouse retina *in vivo* using barcoded transcripts (Figure 3A).

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

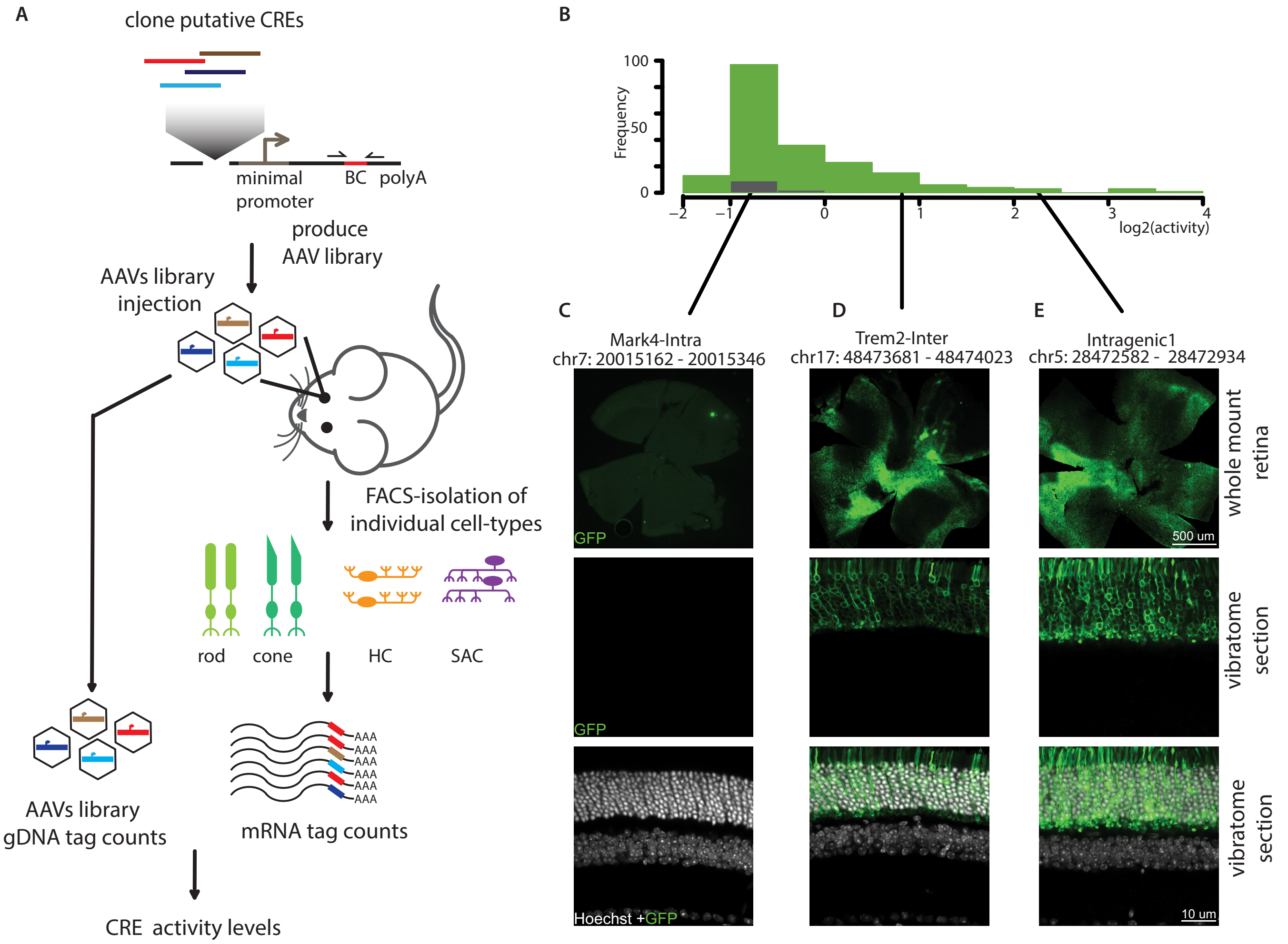


Figure 3. Parallelized reporter assay in specific retinal cell types. (**A**) Schematic representation of the procedure used to perform parallel reporter assays at the resolution of single cell types. Putative CREs are selected based on the detection of cell type-specific low methylation. Subsets of CREs are batch-cloned in front of a minimal promoter driving transcription of a GFP cassette followed by a unique barcode. These libraries are packaged into adeno-associated viruses and injected as pools in retinas of transgenic mice labeled for the cell types of interest. Three weeks following injection, cell populations are FACS sorted and RNA is extracted. CRE activity is determined as function of the barcode counts in the RNA normalized to the barcode counts in the AAV gDNA. (**B**) Histogram representing the distribution of activities observed for putative rod CREs (lowly methylated - green) and control regions (highly methylated - grey). Fragment activity was determined as ratio of barcode abundance in RNA sample versus abundance in the AAV pool used for infection. Displayed are average activity values derived from at least 3 biological replicates. CREs that were tested with an individual GFP reporter system are marked in the plot at their respective activity group. (**C–E**) Comparison of activity levels measured by PRA with trans-membrane GFP reporter signal for individual CREs. Immunohistochemical staining of whole mount and vibratome sections from wild type mouse retinas injected with individual constructs showing no (C), intermediate (D) or high (E) activity in the PRA assay. Green: GFP, white: Hoechst staining of DNA.

Studying individual cell types within a complex tissue poses several technical hurdles. Its success relies on the ability to measure complex libraries of fragments in the target cell type. This creates the necessity to reproducibly deliver diverse DNA libraries and to recover them from cell types under study. This is particularly challenging for cells that are rare within a tissue, which is the case for most cell types of the retina and true for most regions of the brain (2,54,63).

To test the general feasibility of our approach (Figure 3A), we first benchmarked our assay using a library that targets rod photoreceptors as the most abundant retinal cell type (2,54) using photoreceptor-specific methylation as a guide. We designed a pilot library of 384 fragments consisting of regions showing low methylation in rods (Figure 2B) plus 13 negative controls chosen to be highly methylated in rods and devoid of DHS signal in the whole retina (24). Most of the selected sequences arise from CREs that locate distal from promoters (Supplementary Figure S3B) and tend to be short (<600 bp, Supplementary Figure S3C). These sequences were cloned in front of a minimal promoter and a GFP coding sequence plus a 15 basepair randomized barcode at the 3′-end (Supplementary Figure S3A). We aimed for at least 10 barcodes per sequence to exclude barcode specific biases, and enhance technical reproducibility (29,30; Supplementary Figure S3D). We sequenced these plasmids to assign each CRE with its corresponding BCs and excluded barcodes that would associate with more than one CRE. These constructs were then packaged in high titer AAV (serotype 8) and subsequently used for sub-retinal injection into eyes of rod-labeled transgenic mice (Supple-

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

mentary Table S1). AAV allows for efficient infection of all target cell types used here due to its broad tropism (64). Moreover it is able to transduce non-dividing cells, displays low toxicity and supports strong and persistent transgene expression (65,66). Three weeks following injection, rods were isolated by FACS and RNA was extracted from ∼100 000 sorted rods or unsorted total retinas as a control. Barcode containing mRNAs were amplified and sequenced. The efficacy of infection was determined by measuring barcode complexity in each sample and by comparing it to the input viral AAV pool. This revealed an average recovery of 66%, illustrating that a large proportion of the library is present in the isolated cell population (Supplementary Figure S3E). For each of the 258 recovered sequences, we determined their relative activity as a measure of barcode abundance in the RNA relative to its DNA copy number in the viral pool (Figure 3A). This activity measure showed high reproducibility between samples from independently injected mice (Supplementary Figure S3F and G). A majority of the tested putative rod CREs show only basal activity that is comparable to our negative controls (Figure 3B) while ∼25% of the inserted sequences show transcriptional activity above background. The finding that only a minority of putative CREs are autonomous in driving detectable transcription in this ectopic context is expected and in line with previous reports (21,25,28–31). Importantly however, active sequences display expression over a wide dynamic range suggesting that our assay provides a sensitive and quantitative readout (Figure 3B). Moreover, we observed high correlation between activity levels observed in rods with that of total retinas ($R = 0.96$ - Supplementary Figure S3H), in agreement with the fact that rods make up ∼65% of this tissue (2,54).

Out of this library, we independently tested seven constructs *in vivo*, which cover a range of activities measured by PRA (Figure 3C–E). The fragments were cloned individually into a reporter system driving expression of GFP fused to the trans-membrane protein Channelrhodopsin (ChR2). Individual AAV preparations were injected into eyes of wild type mice. ChR2-GFP expression pattern in the retina was analyzed after antibody staining by confocal microscopy three weeks following injection. While unlikely to reflect quantitative expression differences, this validation system should enable to discriminate highly active from inactive constructs, and to identify the involved cell-types though spatial distribution of the fluorescence. We indeed observed a good agreement between PRA-measured activity and the ChR2-GFP signal observed in the retinal sections (Figure 3C–E). Only the fragments with high PRA signal displayed detectable GFP levels which are restricted to the photoreceptor layer. We conclude that PRA accurately reflects the autonomous transcriptional activity of DNA fragments *in vivo* when performed at the level of single cell types.

Autonomous CRE activities in four retinal cell types

Next we aimed to characterize the autonomous activity of putative CREs identified in the four cell types of interest: rods, cones, HCs or SACs (Figure 3A). We designed two additional libraries that cover a spectrum of putative CREs based on their local hypo-methylation (Figure 2B). These libraries contained shared elements but their composition was biased either toward CRE active in photoreceptors (library #2) (Supplementary Figure S4A) or interneurons (library #3) (Supplementary Figure S4B) as measured by differential DNA methylation. After generating viral pools, we infected retinas of adult transgenic mice (P50–150) that carry specific markers for a particular cell type (3,4; Supplementary Table S1). For each experiment, we sorted cells from three independently injected mice. We systematically performed activity measures in the labeled target cell type and contrasted it with its activity in the remaining unlabeled cell types of the retina.

We first aimed to characterize the *in vivo* activity of putative photoreceptor CREs (Figure 2B) in cones and rods separately. We infected retinas with AAVs carrying library #2 that contains elements showing preferential putative activity in photoreceptors based on their differential methylation. As observed for our pilot experiment in rods (Figure 3B), we observed that only a fraction of the tested CREs drive detectable activity in rods (Figure 4A) or cones (Figure 4B). Additionally, we found that the tested set of CREs display very similar activity in rods and cones (Figure 4C), with only subtle differences between the two cell types. Thus, we were unable to separate cones from rods with the tested set of CREs. This is unlikely to be the sole consequence of contamination by rods as this should not affect the detection of cone specific elements. More likely, it could reflect the high similarity in the regulatory networks active in these two cell types (Figure 2A, B). Alternatively, we cannot exclude that the episomal reporter system used here does not accurately recapitulate differences at the chromosomal level between these cell types (67).

The remaining unlabeled cells behaved similarly to rods and cones. This is mostly due to the fact that these cells contain cones or rods depending on the sorted cell type (Figure 4A, B). Additionally, we cannot exclude that some of the labeled cells are sorted into the unlabeled fraction if their signal was low since we aimed for higher stringency in the labeled cells rather than the unlabeled fraction.

Interneurons such as SACs and HCs on the other hand differ largely from photo-receptors in their repertoire of CREs (Figure 2A, B), which would suggest high differential CRE activity. We then constructed a third library combining 5 fragments identified to have a high autonomous activity in photoreceptors (Figure 4A–C) with a set of 365 CREs showing low methylation in interneurons (HCs and/or SACs). We measured the activity of these fragments in SACs or HCs as well as in the respective remaining unlabeled cells of the retina. Contrasting CRE activity in interneurons with the remaining unlabeled total retina (composed of ∼65% rod photoreceptors) revealed differences over three orders of magnitude (Figure 4D, E). In contrast the two types of interneurons show very similar activity profiles for the tested set of fragments (Figure 4F). These observations are in line with the divergence between the tested cell types (Figures 1C and 2A) and identify a set of CRE with differential autonomous activity between photoreceptor and interneurons.

Since only a fraction of our tested sequences were active in our assay, we wondered if any genetic or chromatin features enrich at sequences that show autonomous activ-

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

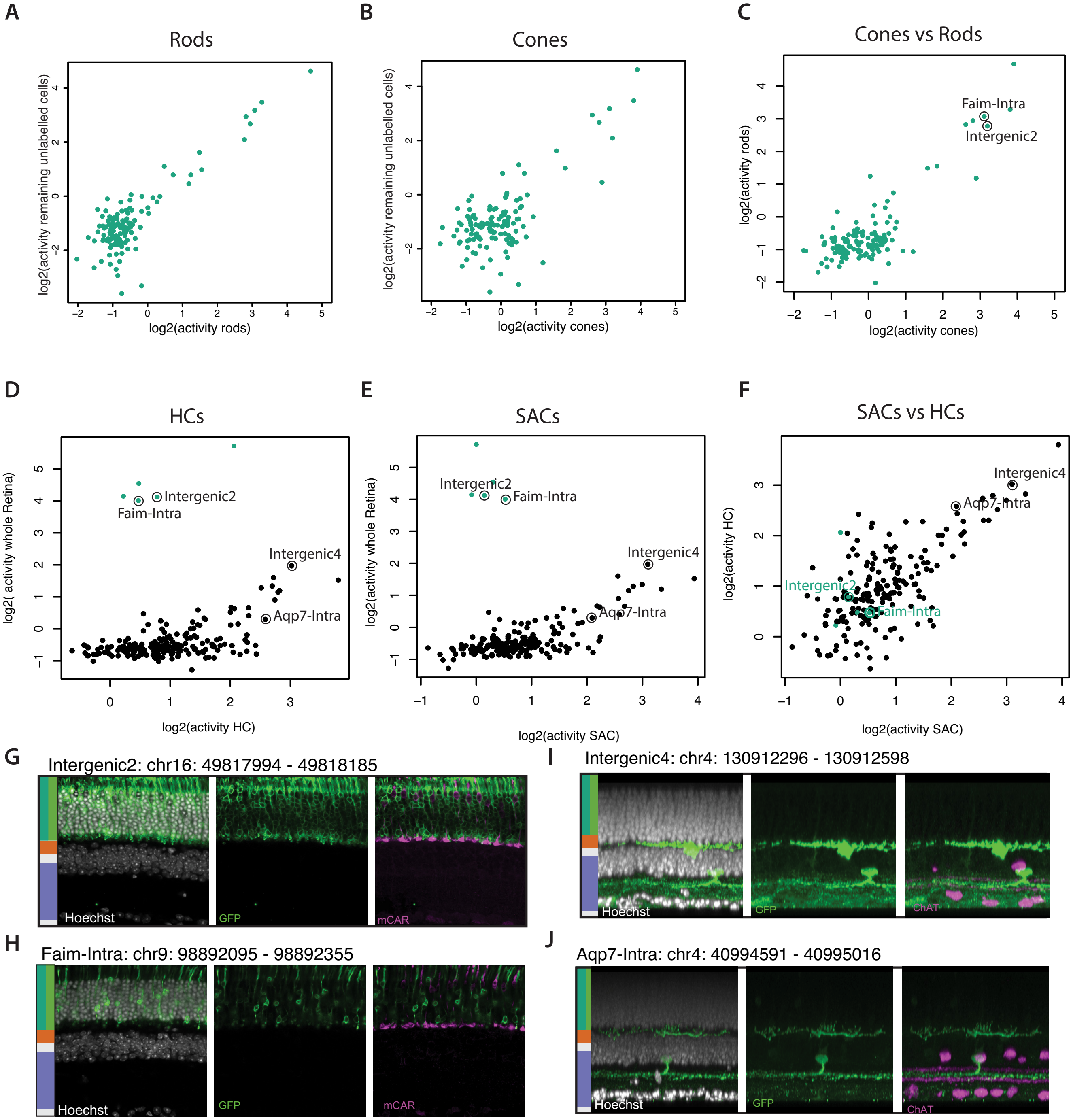


Figure 4. Characterization of the autonomous activity of CREs in distinct cell types of the retina. (**A, B**) Most tested sequences show similar activity in rods and cones. Scatter plot contrasting activity of CREs as measured by PRA in (A) rods or (B) cones against the remaining pool of unlabeled cells from the retina. For each experiment fluorescently labeled cells and remaining unlabeled cells were sorted from retinas infected by the AAV PRA library. Displayed is the normalized activity for each fragment averaged over biological replicates. (**C**) Scatter plot contrasting activity of CREs as measured by PRA in rods versus cones. (**D, E**) Most of the tested sequences show similar activity in HCs and SACs that largely differs from activity observed in whole retina. Similar scatterplot as in (A-B), contrasting PRA activity in (**D**) HCs or (**E**) SACs against the remaining pool of unlabeled cells from the retina. (**F**) Scatter plot contrasting activity of CREs as measured by PRA in HCs versus SACs. (**G-H**) Microscopical validation of the specificity of individual photoreceptor specific CREs in the tissue context. Shown is Chr2-GFP fluorescence for individual reporter constructs with indicated CRE. Immunohistochemical staining of vibratome sections from transgenic mouse (cone-labeled, mCAR) retinas injected with two CREs detected by PRA to be active in photoreceptor but not in interneurons. White, Hoechst; green, GFP; purple, cone marker mCAR. Scale bar, 10 μm. Side bar depicts the expected localization of the different studied cell types based on the considered retina layers. (**I** and **J**) Similar validation as in (G-H) for interneuron specific CREs. Shown is Chr2-GFP fluorescence for individual reporter constructs with indicated CRE. Immunohistochemical staining of retina mounts from transgenic mouse (SAC-labeled, ChAT) retinas injected with two CREs detected by PRA to be active in interneurons but not in photoreceptors. White, Hoechst; green, GFP; purple, cone marker ChAT. Scale bar, 10 μm. Side bar depicts the expected localization of the different studied cell types based on the considered retina layers.

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

ity. Indeed active sequences show a slightly higher enrichment for TF motifs compared to the inactive ones (Supplementary Figure S4C). Enriched motifs include CRX/Otx2 or NR type motifs for photoreceptors and AP2 motifs for interneurons, suggesting that these factors may be important for the activity in the respective cell types. It is important to note however that these are also enriched (though to a lesser degree) in inactive fragments compared to background sequences (Supplementary Figure S4C), indicating that presence of these motifs alone is not sufficient to explain activity. We then asked if methylation or accessibility of their originating sequences would also help to predict their activity (Supplementary Figure S4D, E). None of these chromatin features could clearly discriminate active from inactive fragments. However, we noted that fragments active in photoreceptors tend to show lower methylation levels in cones when compared to inactive ones (Supplementary Figure S4D). Additionally, fragments active in photoreceptors tend to originate from regions showing higher accessibility in whole-retina (Supplementary Figure S4E). Together, this suggests that selecting fragments containing a high number of TF motif occurrences and high levels of chromatin alterations (lower methylation levels and high accessibility) could potentially improve identification of sequences with autonomous activity.

Having determined the relative activity of individual elements at the level of mRNA, we next wanted to ask if their activity is sufficient to autonomously drive the expression levels of the large reporter gene mentioned above (Chr2-GFP) in a specific fashion and at levels that can be detected by *in situ* microscopy. For this we selected 20 CRE sequences and cloned them into a ChR2-GFP reporter for individual testing in the entire retina. Compared to conventional GFP, the utilized fusion protein locates to the cellular membrane, which facilitates the identification of retinal cell types *in vivo*. We observed that only the fragments showing the highest relative activity in PRA led to robustly detectable GFP signal in retinal sections for photoreceptors (Supplementary Figure S4F) and interneurons (Supplementary Figure S4G). From this small subset, we identified fragments driving specific activity in photoreceptors (Figure 4G, H) or in inter-neurons (Figure 4I, J), mirroring the strong differences observed between these cell-types in PRA (Figure 4D, E). In contrast to this clear separation of retinal sub-types (photoreceptors versus inter-neurons), the expression pattern of the tested fragments was rarely limited to a single cell-type. Additionally while several CREs display clear activity in HC and amacrine cells (i.e. Figure 4I, J, Supplementary Table S2), we did not detect Chr2-GFP in starbust amacrine cells, the ChAT positive amacrine cell subtype that we used in the sort. These results likely illustrate the inherent limits of the current screen in which we focused on only four out of the >50 cell types composing the retina. In summary, we show proof of principle that PRA can be used to systematically characterize the autonomous activity pattern of CREs in multiple cell-types and therefore identify short DNA fragments able to confer robust transcriptional activity in different cellular subsets of the retina. A summary of all PRA results is provided as Supplementary files.

Functional dissection of the architecture of two photoreceptor CREs

Analysis of the occurrence of sequence motifs within CREs active in retinal cell types revealed putative TF motifs involved in their regulation (Figure 2C). We aimed to test if and how some of these motifs contribute to CRE activity ectopically. Such information would be relevant to better understand the regulatory landscape that controls cell identity, but could also provide opportunities to effectively modulate the activity of identified elements.

To do so, we selected two elements that we identified as being active in photoreceptors (Figure 4C and Supplementary Table 2) and that contained TF motifs specific for these cell types (Figure 5A, B). For these, we generated a library of sequence mutants to systematically test the effect of deleting each motif instance on CRE activity (Figure 5C, D). Each motif occurrence detected within the sequence was iteratively replaced with a random sequence (Figure 5C, D), generating a collection of individual mutants for each fragment.

Performing PRA with this mutant library identifies diverse effects depending on the mutation. This included reduced activity upon motif randomization as might be expected from removing a binding site for an activating factor. However, we also detect in several cases increased levels of CRE activity upon ablation of TF motifs suggesting removal of a binding site for a repressive factor. Together this suggests that negative and positive regulatory inputs co-exist to modulate precise activity level of these particular CREs.

More specifically we detect reduced CRE activity upon deletion of CRX/Otx2, Sp1 or motifs typically bound by factors containing a homeodomain (Figure 5C, D). This is in agreement with the notion that these motifs act as activators. In contrast, mutating the conserved monomeric NR motif enhanced CRE activity (Figure 5C), arguing that NR mediated repression negatively modulates the activity of this element (Faim-Intra). This evidence for repression seems particularly interesting given that this element already displayed the highest activity in our screen (Figure 4C). Similarly, deletion of the NRL motif enhances CRE activity suggesting that NRL can act as a repressor in this context. This potentially reveals a novel mechanism of action for NRL that was hitherto only known for its activation function at the rhodopsin promoter (56,68). In summary, we show how PRA assays can be used to identify and functionally annotate determinant TF motifs within CREs. Moreover, we demonstrate how discrete sequence changes can be used to rationally enhance or reduce transgene expression levels *in vivo*.

DISCUSSION

Using transcriptome and epigenome profiling, this study identifies a large collection of putative *cis*-regulatory elements active in four distinct cell types of the retina. Additionally, we provide proof of concept for the usage of parallel reporter assays to measure the autonomous transcriptional activity within the retinal tissue at the level of individual cell types. We successfully apply this framework to

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

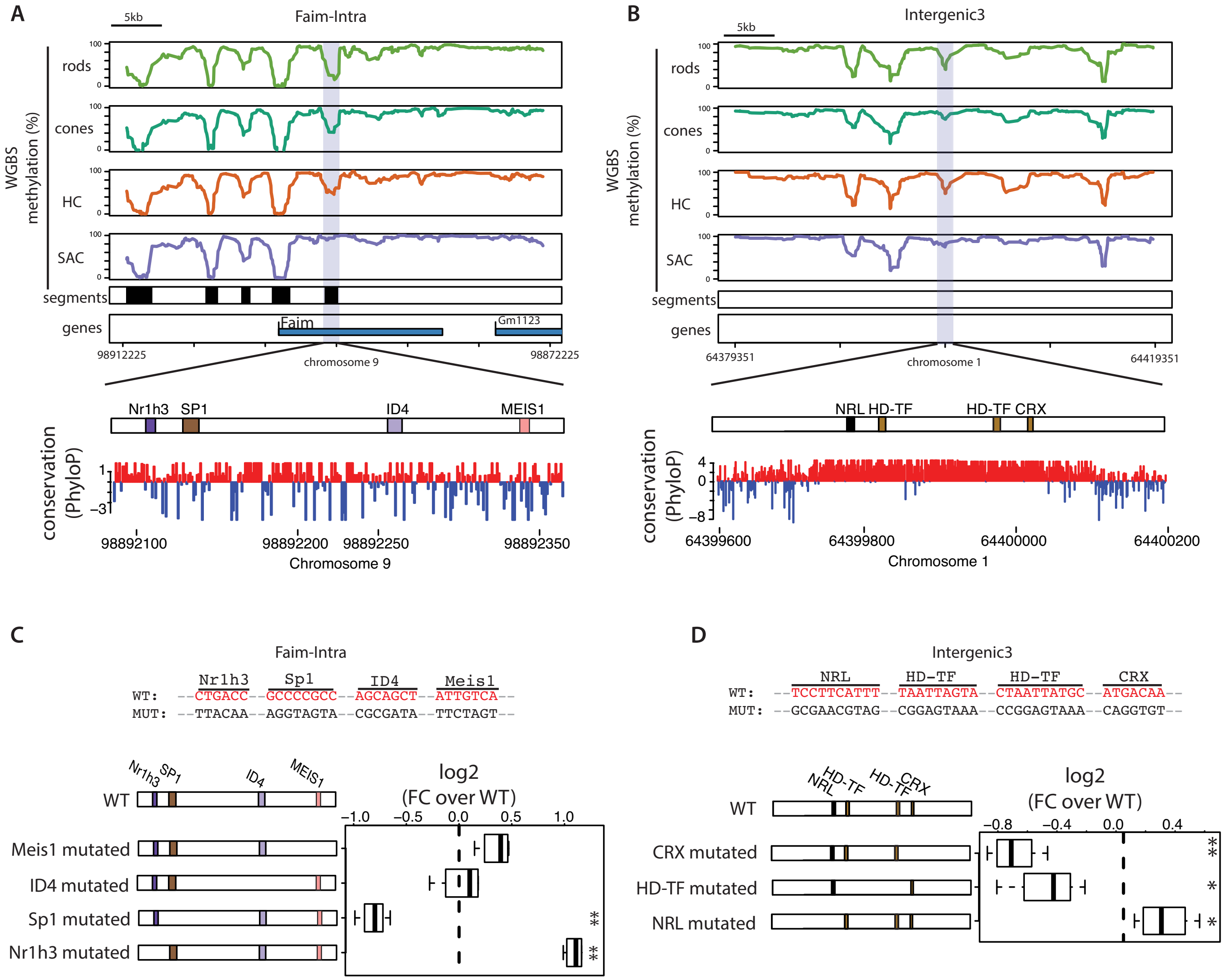


Figure 5. Dissection of the architecture of two photoreceptor CREs. (**A, B**) DNA methylation pattern of the four analyzed cell types of the retina around the two CREs selected for functional dissections (purple box); (A) Faim-Intra and (B) Intragenic 3 enhancers. Black boxes denote regions displaying low methylation in at least one of the analyzed cell types, representing putative retinal CREs. The TF architecture of the CREs is detailed, showing the TF binding sites identified (colored according to the TF family as in Figure 2C) and the PhyloP conservation track for the region. (**C, D**) Effects on activity of systematic deletion of TF motifs for two CREs. Each motif identified in the CRE sequence was iteratively replaced by a randomized sequence (shown in upper panel). Activity of the generated library of mutants was compared to the wild type sequence. Boxplots represents the distribution of individual measurements from different biological replicates. Statistical significance of changes in activity were tested using a bidirectional *t*-test (*$P < 0.05$; **$P < 0.01$).

define the activity pattern of hundreds of short DNA sequences in several cell types of the retina. This effort let to the identification of a small set of short sequences showing preferential activity in different cellular subsets of the retina. We also demonstrate how this technology can be used to dissect the architecture of regulatory regions *in vivo*.

PRA has been previously applied *in vivo* but without discriminating between specific cell types that make up a tissue (23,38,40). Here we use FACS-sorting on a library of mouse lines, where various cell types are fluorescently labeled. This enables reproducible isolation of pure cell populations (3,4), but is inherently constrained by the number of cells available, which is particularly limiting for rare cell types. In order to circumvent these bottlenecks and to derive accurate PRA measures for low cell numbers, we combined high efficiency AAV-based delivery of our libraries with multiplexed measures for each fragment. When contrasting whole tissue with cell-type specific data we observed that whole tissue data only reflects sequence activity in photoreceptors, which is the dominant cell type in mouse retina. Consequently whole tissue analysis failed to capture activities in rare cell types thus demonstrating that cell type isolation and activity assignment is critical.

It has recently been established that DNA hypomethylation at distal regulatory regions coincides with their accessibility and putative activity in many cell types and tissues (18,49). Here we used this epigenetic feature as a guide to nominate CREs active in a cell-type specific manner. In agreement with previous reports (21,25,28–31), most of the hypo-methylated CREs tested failed to autonomously drive

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

transcription in our reporter assay. Using chromatin accessibility data available for some cell types (26), we found that most of these inactive sequences are indeed accessible, ruling out that alterations of DNA methylation dynamics (as observed in rods (26)) could explain this high rate of negatives. Alternatively, we think that this result can likely be explained by the inherent inability of some CREs to function autonomously (21,25,28–31), their incompatibility with the TATA containing minimal promoter used in this assay (69) and/or the truncation of the regulatory element in our systematic library design (21,25,28–31). In any case, this confirms the requirement of a high throughput screening strategy to identify CREs that function autonomously.

Mining of the resulting datasets identified TF binding motifs that associate with cell type specific activity within the retina. Mutational analysis indeed reveals functional relevance for these motifs within a given CRE. One striking observation was the enrichment for NR motifs within sequences active in photoreceptors. This included differential enrichment of monomeric NR motifs between rods and cones, suggesting a role for orphan NRs in distinguishing between these closely related cell types. Importantly deletion of this motif leads to increased activity in an example CRE, which is in agreement with the repressive function assigned to some orphan NR (70). While the motif does not reveal the responsible TF, we note that only a few candidate orphan NRs are differentially expressed between rods and cones, which we hypothesize to be likely candidates (Supplementary Figure S4C). One of these is Nr1h3 (LXRα) a well-known regulator of lipid metabolism in liver. This factor is highly expressed in cones and was recently linked to the development of the zebrafish visual system (71). Another relevant finding relates to the recognition motif for NRL, which is a known decisive factor during photo-receptor differentiation (56,72). Mutations of this motif resulted in enhanced CRE activity, indicating an unexpected repressive role for NRL at least within the tested element. This contrasts with its function at the rhodopsin promoter (56,68,72), suggesting that NRL function could be context specific, in line with recent reports of regulation of its activity through complex interactions (73).

For several inherited retinal diseases, gene-replacement or targeted expression of optogenetic sensors is considered a credible strategy to reverse phenotypes and at least partially restore vision (5–7). Such approach critically relies on short sequences that can drive expression of a transgene from an AAV in a defined cell type *in vivo*. Our parallel measurements defined the expression pattern within four defined cell types for a catalogue of short sequences in the mouse retina. Testing the ability of a subset of these CREs to drive levels of transgene relevant for cellular manipulation highlights the importance of selecting CREs displaying specific but also strong activity patterns. In the current screen, we identified a small set of sequences with preferential expression in different cellular subtypes of the retina, but not identifying unique cell types. Yet only sequences displaying the highest activity level in our parallel assay led to expression levels of a fusion reporter protein considered to be sufficient for cellular manipulation. Thus targeting of disease relevant cell-types of the retina, will ultimately require a better understanding of the sequence features that define cell type specific expression and thus further screening efforts and sequence engineering to more cell types. The current study demonstrates that PRA applied at the cell type specific level provides throughput and sensitivity to contribute to this goal.

DATA AVAILABILITY

The raw and processed data have been submitted to GEO under the accession GSE84589. The processed data can be visualized on the UCSC genome browser at: http://genome-euro.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=krebsarnaud&hgS_otherUserSessionName=Hartl_et_al_retina_2017.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Michael Stadler for his support in developing the PRA analysis strategy, Hubertus Kohler for cell sorting and Birgitte Lucas for low-cell RNA-seq library preparation. The authors would like to thank members of the Schübeler laboratory for helpful discussions and comments on the manuscript.

Author contributions: A.R.K. initiated and designed the study with support from D.S and B.R.. D.H., designed the PRA strategy with support from A.R.K. and D.S.. A.R.K. and D.H. designed and performed genomics and PRA experiments. J.J. generated and injected AAV libraries, sorted cells, performed and analyzed *in vivo* testing of single fragments. D.H. implemented the PRA analysis pipeline with support of A.R.K., D.H and A.R.K analyzed data. D.H., A.R.K. and D.S. wrote the manuscript with help from B.R. and J.J.. D.S. and B.R. supervised the project.

FUNDING

Novartis Research Foundation (to DS and BR); European Research Council (to BR and ReaDMe - REP-667951 to DS); Swiss National Sciences Foundation (to D.S. and B.R.); Swiss National Sciences Foundation Ambizione (PZOOP3_161493 to A.R.K.); Boehringer Ingelheim Fonds (to D.H.). Funding for open access charge: Swiss National Sciences Foundation Ambizione.

Conflict of interest statement. None declared.

REFERENCES

1. Cepko,C. (2014) Intrinsically different retinal progenitor cells produce specific types of progeny. *Nat. Rev. Neurosci.*, **15**, 615–627.
2. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
3. Siegert,S., Scherf,B.G., Del Punta,K., Didkovsky,N., Heintz,N. and Roska,B. (2009) Genetic address book for retinal cell types. *Nat. Neurosci.*, **12**, 1197–1204.

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

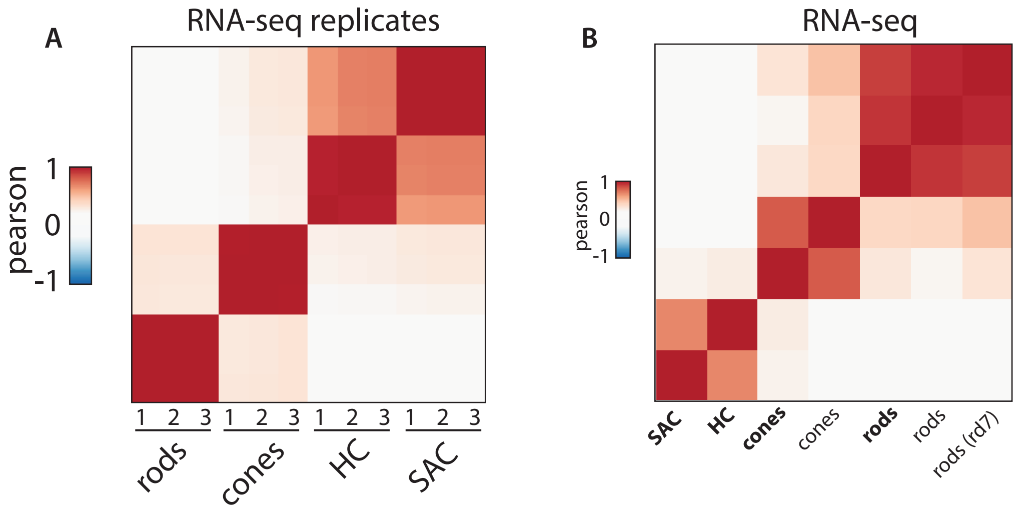
4. Siegert,S., Cabuy,E., Scherf,B.G., Kohler,H., Panda,S., Le,Y.-Z., Fehling,H.J., Gaidatzis,D., Stadler,M.B. and Roska,B. (2012) Transcriptional code and disease map for adult retinal cell types. *Nat. Neurosci.*, **15**, 487–495.
5. Sahel,J.-A.A. and Roska,B. (2013) Gene therapy for blindness. *Annu. Rev. Neurosci.*, **36**, 467–488.
6. Boye,S.E., Boye,S.L., Lewin,A.S. and Hauswirth,W.W. (2013) A comprehensive review of retinal gene therapy. *Mol. Ther.*, **21**, 509–519.
7. Nash,B.M., Wright,D.C., Grigg,J.R., Bennetts,B. and Jamieson,R. V (2015) Retinal dystrophies, genomic applications in diagnosis and prospects for therapy. *Transl. Pediatr.*, **4**, 139–163.
8. Xu,J. and Smale,S.T. (2012) Designing an enhancer landscape. *Cell*, **151**, 929–931.
9. Smith,E. and Shilatifard,A. (2014) Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.*, **21**, 210–219.
10. Smallwood,A. and Ren,B. (2013) Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.*, **25**, 387–394.
11. Calo,E. and Wysocka,J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, **49**, 825–837.
12. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
13. Thurman,R., Rynes,E., Humbert,R., Vierstra,J., Maurano,M., Haugen,E., Sheffield,N., Stergachis,A., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
14. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
15. Thurman,R., Rynes,E., Humbert,R., Vierstra,J., Maurano,M., Haugen,E., Sheffield,N., Stergachis,A., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
16. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
17. Hodges,E., Molaro,A., Dos Santos,C.O., Thekkat,P., Song,Q., Uren,P.J., Park,J., Butler,J., Rafii,S., McCombie,W.R. *et al.* (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell*, **44**, 17–28.
18. Stadler,M.B., Murr,R., Burger,L., Ivanek,R., Lienert,F., Schöler,A., van Nimwegen,E., Wirbelauer,C., Oakeley,E.J., Gaidatzis,D. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.
19. Ziller,M.J., Gu,H., Müller,F., Donaghey,J., Tsai,L.T.-Y., Kohlbacher,O., De Jager,P.L., Rosen,E.D., Bennett,D.A., Bernstein,B.E. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.
20. Mo,A., Mukamel,E.A., Davis,F.P., Luo,C., Henry,G.L., Picard,S., Urich,M.A., Nery,J.R., Sejnowski,T.J., Lister,R. *et al.* (2015) Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron*, **86**, 1369–1384.
21. Ernst,J., Melnikov,A., Zhang,X., Wang,L., Rogov,P., Mikkelsen,T.S. and Kellis,M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.*, doi:10.1038/nbt.3678.
22. Attanasio,C., Nord,A.S., Zhu,Y., Blow,M.J., Li,Z., Liberton,D.K., Morrison,H., Plajzer-Frick,I., Holt,A., Hosseini,R. *et al.* (2013) Fine tuning of craniofacial morphology by distant-acting enhancers. *Science*, **342**, 1241006.
23. Patwardhan,R.P., Hiatt,J.B., Witten,D.M., Kim,M.J., Smith,R.P., May,D., Lee,C., Andrie,J.M., Lee,S.-I., Cooper,G.M. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, **30**, 265–270.
24. Wilken,M.S., Brzezinski,J.A., La Torre,A., Siebenthall,K., Thurman,R., Sabo,P., Sandstrom,R.S., Vierstra,J., Canfield,T.K., Hansen,R.S. *et al.* (2015) DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenet. Chromatin*, **8**, 8.
25. Shen,S.Q., Myers,C.A., Hughes,A.E.O., Byrne,L.C., Flannery,J.G. and Corbo,J.C. (2016) Massively parallel cis -regulatory analysis in the mammalian central nervous system. *Genome Res.*, doi:10.1101/gr.193789.115.
26. Mo,A., Luo,C., Davis,F.P., Mukamel,E.A., Henry,G.L., Nery,J.R., Urich,M.A., Picard,S., Lister,R., Eddy,S.R. *et al.* (2016) Epigenomic landscapes of retinal rods and cones. *Elife*, **5**, 1–29.
27. Solovei,I., Kreysing,M., Lanctôt,C., Kösem,S., Peichl,L., Cremer,T., Guck,J. and Joffe,B. (2009) Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell*, **137**, 356–368.
28. Kheradpour,P., Ernst,J., Melnikov,A., Rogov,P., Wang,L., Alston,J., Mikkelsen,T.S. and Kellis,M. (2013) Systematic dissection of motif instances using a massively parallel reporter assay. *Genome Res.*, doi:10.1101/gr.144899.112.
29. Tewhey,R., Kotliar,D., Park,D.S., Liu,B., Winnicki,S., Reilly,S.K., Andersen,K.G., Mikkelsen,T.S., Lander,E.S., Schaffner,S.F. *et al.* (2016) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, **165**, 1519–1529.
30. Ulirsch,J.C., Nandakumar,S.K., Wang,L., Giani,F.C., Zhang,X., Rogov,P., Melnikov,A., McDonel,P., Do,R., Mikkelsen,T.S. *et al.* (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, **165**, 1530–1545.
31. Kwasnieski,J.C., Fiore,C., Chaudhari,H.G. and Cohen,B.A. (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.*, **24**, 1595–1602.
32. Li,Q., Peterson,K.R., Fang,X. and Stamatoyannopoulos,G. (2002) Review article Locus control regions. *Blood*, **100**, 3077–3086.
33. Mogno,I., Kwasnieski,J.C. and Cohen,B.a (2013) Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.*, doi:10.1101/gr.157891.113.
34. Melnikov,A., Murugan,A., Zhang,X., Tesileanu,T., Wang,L., Rogov,P., Feizi,S., Gnirke,A., Callan,C.G., Kinney,J.B. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, **30**, 271–277.
35. Kwasnieski,J.C., Mogno,I., Myers,C.A., Corbo,J.C. and Cohen,B.A. (2012) Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 19498–19503.
36. White,M.a, Myers,C.a, Corbo,J.C. and Cohen,B.a (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11952–11957.
37. Patwardhan,R.P., Lee,C., Litvin,O., Young,D.L., Pe'er,D. and Shendure,J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.*, **27**, 1173–1175.
38. Shen,S.Q., Myers,C.A., Hughes,A.E., Byrne,L.C., Flannery,J.G. and Corbo,J.C. (2015) Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.*, doi:10.1101/gr.193789.115.
39. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.
40. Gisselbrech,S.S., Barrera,L.A., Porsch,M., Aboukhalil,A., 3rd,P.W.E., Vedenko,A., Palagi,A., Kim,Y., Zhu,X., Busser,B.W. *et al.* (2014) Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Igarss 2014*, **10**, 1–5.
41. Le,Y.-Z., Ash,J.D., Al-Ubaidi,M.R., Chen,Y., Ma,J.-X. and Anderson,R.E. (2004) Targeted expression of Cre recombinase to cone photoreceptors in transgenic mice. *Mol. Vis.*, **10**, 1011–1018.
42. Le,Y.-Z., Zheng,L., Zheng,W., Ash,J.D., Agbaga,M.-P., Zhu,M. and Anderson,R.E. (2006) Mouse opsin promoter-directed Cre recombinase expression in transgenic mice. *Mol. Vis.*, **12**, 389–398.
43. Grieger,J.C., Choi,V.W. and Samulski,R.J. (2006) Production and characterization of adeno-associated viral vectors. *Nat. Protoc.*, **1**, 1412–1428.
44. Busskamp,V., Duebel,J., Balya,D., Fradot,M., Viney,T.J., Siegert,S., Groner,A.C., Cabuy,E., Forster,V., Seeliger,M. *et al.* (2010) Genetic reactivation of cone photoreceptors restores visual responses in Retinitis pigmentosa. *Science*, **329**, 413–417.
45. Trimarchi,J.M., Stadler,M.B., Roska,B., Billings,N., Sun,B., Bartch,B. and Cepko,A.C.L. (2007) Molecular heterogeneity of

Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017

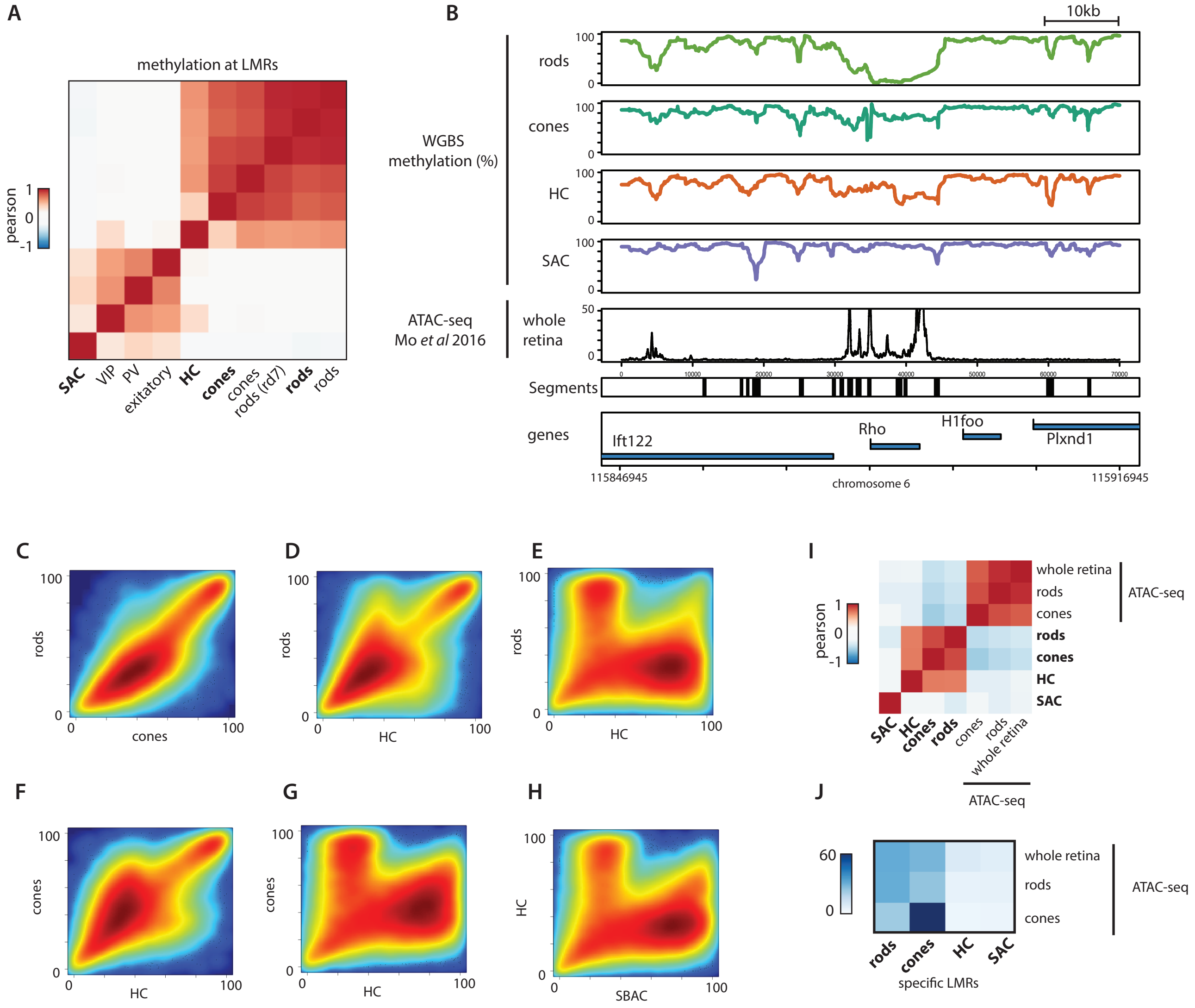
developing retinal ganglion and amacrine cells revealed through single cell gene expression profilin. *J. Comp. Neurol.*, **504**, 287–297.

46. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
47. Gaidatzis,D., Lerch,A., Hahne,F. and Stadler,M.B. (2015) QuasR: quantification and annotation of short reads in R. *Bioinformatics*, **31**, 1130–1132.
48. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
49. Burger,L., Gaidatzis,D., Schübeler,D. and Stadler,M.B. (2013) Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.*, **41**, e155.
50. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
51. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
52. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
53. Kim,J.W., Yang,H.J., Brooks,M.J., Zelinger,L., Karakülah,G., Gotoh,N., Boleda,A., Gieser,L., Giuste,F., Whitaker,D.T. *et al.* (2016) NRL-regulated transcriptome dynamics of developing rod photoreceptors. *Cell Rep.*, **17**, 2460–2473.
54. Jeon,C.J., Strettoi,E. and Masland,R.H. (1998) The major cell populations of the mouse retina. *J. Neurosci.*, **18**, 8936–8946.
55. Blackshaw,S., Harpavat,S., Trimarchi,J., Cai,L., Huang,H., Kuo,W.P., Weber,G., Lee,K., Fraioli,R.E., Cho,S. *et al.* (2004) Genomic analysis of mouse retinal development. *PLoS Biol.*, **2**, E247.
56. Swaroop,A., Kim,D. and Forrest,D. (2010) Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nat. Rev. Neurosci.*, **11**, 563–576.
57. Nishida,A., Furukawa,A., Koike,C., Tano,Y., Aizawa,S., Matsuo,I. and Furukawa,T. (2003) Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. *Nat. Neurosci.*, **6**, 1255–1263.
58. Koike,C., Nishida,A., Ueno,S., Saito,H., Sanuki,R., Sato,S., Furukawa,A., Aizawa,S., Matsuo,I., Suzuki,N. *et al.* (2007) Functional roles of Otx2 transcription factor in postnatal mouse retinal development. *Mol. Cell. Biol.*, **27**, 8318–8329.
59. Peng,G.H., Ahmad,O., Ahmad,F., Liu,J. and Chen,S. (2005) The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes. *Hum. Mol. Genet.*, **14**, 747–764.
60. Deneen,B., Ho,R., Lukaszewicz,A., Hochstim,C.J., Gronostajski,R.M. and Anderson,D.J. (2006) The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord. *Neuron*, **52**, 953–968.
61. Piper,M., Barry,G., Hawkins,J., Mason,S., Lindwall,C., Little,E., Sarkar,A., Smith,A.G., Moldrich,R.X., Boyle,G.M. *et al.* (2010) NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. *J. Neurosci.*, **30**, 9127–9139.
62. Jin,K., Jiang,H., Xiao,D., Zou,M., Zhu,J. and Xiang,M. (2015) Tfap2a and 2b act downstream of Ptf1a to promote amacrine cell differentiation during retinogenesis. *Mol. Brain*, **8**, 28.
63. Tasic,B., Menon,V., Nguyen,T.N.T., Kim,T.T.K., Jarsky,T., Yao,Z., Levi,B.B., Gray,L.T., Sorensen,S.A., Dolbeare,T. *et al.* (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, **19**, 335–346.
64. Stieger,K., Colle,M.-A., Dubreil,L., Mendes-Madeira,A., Weber,M., Le Meur,G., Deschamps,J.Y., Provost,N., Nivard,D., Cherel,Y. *et al.* (2008) Subretinal delivery of recombinant AAV serotype 8 vector in dogs results in gene transfer to neurons in the brain. *Mol. Ther.*, **16**, 916–923.
65. Mueller,C. and Flotte,T.R. (2008) Clinical gene therapy using recombinant adeno-associated virus vectors. *Gene Ther.*, **15**, 858–863.
66. Buning,H., Perabo,L., Coutelle,O., Quadt-Humme,S. and Hallek,M. (2008) Recent developments in adeno-associated virus vector technology. *J. Gene Med.*, **10**, 610–618.
67. Inoue,F., Kircher,M., Martin,B., Cooper,G.M., Witten,D.M., Mcmanus,M.T., Ahituv,N. and Shendure,J. (2017) A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.*, **27**, 38–52.
68. Rehemtulla,A., Warwar,R., Kumar,R., Ji,X., Zack,D.J. and Swaroop,A. (1996) The basic motif-leucine zipper transcription factor Nrl can positively regulate rhodopsin gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 191–195.
69. Zabidi,M.a., Arnold,C.D., Schernhuber,K., Pagani,M., Rath,M., Frank,O. and Stark,A. (2014) Enhancer—-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, **518**, 556–559.
70. Aranda,a, Aranda,a, Pascual,a and Pascual,a (2001) Nuclear hormone receptors and gene expression. *Physiol. Rev.*, **81**, 1269–1304.
71. Pinto,C.L., Kalasekar,S.M., McCollum,C.W., Riu,A., Jonsson,P., Lopez,J., Swindell,E.C., Bouhlatouf,A., Balaguer,P., Bondesson,M. *et al.* (2016) Lxr regulates lipid metabolic and visual perception pathways during zebrafish development. *Mol. Cell. Endocrinol.*, **419**, 29–43.
72. Mears,A.J., Kondo,M., Swain,P.K., Takada,Y., Bush,R.A, Saunders,T.L., Sieving,P.A. and Swaroop,A. (2001) Nrl is required for rod photoreceptor development. *Nat. Genet.*, **29**, 447–452.
73. Kim,J.-W., Jang,S.-M., Kim,C.-H., An,J.-H. and Choi,K.-H. (2012) Transcriptional activity of neural retina leucine zipper (Nrl) is regulated by c-Jun N-terminal kinase and Tip60 during retina development. *Mol. Cell. Biol.*, **32**, 1720–1732.

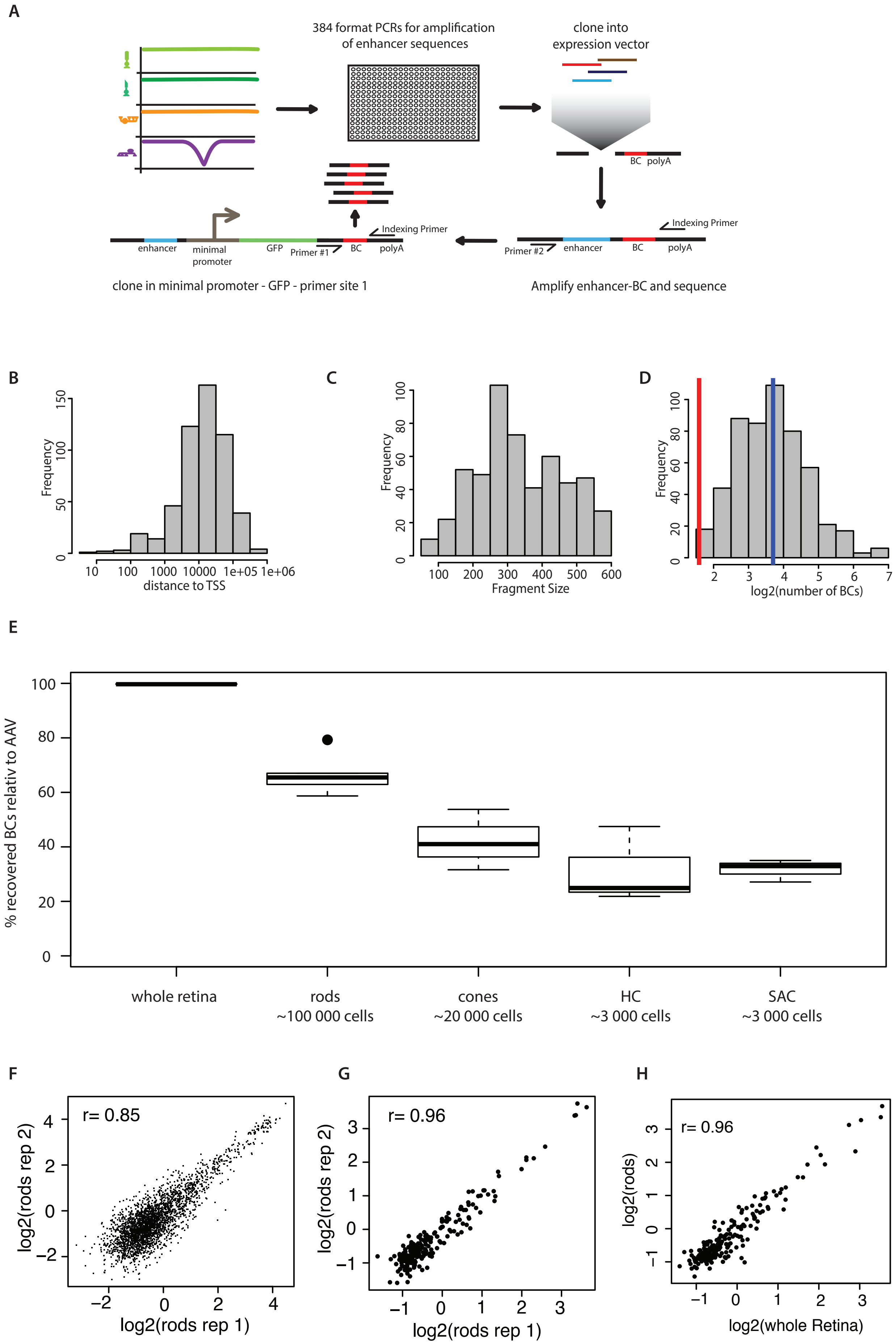
Downloaded from https://academic.oup.com/nar/article-abstract/doi/10.1093/nar/gkx923/4559116
by Novartis Forschungsstiftung / Friedrich Miescher Institute user
on 06 November 2017



Supplementary figure 1:
(A) Reproducibility of cell isolation as measured by RNA-seq. Correlation heatmap comparing RNA-seq signal for biological replicates. Pearson correlation for genic RNA-seq signal for samples issued from independent cell sorts for each cell type studied.
(B) Comparison of the generated expression profiles with photo-receptors sorted by INTACT nuclei-purification procedure. Correlation heatmap comparing RNA-seq signal obtained in this study using whole cell purification (bold labels) with previously generated nuclear RNA-seq datasets for the same cell types (Mo et al, 2016) (normal labels).



Supplementary figure 2:
(A) Comparison of the generated WGBS data with datasets previously generated by INTACT-nuclei purification procedure. Correlation heatmap comparing single CpG methylation levels at LMRs (bold labels) with previously generated datasets for retinal and neuronal cell types (Mo et al, 2015, 2016) (normal labels).
(B) Rod specific hypo-methylation is observed at the Rhodopsin gene locus. Shown is the average smoothed methylation for the indicted cell types over a 70 kb chromosomal region around the rod specific Rhodopsin gene. Read counts for accessibility measure of total retinas (ATAC-seq) is shown as smoothed signal. Black boxes denote regions having low methylation in at least one of the analyzed cell types, representing putative retinal CREs. (C-H) Variation of DNA methylation levels at LMRs across the four retina cell types. Pairwise smoothed scatter plots representing average methylation levels for LMRs in the corresponding cell types. (I) Quantitative comparison of methylation levels at LMRs with chromatin accessibility as measured by ATAC-seq. Correlation heatmap comparing average methylation levels at LMRs (bold labels) with previously generated ATAC-seq datasets in whole-retina and isolated photoreceptors (Mo et al, 2016) (normal labels). (J) Overlap between cell type-specific LMRs identified in isolated cell types with accessibility in whole-retina and photoreceptors. Heatmap depicting the percentage overlap between LMRs detected in isolated cell types (x axis) and ATAC-seq signal in whole retina or isolated photoreceptors (log2(RPKM)>0.5) (y axis) (Mo et al, 2016).


A
384 format PCRs for amplification of enhancer sequences
clone into expression vector
BC
polyA
Indexing Primer
Primer #2
enhancer
BC
polyA
Amplify enhancer-BC and sequence
Indexing Primer
enhancer
minimal promoter
GFP
Primer #1
BC
polyA
clone in minimal promoter - GFP - primer site 1
B
Frequency
0
50
100
150
10
100
1000
10000
1e+05
1e+06
distance to TSS
C
Frequency
0
20
40
60
80
100
100
200
300
400
500
600
Fragment Size
D
Frequency
0
20
40
60
80
100
2
3
4
5
6
7
log2(number of BCs)
E
% recovered BCs relativ to AAV
0
20
40
60
80
100
whole retina
rods ~100 000 cells
cones ~20 000 cells
HC ~3 000 cells
SAC ~3 000 cells
F
r= 0.85
log2(rods rep 2)
−2
0
2
4
−2
0
2
4
log2(rods rep 1)
G
r= 0.96
log2(rods rep 2)
−1
0
1
2
3
−1
0
1
2
3
log2(rods rep 1)
H
r= 0.96
log2(rods)
−1
0
1
2
3
−1
0
1
2
3
log2(whole Retina)

Supplementary figure 3:

(A) Experimental procedure used to construct the PRA libraries.

Methylation maps informed on putative CREs, fragments were PCR amplified in 384-well format and pooled. Pooled fragments were cloned into a vector containing the expression cassette consisting of a multiple cloning site, a random 15bp barcode sequence and a polyA signaling sequence (pA). In order to average out the contribution of barcode specific biases to the signal we aimed for at least ten different barcodes per unique fragment. To link CREs to barcodes the CRE-barcode sequences were amplified using Primer #2 and one of the Indexing primers (Primers #3-11) containing the Illumina flow cell annealing sequences. PCR products were purified and sequenced. Next the vector was cut between the enhancer and BC and a sequence containing a 31bp minimal promoter, CpG free eGFP and the annealing sequence for Primer #1 was cloned in. Subsequently the construct was cut out of the cloning vector into the AAV vector for virus packaging. (Also see methods)

(B) Most tested putative CREs are located kilobases away from genes. Histogram displaying distance distribution of assayed CREs to nearest transcriptional start site. (C) The tested CREs are restricted in size, not exceeding 600bp. Histogram displaying size distribution of assayed CREs.

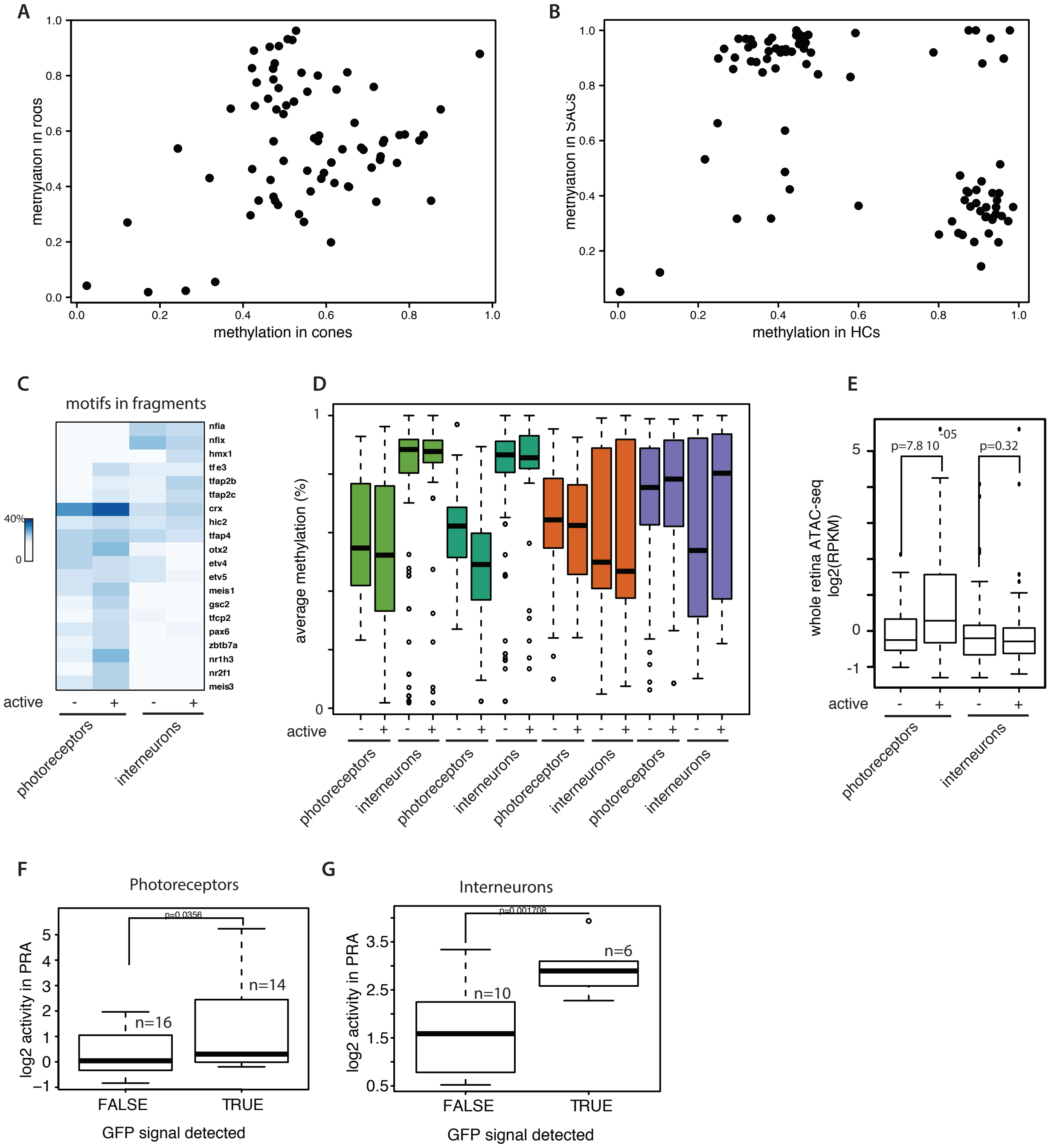
(D) Histogram displaying distribution of number of barcodes per CRE. The red line indicates cutoff for minimal number of Barcodes per CRE in the reporter assay (3 BCs), the blue line represents median BCs per CRE (13 BCs).

(E) Fraction of recovered barcodes scales the number of sampled cells and cell number tested. Boxplot displaying percentage of unique barcodes recovered in the different cell types relative to their representation in the AAV input for all cell types tested. Each box shows % recovered barcodes of all replicates for per cell type.

(F) PRA signal reproducibility at the level of barcodes in sorted rod photoreceptors. Reproducibility of barcode level activity is illustrated here by comparing two independent biological replicates.

(G) Using multiple barcodes to derive CRE activity levels enhance PRA accuracy. PRA signal reproducibility at the level of fragment in sorted rod photoreceptors. Activities of multiple barcodes (median ~13) is used to derive mean activity for each tested fragment to reducing the technical noise of the assay.

(H) The tested set of CRE shows very similar activity between rods and whole retina. Scatterplot comparing activity of CREs at the fragment level in whole Retina and rods.

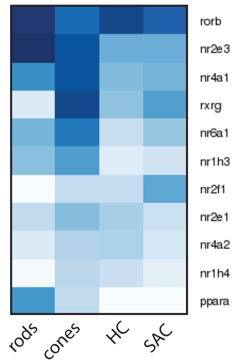


Supplementary figure 4:

(A-B) Scatterplot depicting the average methylation levels of the regions used to design the constructs in (A) library #2 and (B) library #3. Methylation is compared among photoreceptors or interneurons.

(C) Active fragments have slightly more TF binding sites than inactive ones. TF motifs enriched within fragments showing activity in photoreceptors or interneurons. Activity of a fragment was defined by PRA. Shown are motif occurrence frequencies in % for the indicated set of fragments after subtracting motif frequency in a background set of sequences. (D) Methylation levels of endogenous regions from which fragments originate are displayed as a function of their activity in the PRA. (E) Active fragments tend to have higher chromatin accessibility than inactive ones. Same as D but showing chromatin accessibility of endogenous regions relative to their activity in the PRA. ATAC-seq data from whole retina were used (Mo et al, 2016). p-values were derived using a bi-directional t-test. (F-G) Single fragment validation results are in good agreement with PRA activity measures. Boxplot depicting comparing the PRA measured activity for fragments showing or not detectable Chrd2-GFP in (F) photoreceptors (G) interneurons. The fragments where Chr2-GFP is detected by immunostaining have significantly higher PRA activity in the respective cell type. P-value was calculated using an unidirectional t-test.

A



Supplementary figure 5:

(A) Repertoire of nuclear receptors expressed in the studied cell types. Heatmap representing RPKM for nuclear receptors expressed in at least one of the studied cell type. Heatmap was organized by hierarchical clustering.

cell type	mouse strain
rods	b2-Cre x B6.Cg-Gt(ROSA)26Sortm9(CAG-tdTomato)Hze/J # 007913
cones	d4-Cre x B6.Cg-Gt(ROSA)26Sortm9(CAG-tdTomato)Hze/J # 007911
horizontal cells	B6Cf1-Tg(Gja10-Cre, #14)BR x B6.Cg-Gt(ROSA)26Sortm9(CAG-tdTomato)Hze/J # 007917
starburst amacrine cells	129S6-Chat_tm1(cre)Lowl/J x B6.Cg-Gt(ROSA)26Sortm9(CAG-tdTomato)Hze/J # 007921

Supplementary Table 1:
Table of mouse strains used to FACS sort specific cell types from the whole retina (Siegert *et al*, 2009).

chromosome	start	end	strand	ID	SACMeans	HCMeans	wholeMeans	rodMeans_lib1	coneMeans_lib2	rodsMeans_lib2	GFP signal detected in	csPRA prediction
chr16	49817994	49818185	-	144-/Intergenic2	0.1448962	0.7817183	4.1180784	3.130001537	3.193095825	2.777570411	Photoreceptors	Photoreceptors
chr17	48473681	48474023	-	217-/Trem2-Intra	NA	NA	NA	0.919568521	NA	NA	Photoreceptors	Photoreceptors
chr18	60942592	60942910	+	218+	0.2246052	0.5558612	0.4719649	0.570810657	0.547496002	-0.502445419	negative	Photoreceptors
chr4	153317537	153317650	+	259+	NA	NA	NA	-0.922875571	0.086656356	-0.962055755	negative	Photoreceptors
chr5	28472582	28472934	-	128-/Intergenic1	NA	NA	NA	2.447912426	1.585804797	1.486156309	Photoreceptors	Photoreceptors
chr7	20015162	20015346	+	262+	NA	NA	NA	-0.834708042	NA	NA	negative	negative
chr9	108433361	108433572	+	108+	NA	NA	NA	-0.234098952	NA	NA	negative	negative
chr9	64799176	64799667	-	112-	1.3081915	0.6281876	5.2348589	NA	NA	NA	Photoreceptors	Photoreceptors
chr9	98892095	98892355	-	107-/Faim-Intra	0.5212305	0.4705372	4.0082861	3.694082443	3.106615399	3.071354462	rods	Photoreceptors
chr10	75067780	75068255	+	524+	NA	NA	NA	NA	0.307958584	-0.642933092	Photoreceptors, HC, Mueller	Photoreceptors
chr10	95018573	95019092	-	283-	NA	NA	NA	NA	-0.195550916	-0.778190793	Photoreceptors	Photoreceptors
chr14	68252390	68252858	+	427+	NA	NA	NA	NA	-0.320660988	-0.740342556	Photoreceptors, Mueller	negative
chr1	64399627	64400178	-	483-	NA	NA	NA	NA	0.278933508	-1.073349105	Photoreceptors, Mueller	Photoreceptors
chr3	30669079	30669564	+	360+	NA	NA	NA	NA	0.219970671	-0.8625043	Photoreceptors, Mueller	Photoreceptors
chr3	57348332	57348893	+	506+	NA	NA	NA	NA	-0.012080873	-0.650201155	Photoreceptors, Mueller	Photoreceptors
chr5	53662011	53662512	-	357-	NA	NA	NA	NA	-0.161178677	-0.582074305	Photoreceptors, Mueller	Photoreceptors
chr7	87193812	87194240	-	330-	NA	NA	NA	NA	-0.060127345	-1.13118179	Photoreceptors, Mueller	Photoreceptors
chr8	49596516	49597041	+	359+	NA	NA	NA	NA	0.302115742	-0.739830871	Photoreceptors, ganglion cells	Photoreceptors
chr10	114602962	114603205	-	3274-	0.57325865	2.24826702	-0.342795558	NA	NA	NA	Mueller	Interneurons
chr12	75335993	75336295	+	3285+	3.33951396	2.82205884	1.1984568	NA	NA	NA	negative	Interneurons
chr1	38455370	38455657	-	3192-	2.5687756	2.73769384	0.900204402	NA	NA	NA	Mueller, Glia cells	Interneurons
chr3	131727036	131727461	+	3010+	1.65020139	0.56919204	-0.658386556	NA	NA	NA	ganglion cells	negative
chr4	100706634	100706916	-	3046-	1.55601072	1.34055942	-0.000979624	NA	NA	NA	negative	negative
chr4	103956384	103956713	-	3215-	2.99481611	2.69555303	1.342116548	NA	NA	NA	Mueller + HC	Interneurons
chr4	130912296	130912598	-	3024-/Intergenic4	3.09691086	3.01884932	1.965822194	NA	NA	NA	HC + Amacrine cells	Interneurons
chr4	40994591	40995016	-	3326-/Aqp7-Intra	2.09420918	2.58195659	0.293942264	NA	NA	NA	Amacrine cells + HC + Mueller	Interneurons
chr6	127582881	127583163	-	3009-	1.61731687	0.31911029	-0.614659127	NA	NA	NA	negative	negative
chr6	48361745	48362148	+	3149+	3.93440563	3.79679328	1.521460404	NA	NA	NA	Amacrine cells	Interneurons
chr9	31656883	31657191	-	3259-	2.75159521	2.79424503	1.151503338	NA	NA	NA	Amacrine cells + Photoreceptors	Interneurons
chr9	50142608	50142868	+	3260+	1.48726742	2.2763465	-0.191911937	NA	NA	NA	HC	Interneurons

Supplementary Table 2:
Table of individually tested enhancers and their activity pattern in the whole retina and log2 activity in csPRA. Second last column shows in which cell types GFP signal was detected under the microscope when single CREs were tested. 'negative' means no GFP signal could be found. The last column shows in which cell type CREs are expected to be active in csPRA. 'negative' means that activity is below threshold in csPRA.

Sequences of mutated CREs:

>107_WT
GAGGCTTTCACTGACCTTTCCATGTACGAGACTAGCCCCGCCCAGAGTGCTGTCTGGGATTAACTGTTACAGTCT
TAATTGTATTAGTATTTAGGTGACTTTGGGATTATTAGACTATTCCTGTGCATAGCTGCCTCTTGAGGGGAGAGC
CGGGAGAGAGAAGCAGCTGCATGTGCTGTGAGTAGGACATCTGGGGGCATCACTTTACCATCTCATAGTTCTA
GGGCCCTTGAAAACCTGGTATTGTCATGCTGGAAGGGT
>107_all_MEIS1_mut
GAGGCTTTCACTGACCTTTCCATGTACGAGACTAGCCCCGCCCAGAGTGCTGTCTGGGATTAACTGTTACAGTCT
TAATTGTATTAGTATTTAGGTGACTTTGGGATTATTAGACTATTCCTGTGCATAGCTGCCTCTTGAGGGGAGAGC
CGGGAGAGAGAAGCAGCTGCATGTGCTGTGAGTAGGACATCTGGGGGCATCACTTTACCATCTCATAGTTCTA
GGGCCCTTGAAAACCTGGTTTCTAGTTGCTGGAAGGGT
>107_all_SP1_mut
GAGGCTTTCACTGACCTTTCCATGTACGAGACTAAGGTAGTACAGAGTGCTGTCTGGGATTAACTGTTACAGTC
TTAATTGTATTAGTATTTAGGTGACTTTGGGATTATTAGACTATTCCTGTGCATAGCTGCCTCTTGAGGGGAGAG
CCGGGAGAGAGAAGCAGCTGCATGTGCTGTGAGTAGGACATCTGGGGGCATCACTTTACCATCTCATAGTTCT
AGGGCCCTTGAAAACCTGGTATTGTCATGCTGGAAGGGT
>107_all_ID4_mut
GAGGCTTTCACTGACCTTTCCATGTACGAGACTAGCCCCGCCCAGAGTGCTGTCTGGGATTAACTGTTCAGTCTT
AATTGTATTAGTATTTAGGTGACTTTGGGATTATTAGACTATTCCTGTGCATAGCTGCCTCTTGAGGGGAGAGCC
GGGAGAGAGACGCGATAGCATGTGCTGTGAGTAGGACATCTGGGGGCATCACTTTACCATCTCATAGTTCTAG
GGCCCTTGAAAACCTGGTATTGTCATGCTGGAAGGGT
>107_all_Nr1h3_mut
GAGGCTTTCATTACAATTTCCATGTACGAGACTAGCCCCGCCCAGAGTGCTGTCTGGGATTAACTGTTACAGTCT
TAATTGTATTAGTATTTAGGTGACTTTGGGATTATTAGACTATTCCTGTGCATAGCTGCCTCTTGAGGGGAGAGC
CGGGAGAGAGAAGCAGCTGCATGTGCTGTGAGTAGGACATCTGGGGGCATCACTTTACCATCTCATAGTTCTA
GGGCCCTTGAAAACCTGGTATTGTCATGCTGGAAGGGT
>483_WT
CGGGGCAGGAGTGAGTCATTCACACAGAGGCGGGTCAGAGAGTAGTGGTCACTCTTCAGCTTACAGCTCTCTC
TAGTCCTCTATCCAGACTCTAGTTTCATGAACTTTGTAGTTAGACATTTTTCCTAGTGAATATTTATTACCCCCCAC
TGTAATCCTTCATTTAACATAATATAAAATTTGTGGAAAGGGAGAGTAATTAGTAATAAATCATCATCTCATCCA
TTAGCAGTAAATAATGCCACTTTATCAAAGTCACAGCCATCGAACAGGCGGCTAGAGGTGGTTATGTATGCCAC
CCGACTGGAAGCAGGCCAAAAGCAAACCGCAGCCCCCGTTTATTATCCTAATTATGCCCTAATACGATGCCATC
TTTTTCTCCTATAAACTTGATGACAATAAAAGGGTAACAATGAAAATTGGCAGGGTAAGTGAGCAAGGAAGAT
AGGCTGGGAAACCACCTAGCCCCACCGGCTACCA
GCCTGAGTCCTGAGGCTGAAAGGGCTGAAAACCCCATGGGAATGAAATGGAGCAGGGGACTCAAGTGGTTGG
>483_all_MEIS1_mut
CGGGGCAGGAGTGAGTCATTCACACAGAGGCGGGTCAGAGAGTAGTGGTCACTCTTCAGCTTACAGCTCTCTC
TAGTCCTCTATCCAGACTCTAGTTTCATGAACTTTGTAGTTAGACATTTTTCCTAGTGAATATTTATTACCCCCCAC
TGTTCTAGTTCATTTAACATAATATAAAATTTGTGGAAAGGGAGAGTAATTAGTAATAAATCATCATCTCATCCA
TTAGCAGTAAATAATGCCACTTTATCAAAGTCACAGCCATCGAACAGGCGGCTAGAGGTGGTTATGTATGCCAC
CCGACTGGAAGCAGGCCAAAAGCAAACCGCAGCCCCCGTTTATTATCCTAATTATGCCCTAATACGATGCCATC
TTTTTCTCCTATAAACTTGATGACAATAAAAGGGTAACAATGAAAATTGGCAGGGTAAGTGAGCAAGGAAGAT
AGGCTGGGAAACCACCTAGCCCCACCGGCTACCA
GCCTGAGTCCTGAGGCTGAAAGGGCTGAAAACCCCATGGGAATGAAATGGAGCAGGGGACTCAAGTGGTTGG
>483_all_CRX_mut
CGGGGCAGGAGTGAGTCATTCACACAGAGGCGGGTCAGAGAGTAGTGGTCACTCTTCAGCTTACAGCTCTCTC
TAGTCCTCTATCCAGACTCTAGTTTCATGAACTTTGTAGTTAGACATTTTTCCTAGTGAATATTTATTACCCCCCAC
TGTAATCCTTCATTTAACATAATATAAAATTTGTGGAAAGGGAGAGTAATTAGTAATAAATCATCATCTCATCCA
TTAGCAGTAAATAATGCCACTTTATCAAAGTCACAGCCATCGAACAGGCGGCTAGAGGTGGTTATGTATGCCAC
CCGACTGGAAGCAGGCCAAAAGCAAACCGCAGCCCCCGTTTATTATCCTAATTATGCCCTAATACGATGCCATC
TTTTTCTCCTATAAACTTGCAGGTGTTAAAAGGGTAACAATGAAAATTGGCAGGGTAAGTGAGCAAGGAAGAT
AGGCTGGGAAACCACCTAGCCCCACCGGCTACCA
GCCTGAGTCCTGAGGCTGAAAGGGCTGAAAACCCCATGGGAATGAAATGGAGCAGGGGACTCAAGTGGTTGG
>483_all_NRL_mut

```
CGGGGCAGGAGTGAGTCATTCACACAGAGGCGGGTCAGAGAGTAGTGGTCACTCTTCAGCTTACAGCTCTCTC
TAGTCCTCTATCCAGACTCTAGTTTCATGAACTTTGTAGTTAGACATTTTTCCTAGTGAATATTTATTACCCCCCAC
TGTAAGCGAACGTAGAACATAATATAAAATTTGTGGAAAGGGAGAGTAATTAGTAATAAATCATCATCTCATCC
ATTAGCAGTAAATAATGCCACTTTATCAAAGTCACAGCCATCGAACAGGCGGCTAGAGGTGGTTATGTATGCCA
CCCGACTGGAAGCAGGCCAAAAGCAAACCGCAGCCCCCGTTTATTATCCTAATTATGCCCTAATACGATGCCAT
CTTTTTCTCCTATAAACTTGATGACAATAAAAGGGTAACAATGAAAATTGGCAGGGTAAGTGAGCAAGGAAGA
TAGGCTGGGAAACCACCTAGCCCCACCGGCTACCA
GCCTGAGTCCTGAGGCTGAAAGGGCTGAAAACCCCATGGGAATGAAATGGAGCAGGGGACTCAAGTGGTTGG
>483_all_En2_mut
CGGGGCAGGAGTGAGTCATTCACACAGAGGCGGGTCAGAGAGTAGTGGTCACTCTTCAGCTTACAGCTCTCTC
TAGTCCTCTATCCAGACTCTAGTTTCATGAACTTTGTAGTTAGACATTTTTCCTAGTGAATATTTATTACCCCCCAC
TGTAATCCTTCATTTAACATAATATAAAATTTGTGGAAAGGGAGAGCGGAGTAAACTAAATCATCATCTCATCCA
TTAGCAGTAAATAATGCCACTTTATCAAAGTCACAGCCATCGAACAGGCGGCTAGAGGTGGTTATGTATGCCAC
CCGACTGGAAGCAGGCCAAAAGCAAACCGCAGCCCCCGTTTATTATCCCGGAGTAAACCTAATACGATGCCATC
TTTTTCTCCTATAAACTTGATGACAATAAAAGGGTAACAATGAAAATTGGCAGGGTAAGTGAGCAAGGAAGAT
AGGCTGGGAAACCACCTAGCCCCACCGGCTACCA
GCCTGAGTCCTGAGGCTGAAAGGGCTGAAAACCCCATGGGAATGAAATGGAGCAGGGGACTCAAGTGGTTGG
```

Primers used for csPRA library preparation:

Name	Sequence
Primer #1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Primer #2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCACTGGGAGAAGAGGAAGTCAAA

Indexing Primers:

Name	Sequence	Index
Primer #3	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	1
Primer #4	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	2
Primer #5	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	6
Primer #6	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	12
Primer #7	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	4
Primer #8	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	7
Primer #9	CAAGCAGAAGACGGCATACGAGATGACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	5
Primer #10	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	8
Primer #11	CAAGCAGAAGACGGCATACGAGATGTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	9

References:

1. Siegert, S. *et al.* Genetic address book for retinal cell types. *Nat. Neurosci.* **12,** 1197–1204 (2009).
2. Siegert, S. *et al.* Transcriptional code and disease map for adult retinal cell types. *Nat. Neurosci.* **15,** 487–95, S1-2 (2012).
3. Mo, A. *et al.* Epigenomic landscapes of retinal rods and cones. *Elife* **5,** 1–29 (2016).
4. Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* **86,** 1369–1384 (2015).

4 General discussion and Conclusions

Within this thesis design principles of promoter and enhancer activity in mammalian genomes were investigated using the mouse as a model system. This has been achieved through the development of an advanced reporter assay for transcriptional activity. This allowed to test the activity of regulatory regions integrated into genomic DNA or within specific, low abundant cell types in the retina.

In the first study, we show that CpG density contributes to CGI promoter activity in combination with the binding of specific TFs. This finding is likely to be generalizable to other vertebrates and their CGI promoters.

The second project provided proof of concept for how active regulatory regions can be identified and tested for their transcriptional activity in high-throughput in individual cell types of the retina. This strategy has the potential to be used for other cell types and tissues in a variety of organisms.

4.1 Design principles of CpG island promoter activity

CGIs were identified nearly three decades ago based on their lack of DNA methylation (Bird *et al*, 1985). Since then, their high CpG density has been linked to active chromatin and transcriptional activity by correlative means (Weber *et al*, 2007; Guenther *et al*, 2007; Thomson *et al*, 2010; Deaton & Bird, 2011; Fenouil *et al*, 2012). Recently, attempts were taken to functionally assay the role of CpGs in CGIs (Lienert *et al*, 2011; Krebs *et al*, 2014; Wachter *et al*, 2014). However, to what extend and how high CpG density contributes to transcriptional activity of CGI promoters remained unclear.

Our initial computational analysis showed that CpG density and motif occurrence perform well in predicting binding of specific TFs. However, the non-uniform distribution of CpGs within mammalian genomes and the lack of CpG dense regions that are not under functional selection makes it inherently difficult to assign a functional role to CpGs. To circumvent this problem, we assayed a large number of promoter mutants for their transcriptional activity. These assays show that CpG density does indeed functionally contribute to transcriptional activity of CGI promoters in combination with the presence of more complex TF motifs.

Our data indicate that the CpG density required for transcriptional activity is around 0.6 (OE) or higher which we observed in most promoter mutants across all experiments. This value is surprisingly similar to the currently used CGI definition. The same threshold can also be deduced from ChIP-seq data of four TFs that were mapped genome-wide within this study. More motifs are occupied when the CpG density is >0.6 while much fewer are bound at lower densities. We speculate that these differences are linked to accessibility of the regions. In agreement with this, accessibility increases with CpG density. This is consistent with published data showing that CpG and GC rich artificial sequences display marks of accessible chromatin (Wachter *et al*, 2014; Lienert *et al*, 2011; Krebs *et al*, 2014).

The putative link between CpG density and accessibility raises the question if accessibility decreases upon CpG depletion in mutant constructs with reduced CpG density. To answer this question, changes in accessibility could be tested on the mutant CpG density constructs using NOMe-seq (Kelly *et al*, 2012; Nabilsi *et al*, 2014; Krebs *et al*, 2017). Additionally, histone modifications could play a role in changes in activity observed upon mutation of CpGs. For example, it was shown that CGI promoters are depleted in H3K36me2, catalyzed by the ZF-CxxC domain containing histone demethylase KDM2A (Blackledge *et al*, 2010). One potential scenario could be that unmethylated CpGs mediate demethylation of H3K36me2 at the promoter, thus preventing H3K36me2 from interfering with transcriptional initiation. This model could be further tested by performing ChIP on individual CpG mutant constructs.

In non-transformed cells, DNA methylation does not occur at CpG rich DNA sequences but only at those with lower CpG density (Lienert *et al*, 2011; Krebs *et al*, 2014). Thus, we wondered how DNA methylation would affect the transcriptional activity of CpG density mutant promoters. Our results argue that removal of DNA methyltransferases leads to increased activity of mutants with low CpG densities. One might speculate that high CpG density is required at CGIs to prevent DNA methylation which would allow binding of methylation sensitive TFs. A protective function of high CpG density against DNA methylation could also explain why high CpG density together with motif occurrence is such a good predictor of TF binding.

Taken together our results suggest that high CpG density at CGIs is not only a relict of evolution due to its unmethylated state in the germline. Rather, CpGs are required for high transcriptional activity of CGI promoters.

4.2 *Cis*-regulatory landscape of four cell types of the retina

The identification of *cis*-regulatory elements is essential in order to understand the transcriptional regulatory principles that control cell-type specification. Using transcriptome and epigenome profiling, we identified a large collection of putative *cis*-regulatory elements active in four distinct cell types of the retina. Additionally, we provide proof of concept for the usage of parallel reporter assays to measure the autonomous transcriptional activity of regulatory regions in individual retinal cell types. We successfully applied this assay to quantify activity of hundreds of short DNA sequences in four individual cell types. This effort allowed us to identify a small set of short sequences that are preferentially active in different cellular subsets of the retina. Additionally, we demonstrated how this technology can be employed for dissection of the architecture of regulatory regions *in vivo.*

Parallel reporter assays have been previously applied *in vivo* but without discriminating between specific cell types that make up a tissue (Gisselbrech *et al*, 2014; Shen *et al*, 2015; Patwardhan *et al*, 2012). We used FACS-sorting on a library of mouse lines with various fluorescently labeled retinal cell types enabling reproducible isolation of pure cell populations (Siegert *et al*, 2009, 2012). However, this approach is inherently constrained by the number of cells available, which is limiting in the case of rare cell types. To circumvent these bottlenecks and to derive accurate parallel reporter assay measures for low cell numbers, we combined high efficiency Adeno-associated virus-based delivery of our libraries with multiple measures for each fragment. Contrasting whole tissue with cell-type specific data demonstrated that whole tissue data only reflects sequence activity in photoreceptors, which is the dominant cell type in mouse retina. Consequently, analysis on whole tissue failed to capture activities in rare cell types, demonstrating that cell type isolation and activity assignment is critical.

Besides the potential to advance our understanding of regulatory principles, the introduced experimental strategy is capable to identify candidate regulatory elements controlling transgene activity for gene therapy. For several inherited retinal diseases, gene-replacement or targeted expression of transgenes is considered a credible strategy to at least partially restore vision (Boye *et al*, 2013; Nash *et al*, 2015; Sahel & Roska, 2013). Ideally such regulatory elements are short and specifically active in the target cell type. We determined the expression pattern of putative regulatory regions within four cell types leading to a catalogue of short sequences that drive expression in the mouse retina. This screen allowed us to identify a small set of sequences that are preferentially expressed in different cellular subtypes of the retina, but not uniquely in one cell type. Targeting of disease relevant cell-types of the retina specifically, will require a better understanding of the sequence features defining cell type specific expression. Therefore, further screening in more cell types is required.

Additionally, we demonstrated that mutation of TF motifs can be utilised to tune transcriptional activity of identified *cis*-regulatory elements. Such mutations could be utilised to obtain accurate expression levels of the transgene for restoration of cellular function.

Taken together, the work presented here advances our knowledge about location and regulation of regulatory regions that function in specialised cell types and also provides insight into the regulation of CGI promoters that tend to be ubiquitously expressed.

5 References

Acland GM, Aguirre GD, Ray J, Zhang Q, Aleman TS, Cideciyan A V, Pearce-kelling SE, Anand V, Zeng Y, Maguire AM, Jacobson SG, William W, Bennett J, Acland GM, Aguirre GD, Ray J, Zhang Q, Aleman TS, Cideciyan A V, Pearce-kelling SE, et al (2001) Gene therapy restores vision in a canine model of childhood blindness. *Nat Genet* **28:** 92–95

Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H & Shiroishi T (2009) Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* **16:** 47–57

Antequera F & Bird AP (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* **90:** 11995–9

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC & Stark A (2014) Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat. Genet.*

Arnone MI & Davidson EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124:** 1851–1864

Bainbridge JWB, Tan MH & Ali RR (2006) Gene therapy progress and prospects: the eye. *Gene Ther.* **13:** 1191–1197

Bainbridge JWBB, Smith AJ, Barker SS, Robbie S, Henderson R, Balaggan K, Viswanathan A, Holder GE, Stockman A, Tyler N, Petersen-Jones S, Bhattacharya SS, Thrasher AJ, Fitzke FW, Carter BJ, Rubin GS, Moore AT & Ali RR (2008) Effect of gene therapy on visual function in Leber's congenital amaurosis. *N. Engl. J. Med.* **358:** 2231–2239

Bajic VB, Tan SL, Christoffels A, Schönbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, Kawai J, Hume DA, Carninci P & Hayashizaki Y (2006) Mice and men: their promoter properties. *PLoS Genet.* **2:** e54

Banerji J, Olson L & Schaffner W (1983) A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33:** 729–740

Banerji J, Rusconi S & Schaffner W (1981) Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27:** 299–308

Barnes DE & Lindahl T (2004) Repair and genetic consequences of endogenous

DNA base damage in mammalian cells. *Annu. Rev. Genet.* **38:** 445–76

Barth TK & Imhof A (2010) Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem. Sci.* **35:** 618–626

Baubec T, Ivánek R, Lienert F & Schübeler D (2013) Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. *Cell* **153:** 480–92

Benoist C & Chambon P (1981) In vivo sequence requirements of the SV40 early promotor region. *Nature* **290:** 304–310

Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21:** 611–26

Bird A (2011) The dinucleotide CG as a genomic signalling module. *J. Mol. Biol.* **409:** 47–53

Bird A, Taggart M, Frommer M, Miller OJ & Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40:** 91–99

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8:** 1499–1504

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ & Klose RJ (2010) CpG islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell* **38:** 179–90

Blattner FR (1997) The Complete Genome Sequence of Escherichia coli K-12. *Science (80-.).* **277:** 1453–1462

Boulard M, Edwards JR & Bestor TH (2015) FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat. Genet.* **47:** 1–9

Boye SE, Boye SL, Lewin AS & Hauswirth WW (2013) A comprehensive review of retinal gene therapy. *Mol. Ther.* **21:** 509–19

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A & Cedar H (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371:** 435–8

Buenrostro JD, Wu B, Chang HY & Greenleaf WJ (2015) ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015:** 21.29.1-21.29.9

Butler JEF & Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16:** 2583–92

Calo E & Wysocka J (2013) Modification of Enhancer Chromatin: What, How, and Why? *Mol. Cell* **49:** 825–837

Campanero MR, Armstrong MI & Flemington EK (2000) CpG methylation as a mechanism for the regulation of E2F activity. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 6481–6486

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple C a M, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, et al (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38:** 626–35

Cedar H (1988) DNA methylation and gene activity. *Cell* **53:** 3–4

Chen T, Ueda Y, Dodge JE, Wang Z & Li E (2003) Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* **23:** 5594–605

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee JH, Skalnik D & Bird A (2012) Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev.* **26:** 1714–1728

Cohen NM, Kenigsberg E & Tanay A (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145:** 773–86

Courey AJ & Tjian R (1988) Analysis of Sp1 in vivo reveals mutiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* **55:** 887–898

Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen YD, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG & Collins FS (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16:** 123–131

Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev.* **25:** 1010–22

Domcke S, Bardet AF, Ginno PA, Hartl D, Burger L & Schübeler D (2015) Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528:** 575–579

Emili A, Greenblatt J & Ingles CJ (1994) Species-specific interaction of the glutamine-rich activation domains of Sp1 with the TATA box-binding protein. *Mol. Cell. Biol.* **14:** 1582–93

ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M & Bernstein BE (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–9

Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I & Andrau J-C (2012) CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* **22:** 2399–408

Flanagan JF, Mi L-Z, Chruszcz M, Cymborowski M, Clines KL, Kim Y, Minor W, Rastinejad F & Khorasanizadeh S (2005) Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* **438:** 1181–5

Gardiner-Garden M & Frommer M (1987) CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–282

Gill G, Pascal E, Tseng ZH & Tjian R (1994) A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc. Natl. Acad. Sci. U. S. A.* **91:** 192–6

Gisselbrech SS, Barrera LA, Porsch M, Aboukhalil A, Estep III PW, Vedenko A, Palagi A, Kim Y, Zhu X, Busser BW, Gamble CE, Iagovitina A, Singhania A, Michelson AM & Bulyk ML (2014) Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Igarss 2014* **10:** 1–5

Grosschedl R & Birnstiel ML (1980a) Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **77:** 1432–6

Grosschedl R & Birnstiel ML (1980b) Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **77:** 7102–6

Guenther MG, Levine SS, Boyer LA, Jaenisch R & Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88

Haberle V & Lenhard B (2016) Promoter architectures and developmental gene regulation. *Semin. Cell Dev. Biol.* **57:** 11–23

Hah N, Murakami S, Nagari A, Danko CG & Kraus WL (2013) Enhancer transcripts mark active estrogen receptor binding sites Enhancer transcripts mark active estrogen receptor binding sites. **:** 1210–1223

Han L, Su B, Li W-H & Zhao Z (2008) CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol.* **9:** R79

Hauswirth WW, Aleman TS, Kaushal S, Cideciyan A V, Schwartz SB, Wang L, Conlon TJ, Boye SL, Flotte TR, Byrne BJ & Jacobson SG (2008) Treatment of leber congenital amaurosis due to RPE65 mutations by ocular subretinal injection of adeno-associated virus gene vector: short-term results of a phase I trial. *Hum. Gene Ther.* **19:** 979–90

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov V V, Stewart R, Thomson J a, Crawford GE, Kellis M, et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108–112

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE & Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39:** 311–8

Hendrich B & Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18:** 6538–47

Hendrich B, Hardeland U, Ng HH, Jiricny J & Bird A (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401:** 301–304

Hendrich B & Tweedie S (2003) The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19:** 269–277

Hermann A, Gowher H & Jeltsch A (2004) Biochemistry and biology of mammalian DNA methyltransferases. *Cell. Mol. Life Sci.* **61:** 2571–2587

Ho KL, McNae IW, Schmiedeberg L, Klose RJ, Bird AP & Walkinshaw MD (2008) MeCP2 Binding to DNA Depends upon Hydration at Methyl-CpG. *Mol. Cell* **29:** 525–531

Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR, Smith AD & Hannon GJ (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell* **44:** 17–28

Iguchiariga SMM & Schaffner W (1989) CpG Methylation of the Camp-Responsive Enhancer Promoter Sequence TGACGTCA Abolishes Specific Factor Binding As Well As Transcriptional Activation. *Genes Dev.* **3:** 612–619

Ioshikhes IP & Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **26:** 61–3

Jacob F & Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3:** 318–56

John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL & Stamatoyannopoulos JA (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43:** 264–8

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E & Kivioja T (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell* **152:** 327–339

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E & Taipale J (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527:** 384–8

Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13:** 484–92

Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116:** 247–57

Kelly TK, Liu Y, Lay FD, Liang G, Berman BP & Jones PA (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22:** 2497–506

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito

H, Worley PF, Kreiman G & Greenberg ME (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182–7

Kornberg RD & Thomas JO (1974) Chromatin Structure : Oligomers of the Histones. *Science (80-.).* **184:** 865–868

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128:** 693–705

Krebs A, Dessus-Babus S, Burger L & Schübeler D (2014) High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife* **3:** 1–18

Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L & Schübeler D (2017) Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell***:** 1–12

Kwasnieski JC, Mogno I, Myers CA, Corbo JC & Cohen BA (2012) Complex effects of nucleotide variants in a mammalian cis -regulatory element.

de la Serna IL, Ohkawa Y, Berkes C a, Bergstrom D a, Dacwag CS, Tapscott SJ & Imbalzano AN (2005) MyoD targets chromatin remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex. *Mol. Cell. Biol.* **25:** 3997–4009

Larsen F, Gundersen G, Lopez R & Prydz H (1992) CpG islands as gene markers in the human genome. *Genomics* **13:** 1095–1107

Levo M, Avnit-Sagi T, Lotan-Pompan M, Kalma Y, Weinberger A, Yakhini Z & Segal E (2017) Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. *Mol. Cell* **65:** 604–617.e6

Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD & Patel DJ (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442:** 91–95

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F & Schübeler D (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **43:** 1091–7

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar a H, Thomson J a, Ren B & Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315–22

Long HK, Blackledge NP & Klose RJ (2013) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.* **41:** 727–40

Loyola A, LeRoy G, Wang YH & Reinberg D (2001) Reconstitution of recombinant chromatin establishes a requirement for histone-tail modifications during chromatin assembly and transcription. *Genes Dev.* **15:** 2837–2851

Macleod D, Charlton J, Mullins J & Bird AP (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8:** 2282–2292

Maguire AM, Simonelli F, Pierce EA, Pugh EN, Mingozzi F, Bennicelli J, Banfi S, Marshall KA, Testa F, Surace EM, Rossi S, Lyubarsky A, Arruda VR, Konkle B, Stone E, Sun J, Jacobs J, Dell'Osso, Hertle R, Ma J, et al (2008) Safety and Efficacy of Gene Transfer for Leber{\textquoteright}s Congenital Amaurosis. *N. Engl. J. Med.* **358:** 2240

Masland RH (2001) The fundamental plan of the retina. *Nat. Neurosci.* **4:** 877–886

Maston GA, Evans SK & Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7:** 29–59

Meehan RR, Lewis JD, McKay S, Kleiner EL & Bird AP (1989) Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58:** 499–507

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, Kellis M, Lander ES & Mikkelsen TS (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30:** 271–7

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560

Millar CB (2002) Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice. *Science (80-.).* **297:** 403–405

Miller JA & Widom J (2003) Collaborative Competition Mechanism for Gene Activation In Vivo Collaborative Competition Mechanism for Gene Activation In Vivo. **23:** 1623–1632

Mogno I, Kwasnieski JC & Cohen BA (2013) Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.*

Mohn F & Schübeler D (2009) Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.* **25:** 129–36

Müller H-P & Schaffner W (1990) Transcriptional enhancers can act in trans. *Trends Genet.* **6:** 300–304

Myers RM, Tilly K & Maniatis T (1986) Fine structure genetic analysis of a beta-globin promoter. *Science (80-.).* **232:** 613–618

Nabilsi NH, Deleyrolle LP, Darst RP, Riva A, Reynolds BA & Kladde MP (2014) Multiplex mapping of chromatin accessibility and DNA methylation within targeted single molecules identifies epigenetic heterogeneity in neural stem cells and glioblastoma. *Genome Res.* **24:** 329–339

Nan X, Campoy FJ & Bird A (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88:** 471–481

Nash BM, Wright DC, Grigg JR, Bennetts B & Jamieson R V (2015) Retinal dystrophies, genomic applications in diagnosis and prospects for therapy. *Transl. Pediatr.* **4:** 139–163

Neddermann P & Jiricny J (1993) The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells. *J. Biol. Chem.* **268:** 21218–21224

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, et al (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489:** 83–90

Ohki I, Shimotake N, Fujita N, Jee JG, Ikegami T, Nakao M & Shirakawa M (2001) Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* **105:** 487–497

Oka S, Shiraishi Y, Yoshida T, Ohkubo T, Sugiura Y & Kobayashi Y (2004) NMR structure of transcription factor Sp1 DNA binding domain. *Biochemistry* **43:**

16027–35

Palazzo AF & Gregory TR (2014) The Case for Junk DNA. *PLoS Genet.* **10:**

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, Ahituv N, Pennacchio L a & Shendure J (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30:** 265–70

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D & Shendure J (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27:** 1173–5

Prendergast GC & Ziff EB (1991) Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science* **251:** 186–9

Reiter F, Wienerroither S & Stark A (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* **43:** 73–81

Richmond TJ & Davey CA (2003) The structure of DNA in the nucleosome core. *Nature* **423:** 145–150

Roosing S, Thiadens AAHJ, Hoyng CB, Klaver CCW, den Hollander AI & Cremers FPM (2014) Causes and consequences of inherited cone disorders. *Prog. Retin. Eye Res.* **42:** 1–26

Sahel J-AA & Roska B (2013) Gene therapy for blindness. *Annu Rev Neurosci* **36:** 467–488

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y & Hume D a (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* **8:** 424–36

Saxonov S, Berg P & Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* **103:** 1412–7

Schones D, Cui K, Cuddapah S, Roh T, Barski A, Wang Z, Wei G & Zhao K (2008) Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132:** 887–898

Schübeler D (2015) Function and information content of DNA methylation.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A & Segal E (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30:** 521–30

Shen J cheng, Rideout WM & Jones PA (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22:** 972–976

Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG & Corbo JC (2015) Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.***:** 1–18

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov V V & Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* **488:** 116–20

Shlyueva D, Stampfel G & Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*

Siegert S, Cabuy E, Scherf BG, Kohler H, Panda S, Le Y-Z, Fehling HJ, Gaidatzis D, Stadler MB & Roska B (2012) Transcriptional code and disease map for adult retinal cell types. *Nat. Neurosci.* **15:** 487–95, S1-2

Siegert S, Scherf BG, Del Punta K, Didkovsky N, Heintz N & Roska B (2009) Genetic address book for retinal cell types. *Nat. Neurosci.* **12:** 1197–1204

Smallwood A & Ren B (2013) Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.* **25:** 387–394

Smith E & Shilatifard A (2014) Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.* **21:** 210–219

Song J, Teplova M, Ishibe-Murakami S & Patel DJ (2012) Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science* **335:** 709–12

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK & Schübeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490–5

Suzuki MM, Kerr ARW, De Sousa D & Bird A (2007) CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res.* **17:** 625–631

Svaren J & Hörz W (1997) Transcription factors vs nucleosomes: Regulation of the PHO5 promoter in yeast. *Trends Biochem. Sci.* **22:** 93–97

Swaroop A, Kim D & Forrest D (2010) Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nat.*

Rev. Neurosci. **11:** 563–576

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L & Rao A (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324:** 930–5

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R & Bird A (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082–6

Thurman R, Rynes E, Humbert R, Vierstra J, Maurano M, Haugen E, Sheffield N, Stergachis A, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield T, Diegel M, Dunn D, Ebersol A, Frum T, et al (2012) The accessible chromatin landscape of the human genome. *Nature* **489:** 75–82

Tsukada Y, Fang J, Erdjument-Bromage H, Warren ME, Borchers CH, Tempst P & Zhang Y (2006) Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439:** 811–816

Venter JC, Adams MD, Myers EW, Li PW, RJ M, Sutton GG, Smith HO, Yandell M, A EC, Holt RA, Gocayne, Jeannine D Amanatides P, Ballew RM, Huson DH, Wortman JR, MEW, Zhang Q Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder KCD, N Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, LAJ, Mobarry C Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di RK, Francesco V Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei DP, YD Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun LZY, JT Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, Yao A, Ye J, Zhan M, Zhang WQ, Zhang HY, Zhao Q, Zheng LS, Zhong F, Zhong WY, Zhu WZY, SPC Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An HJ, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, ZSY, Carver A Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, CA, et al (2001) The Sequence of the Human Genome. *Science (80-.).* **291:** 1304–1351

Voo KS, Carlone DL, Jacobsen BM, Flodin A & Skalnik DG (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol.Cell Biol.* **20:** 2108–2121

Voss TC & Hager GL (2013) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.* **15:** 69–81

Voss TC, Schiltz RL, Sung M-H, Yen PM, Stamatoyannopoulos J a, Biddie SC, Johnson T a, Miranda TB, John S & Hager GL (2011) Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell* **146:** 544–54

Wachter E, Quante T, Merusi C, Arczewska A, Stewart F, Webb S & Bird A (2014) Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife* **3:** 1–16

Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG & Fu X-D (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474:** 390–394

Watt F & Molloy PL (1988) Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev.* **2:** 1136–1143

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M & Schübeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39:** 457–66

White MA, Myers CA, Corbo JC & Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis - regulatory function of ChIP-seq peaks.

Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, Wu C & Allis CD (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling.

Nature **442:** 86–90

Xu J & Smale ST (2012) Designing an enhancer landscape. *Cell* **151:** 929–931

Yoshida K, Watanabe D, Ishikane H, Tachibana M, Pastan I & Nakanishi S (2001) A key role of starburst amacrine cells in originating retinal directional selectivity and optokinetic eye movement. *Neuron* **30:** 771–780

Yun M, Wu J, Workman JL & Li B (2011) Readers of histone modifications. *Cell Res.* **21:** 564–78

Zhang Y, Ng HH, Erdjument-Bromage H, Tempst P, Bird A & Reinberg D (1999) Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13:** 1924–1935

Zhu J, He F, Hu S & Yu J (2008) On the nature of human housekeeping genes. *Trends Genet.* **24:** 481–4

Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A & Meissner A (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500:** 477–81

6 Curriculum vitae

EDUCATION

International PhD Program Friedrich Miescher Institute, Basel, Switzerland

07/2013 – present

PhD in Genetics

Thesis title: “Design principles of promoter and enhancer activity in mammalian genomes”

University of Vienna, Vienna, Austria 10/2010 – 12/2012

MSc in Molecular Biology

Thesis title: “Histone H3 phosphorylation in metastasis”

Karl-Franzens-University, Graz, Austria 10/2007 – 05/2010

BSc in Molecular Biology

Europagymnasium Auhof, Linz, Austria 09/1998 – 06/2006

High School diploma

WORK EXPERIENCE

Friedrich Miescher Institute, Basel, Switzerland 07/2013 – present

PhD student

Laboratory of Prof. Dirk Schübeler

1. Project: Design principles of CpG Island promoter activity
2. Project: Cis-regulatory landscape of four cell types of the retina

Max F. Perutz Laboratories, Vienna, Austria 08/2011 – 10/2012

Master student

Laboratory of Prof. Christian Seiser

1. Project: Histone H3 phosphorylation in metastasis
2. Project: H3S28 phosphorylation in cellular stress

FELLOWSHIPS

-Boehringer Ingelheim Fonds Fellowship 08/2014 – 07/2016

SCIENTIFIC PRESENTATIONS

-Talk EPD 30th Anniversary Symposium

29th – 30th September 2016, Lausanne, Switzerland

-Talk BIF Summer Seminar

27th August – 2nd September 2016, Hirschegg, Austria

-Talk BIF Summer Seminar

8th – 14th August 2015, Hirschegg, Austria

COURSES

-Scientific Communication and Communication of Science to Lay Audiences

15th – 20th May 2015, BIF, Lautrach, Germany

-Statistical Techniques for Genomics/Life Sciences

9th – 12th February 2015, Swiss Institute of Bioinformatics, Basel, Switzerland

-Advanced RNA-Seq and ChIP-Seq Data Analysis Course

12th – 15th May 2014, EMBL-EBI, Hinxton, UK

PUBLICATIONS

1. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015; 528(7583):575–579. doi:10.1038/nature16462.
2. Sawicka A, Hartl D, Goiser M, et al. H3S28 phosphorylation is a hallmark of the transcriptional response to cellular stress. Genome Res. 2014; 24(11):1808–1820. doi:10.1101/gr.176255.114.