Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics of

the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M. Sc. IGNACIO IBARRA DEL RÍO

Born in: Santiago de Chile, 1989

Oral examination: 16 September 2019

Integrative modeling of transcription factor cooperativity and its effects on phenotypic variability

Referees:

Prof. Dr. Stefan Wiemann

Dr. Wolfgang Huber

Abstract

The regulation of biological processes relies on a complex nucleotide code embedded in our DNA. Its decoding and interpretation is the main task of Transcription Factors (TFs), which altogether enable the recognition and modulation of gene expression. Whenever factors bind to DNA, a set of additional criteria and conditions need to be satisfied, such as TF concentration, DNA openness, and cooperativity with other binding factors. Such combinations of DNA-bound TFs, as well as their structural and functional cooperativity, allow a fine-grained control of gene expression due to subtle changes in specificity in both DNA recognition and functional outcomes.

This thesis explores the prediction of structural TF cooperativity and its biological consequences. Through integration of publicly available TF binding data, we explore the prediction of determinants of TF-cooperativity across TF families, and validate our observations. By incorporating multi-omics data we set up a framework for annotation and scoring of ontologies associated to TF-TF binding, validating our findings using cancer expression data. Additionally, this thesis explores functional cooperativity between TFs in the context of neuronal activity, delineating rules that determine gene expression programs through up-regulation of specific TFs and their combinations. Finally, we investigate the TF-interactions in cell reprogramming, predicting and validating novel interactions between TF activators and repressors. Altogether, this dissertation provides an extensive set of insights to better understand the complex interplay between TFs cooperativity and phenotypes.

Zusammenfassung

Die Regulation biologischer Prozesse beruht auf einem komplexen Code aus Nukleotiden, der in unserer DNA eingebettet ist. Das Entschlüsseln sowie die Interpretation dieses Codes stellt die Hauptaufgabe der Transkriptionsfaktoren (Transcription Factors, TFs) dar, die insgesamt die Erkennung und Modulation der Genexpression ermöglichen. Wenn diese Faktoren an DNA binden, müssen allerdings eine Reihe zusätzlicher Bedingungen erfüllt sein, wie z.B. TF-Konzentration, DNA-Offenheit und Kooperativität mit anderen bindenden Faktoren. Solche Kombinationen von DNA-bindenden TFs, sowie deren strukturelle oder funktionelle Eigenschaften, ermöglichen eine genauere Kontrolle der Genexpression, aufgrund geringfügige Veränderungen in der Spezifität sowohl bei der DNA-Erkennung als auch in der funktionellen Auswirkung.

Die vorliegende Dissertation untersucht Ansätze zur Vorhersage von struktureller TF-Kooperatitvität und ihrer biologischen Folgen. Durch die Integration von öffentlich verfügbaren Daten, die TF-Interaktionen umfassen, untersuchen wir die Vorhersage von spezifischen Eigenschaften, die die Kooperativität zwischen TF-Familien begründen, und validieren unsere Beobachtungen in nachfolgenden Experimenten. Durch die Einbeziehung von Multi-Omics-Daten erstellen wir ein Framework für die Annotation und Bewertung von Ontologien für TF-TF Interaktionen und bestätigen unsere Beobachtung anhand von Expressionsdaten von Krebs-Patienten. Darüber hinaus untersucht die Dissertation die funktionelle Kooperativität zwischen Transkriptionsfaktoren im Kontext von neuronaler Aktivität und gibt Einblicke wie die

Regulation spezifischer Faktoren und deren Kombination die Genexpression beeinflussen und bestimmen. Die vorliegende Arbeit nimmt ebenso die Wechselwirkungen von Transkriptionsfaktoren im Zusammenhang mit der Zellprogrammierung in Augenschein, sowie auch die Vorhersage und Validierung von spezifischen TF-Aktivatoren und -Repressoren. Insgesamt stellt die Dissertation damit eine umfassende Studie dar, die mit neuen Einblicken und Ansätzen das komplexe Zusammenspiel von Transkriptionsfaktoren, DNA-Erkennung und Phänotypen beschreibt.

Acknowledgements

The collective work written in this dissertation would not have been possible without the constant support of multiple people during my PhD.

First, I would like to acknowledge my advisor Judith Zaugg for letting me work on these ideas and her constant guidance and support. I also thank my TAC committee members Christoph Müller, Stefan Wiemann and Wolfgang Huber for their feedback and their time.

I am grateful to all the people who participated in these projects and actively improved the overall quality of this work. For the second Chapter, I thank Britta Velten and Bernd Klaus for their statistical support and patience in explaining me how to improve my analyses. I thank Janosch Hennig, Nele M. Hollmann and Sandra Augsten for actively collaborating in the experimental validations. I also thank Constantin Ahlmann-Eltze, who worked with me in the exploratory phases of this work.

I thank Vikram Ratnu who acquired the experimental data for the third Chapter of this dissertation. I also thank Kyung-Min Noh for her advice and her extended availability to discuss and propose new directions. I thank Lucía Gordillo for carrying out final experiments.

I thank Moritz Mall who was guiding the project presented in the fourth Chapter of this dissertation. He was truly throughout the entire project, and supporting when results were not coming up in due time. Additionally, I thank Juan Segarra for experimental support and expertise.

I want to thank the mentor and friendship of my Bachelor and M.Sc. advisor Francisco Melo, who helped me prepare for the PhD program at EMBL. Additionally, thanks to

Damien Devos and M.S. Mashusudhan, who trained me during my pre-PhD phase. I also thank Matthew Weirauch, who helped develop my computational skills through an early collaborative project.

Friendships have been an important part of my time in Heidelberg. I am grateful to Paul Igor Costea for his friendship, support, and harsh criticism on virtually everything. I am also thankful to Deepikaa Menon, Sean Powell, and Louis Pedro Coehlo for critical conversations and their sharp advice. I thank all my predoc fellows who went through the same process along with me, especially to Renato Alves for his relentless coaching in Computer Science and his joyful personality. I am particularly grateful to Mariana Ruiz-Velasco for the running sessions, and the friendship we developed during our PhD years.

I thank my lifetime friends in Chile, for never stop being a beacon of light, even in times of not seeing each other for months or even years. Seeing you getting married, having children and moving on with your life has been a driving force for me while in Germany.

I am especially grateful to Natalie Romanov for being an important pillar in the development and writing of this dissertation.

I thank my Grandmother for her love and almost weekly contact during my PhD. Finally, I thank my parents for their love, and support.

Contents

List of Figures

Supplementary Material

List of Publications

Ignacio L. Ibarra*, Nele Merret Hollmann, Bernd Klaus, Sandra Augsten, Britta Velten, Janosch Hennig & Judith B. Zaugg (2019). Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Submitted.*

Ignacio L. Ibarra*, **Vikram S. Ratnu*.** Lucia Gordillo, Luca Mariani, Katy Weinand, Martha L. Bulyk, Judith B. Zaugg & Kyung-Min Noh (2019). BDNF promotes enhancer accessibility essential for gene activation and exon usage in neurons. *In preparation.*

Ignacio L. Ibarra*, **Segarra Juan***, Judith Zaugg & Moritz Mall. Systematic classification of terminal repressors in multiple cell lineages. *In preparation.*

Glossary

Arc	Activity-regulated cytoskeleton-associated protein	**KLF**	Kruppel-like factors
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing	**Myt1l**	myelin transcription factor 1 like
AUC	Area under the curve	**NMR**	Nuclear Magnetic Resonance
BDNF	Brain-derived neurotrophic factor	**PCA**	Principal Component Analysis
bZIP	Basic Leucine Zipper Domain	**PR**	Precision-Recall
CAP-SELEX	Consecutive Affinity Purification SELEX	**Prox1**	Prospero homeobox protein 1
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing	**RNA-seq**	RNA sequencing
CLL	Chronic lymphocytic leukemia	**ROC**	Receiver Operating Characteristic
CPE	Carboxypeptidase E precursor	**SELEX**	Systematic evolution of ligands by exponential enrichment
CTCF	CCCTC-binding factor	**T2D**	Type 2 Diabetes
DBD	DNA-binding domain	**T2D**	Type 2 Diabetes
EGR	Early growth response	**Tbx5**	T-box transcription factor
HIC1	Hypermethylated in cancer 1 protein	**TF**	Transcription Factor
HT-SELEX	High-Throughput SELEX	**TR**	Terminal Repressor
ITC	Isothermal Titration Calorimetry	**Trio**	Trio Rho Guanine Nucleotide Exchange Factor
KCl	Potassium Chloride	**TSS**	Transcription Start Site

The contents of this dissertation are the result of my own work. Results provided by experiments done in collaboration are adapted and acknowledged in the beginning of each Chapter

Chapter 1

Introduction

The development and maintenance of all biological events that constitute an organism requires a precise and robust control of all mechanistic steps involved to ensure the correct information flow. These control mechanisms are also pivotal to provide a dynamic yet precise response to environmental stimuli relevant for survival and reproduction. Evolution has allowed selection of millions of independent controlling strategies in different biological contexts, enabling the diversity in life we see in our world today. Concurrently, biological information flows and systems are also prone to defects that are associated with developmental failures and disease.

The functional layers involved in sensing and responding to stimuli are important for translation of environmental signals into specific molecular mechanisms. Every complex and dynamic behavior, such as the fight-or-flight stress response or social behavior in bees [**Bloch et al 2011**], has an underlying molecular basis. At the same time, thousands of biochemical reactions take place in cells, ensuring homeostasis, survival and reproduction. Due to the importance of these reactions, a proper coordination is pivotal for ensuring specificity of the molecular control.

The landscape of all possible reactions in an organism has been described over decades in Biochemistry research. Millions of these chemical steps dynamically occur at the same time. At the most rudimentary level these reactions rely on binary interactions between

metabolites and bio-molecules such as proteins, nucleic acids and lipids. The sum of all these interactions and their combinations define the molecular switches that allow signal transmission in the cell.

1.1. Protein-DNA interactions and Transcription Factors

Interactions between proteins and nucleic acids are important in a plethora of major biochemical processes, and arguably represent the most sensible regulatory control mechanism for genome readout. Such interactions maintain the tightness in genome architecture, DNA replication in cell division and control of gene expression. It is therefore pivotal that these biochemical reactions to remain efficient, precise and robust across generations. To date, different types of proteins DNA-interacting proteins are known [**Rohs et al 2010**] and their crucial functional roles manifest in their strong conservation levels in Prokaryotes [**van Hijum et al 2009**; **Santos-Zabaleta et al 2018**] and Eukaryotes [**Nitta et al 2015; de Mendoza et al 2013**].

Transcription factors (TFs) represent one of the most studied types of DNA-interacting proteins. These sequence-specific regulatory proteins exert control on gene expression by recruiting transcriptional complexes and chromatin modifiers, which effectively enhance or suppress transcription of nearby genes. Globally, TFs are classified based on the structural conservation of their DNA-binding domains (DBDs), which represent early determinants of gene expression control, and simultaneously evolved with specific roles in the three main kingdoms of life. Several of these DBDs have specialized through gene duplication and divergence in Eukarya, allowing functional diversification. For example, one family of TFs, the Zinc-Fingers (C2H2 ZF), contains approximately 800 members in humans. Current surveys estimate the number of TFs in humans to be around 2000 in total

[**Lambert et al 2017**]. Molecular processes such as splicing would mean that these numbers might be much higher with up to 8000 according to current estimates [**Inukai et al 2017**].

To understand how these regulatory factors recognize DNA, the protein-DNA interactions themselves need to be thoroughly investigated. Structural determination of TF-DNA complexes have revealed over-represented chemical interactions between TFs and DNA [**Angarica et al 2008**], restricting the number of biological features defining the protein-DNA recognition code. These interactions have been classified as 'Base-readout' or 'Shape-readout' according to how they bind the DNA molecule [**Rohs et al 2010**] (**Fig 1.1**). The first readout-mode is described by base interactions mediated by hydrogen bonds or Van der Waals contacts, while the second one is related to proteins 'reading' the overall structure of the DNA, including its electron density, electrostatic potential e.g. in the minor groove, and DNA-backbone.

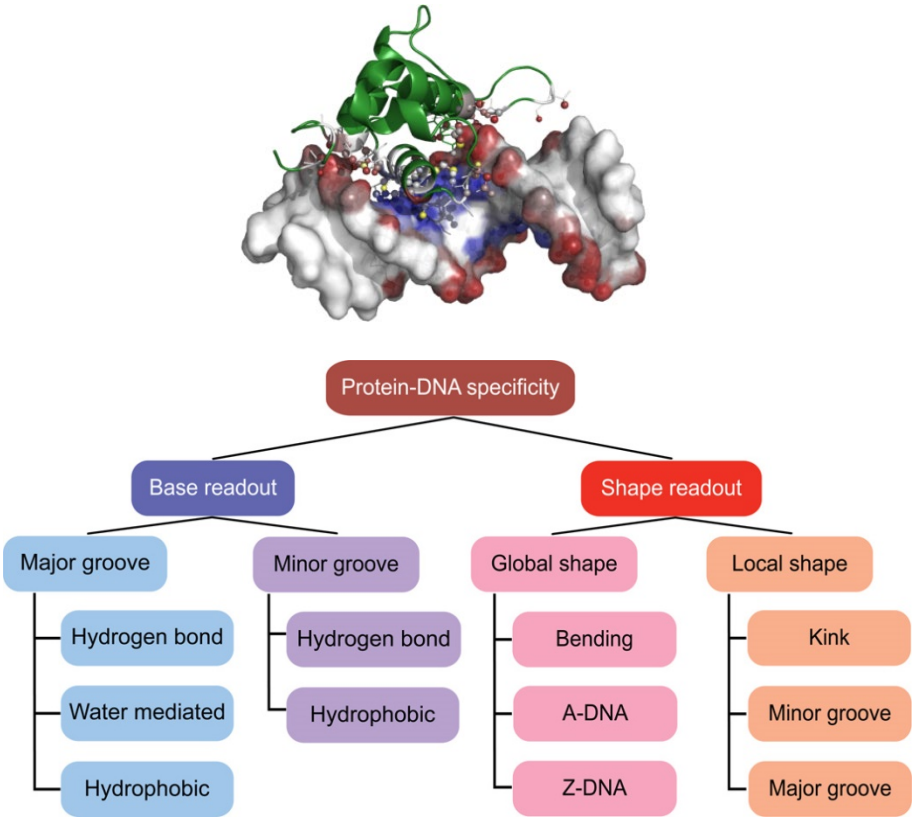

Figure 1.1. Classification of protein-DNA interactions that confer specificity

(top) Protein-DNA complex between Transcription Factor FOXO1 (green) and its DNA binding site (gray). Interactions with the major groove bases (blue) indicate Base readout. The interactions between Forkhead wings and the neighboring DNA minor grooves are indicative of Shape readout. Visualization generated with PDIviz [**Ribeiro et al 2015**] *(bottom)* Classification of interactions types that determine recognition of DNA by proteins. Base readout is based on contacts residues and bases through hydrogen bonds, water mediated or hydrophobic contacts. Shape readout is based on the local and global structure of DNA, and it determines additional specificity. *Adapted from* [**Rohs et al 2010**].

1.2. Methods to explore TF binding specificities

Multiple experimental methods have been generated to study TF binding specificity. Techniques are primarily classified based on their readouts by microarray or deep-sequencing methods. Microarray-based methods are prominently represented in the Protein Binding Microarray (PBM) technique [**Berger et al 2006**], in which a TF of interest is interrogated against a microarray containing DNA-oligos of known sequences. The TF is coupled to a Glutathione S-transferase (GST)-tagged domain, which is ultimately detectable with a fluorophore-tagged antibody (**Fig 1.2**). Binding efficiency can thereby be assessed by relative fluorescence across the array. This approach has been pioneered by the group of Martha Bulyk and has been used to systematically generate data for hundreds of Metazoan TFs.

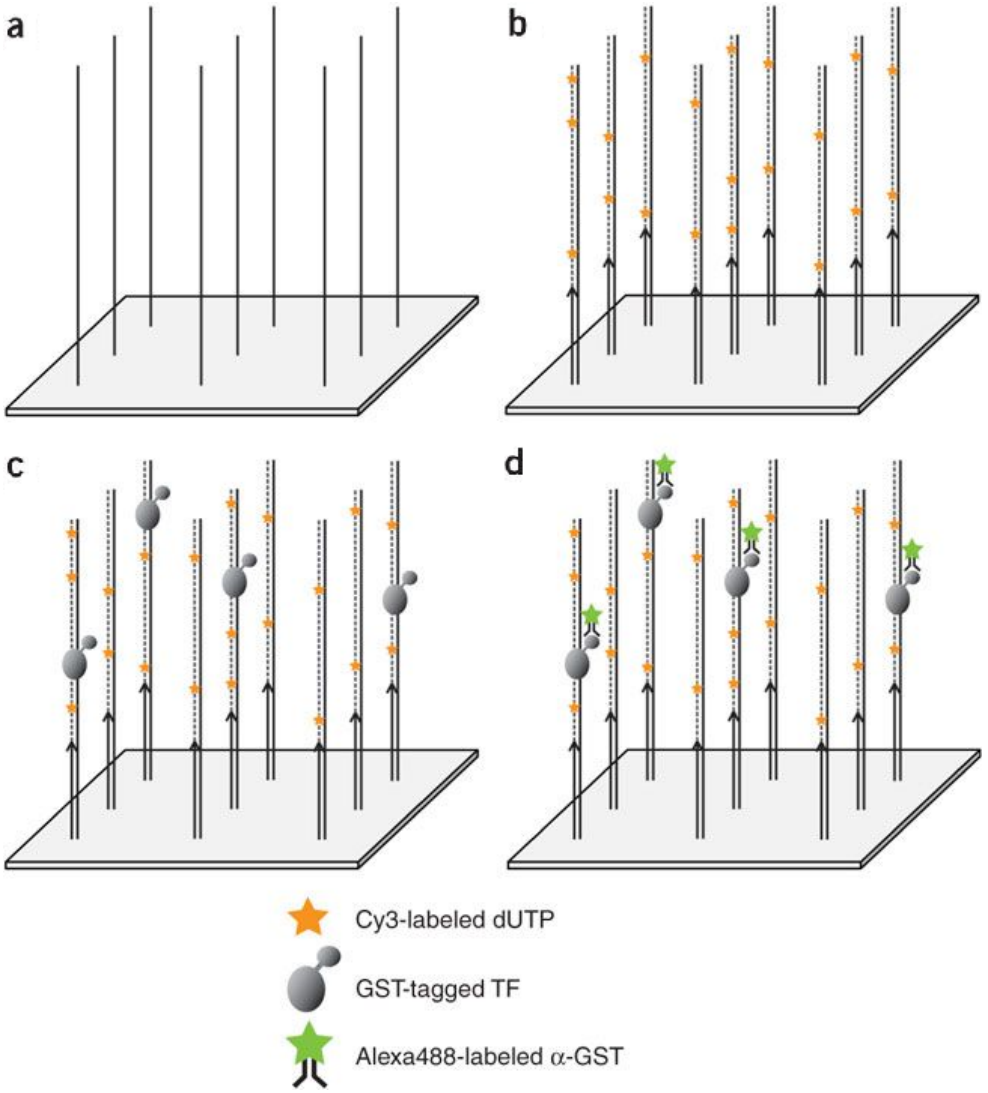

Figure 1.2. Description of Universal Protein Binding Microarray experiment

(a) A single-stranded DNA microarray purchased from a provider is double-stranded using a low concentration of fluorescently labeled dUTP. **(c)** A GST-tagged TF is tested for binding to the DNA sequences in the microarray, using as a detection system a **(d)** fluorophore-conjugated antibody against GST. *Adapted from* [**Berger and Bulyk 2009**].

Among deep-sequencing based methods, the systematic evolution of ligands by exponential enrichment (SELEX) is a practical way to generate TF-binding specificity readouts for TFs based on selection and sequencing of captured TF-specific DNA reads [**Jolma et al 2010**] **(Fig 1.3)**. DNA reads are separated on a DNA gel (SELEX-seq; Spec-seq) [**Riley et al 2014; Stormo et al 2015**], or capture by column purification using the tagged TF of interest. Once reads are obtained, the experiment cycle can be repeated several times by re-amplifying the DNA-oligos and re-capturing with the TF, increasing the fraction of DNA sequences with high-affinity TF binding sites. Approaches based on SELEX for studying TF binding have been improved in recent years, allowing the high-throughput interrogation of hundreds of TFs in the same experiment (HT-SELEX) [**Jolma et al 2013**]. Additional adaptations of this experiment have allowed the interrogation of TF combinations (CAP-SELEX) [**Jolma et al 2015**], effects of methylation (methyl-SELEX) [**Yin et al 2013**], and addition of nucleosomes (NCAP-SELEX) [**Zhu et al 2018**]. One limitation of these studies is that the overall coverage per DNA-sequence is lower when multiple TFs are multiplexed in the same deep-sequencing run. To overcome this, conventional SELEX-seq with a higher sequencing deep has been adopted to explore the DNA-binding for longer DNA footprints [**Zhang et al 2018**]. Adaptations of SELEX-seq techniques that include DNA-modifications can be used to explore TF sensitivity to epigenetic variation (EpiSELEX-seq) [**Kribelbauer et al 2017**].

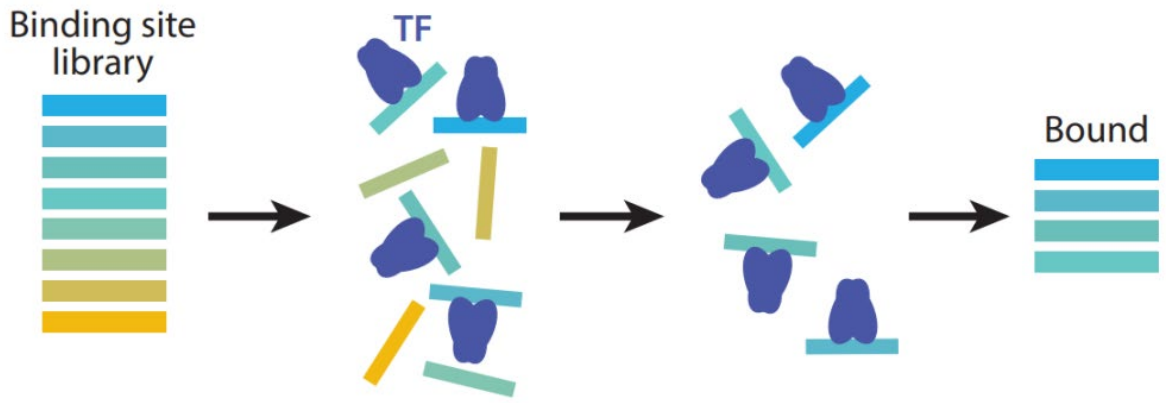

Figure 1.3. Description of HT-SELEX experiment

From a prepared DNA library with randomized DNA sequences of theoretically equimolar concentrations (Binding site library), a purified DNA-binding domain for a TF of interest (TF) is added and used for DNA selection based on protein-DNA interactions (Bound). Comparisons between Bound and the initial Binding Sites Library allow the assessment of features that confer protein-DNA specificity for the studied TF. *Adapted from* [**Kinney et al 2019**].

1.3. Modeling and prediction of TF-DNA binding interactions

Research efforts in the field have tried to conceptualize interactions between TFs and DNA using a set of computational approaches that summarize TF-DNA interfaces and electron density at such interfaces into simplified and interpretable models. One of the earliest models proposed are the Position Weight Matrices (PWMs). These multinomial models were early described by Gary Stormo [**Stormo et al 1982]**, and adapted by Thomas Schneider into a simplified visualization posited as Sequence Logos [**Schneider et al 1990**] (**Fig 1.4)**. Due to its simplicity and fast interpretability such representations are commonplace in the field. However, they have been shown to not fully capture the complexity of protein-DNA experimental data [**Weirauch et al 2011**], highlighting the need of identifying the rules guiding these interactions, and ultimately introducing better models for community interpretation.

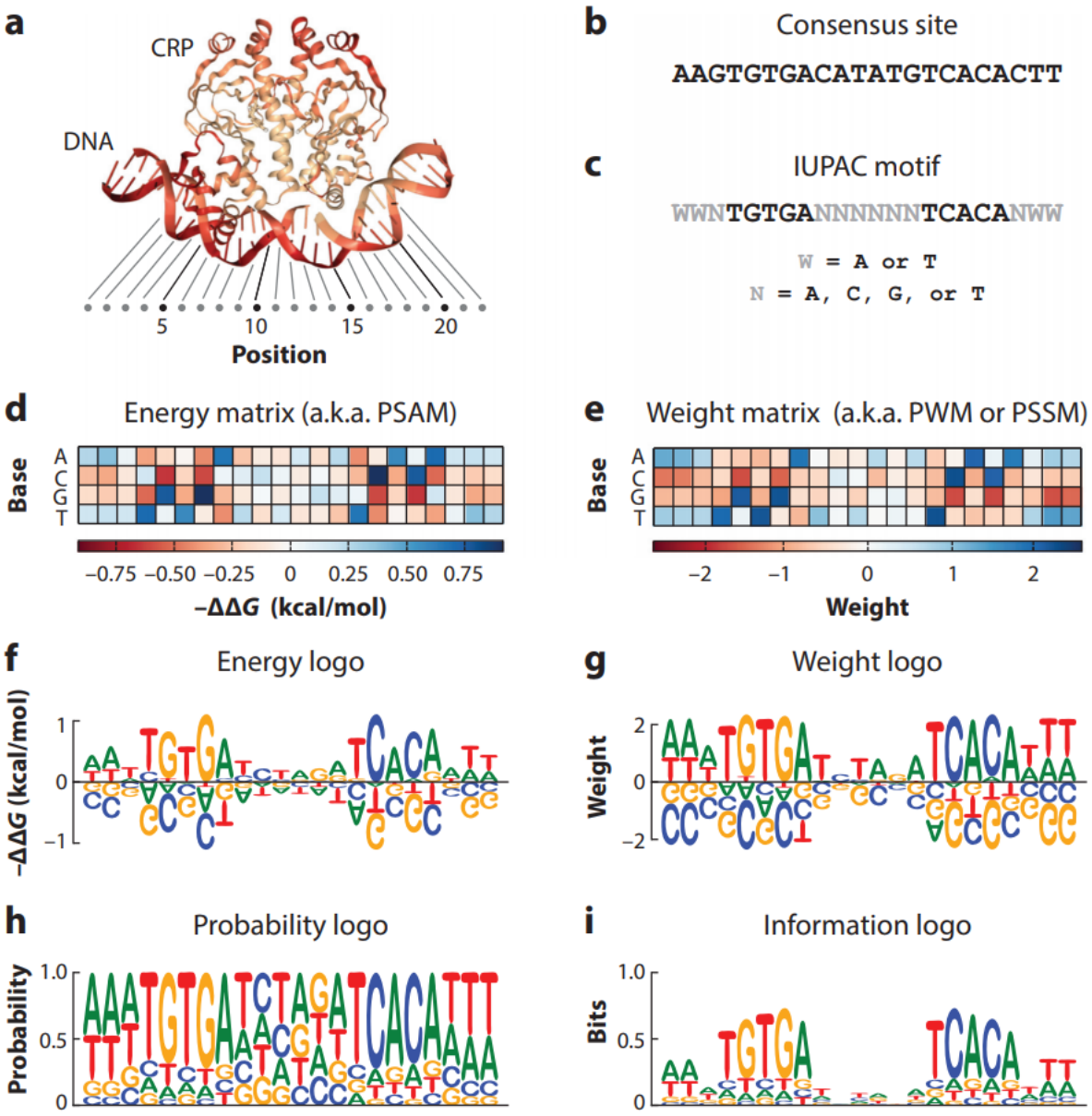

Figure 1.4. Additive models to display protein-DNA recognition.

(a) Protein-DNA complex formed by the cAMP receptor protein (CRP) binding to its cognate binding site (PDB ID: 1CPG). Relative positions indicate DNA base pairs used for following visualizations. **(b)** Consensus binding site of CRP based on crystal structure **(c)** IUPAC motif based on DNA variants that are bound by CRP. **(d)** Energy matrix. Each weight represents the change in ΔG (-ΔΔG) expected when mutating the highest affinity DNA sequence in one position. **(d)** Weight matrix summarizing the relative increase or decrease in the expected probabilities for a selected set of binding sites with respect to a background distribution (e.g. genome-wide GC content). **(f)** Logo visualization for *(d)*. Letter heights indicate -ΔΔG values. **(g)** Logo visualization for *(e)*. Letter heights indicate weights increase or decrease. **(h)** Probability logo indicates the nucleotide probabilities for each position in the set of positive sequences. **(i)** Information logo summarizing the reduction of entropy in each position in *(h)* and highlighting the positions with the largest information values. *Adapted from* [**Kinney et al 2019**].

Historically, several such models for protein-DNA specificities *in vitro* have been proposed and made available [**Weirauch et al 2011**]. These models allow the quantification of DNA enrichment upon addition of a protein in sequencing data ("relative affinities" in SELEX data), or fluorescent signal for a DNA-oligo containing potential binding sites of interest in Protein Binding Microarray data. Proposed models are the ones with the highest held-out performances, and summarize the most important DNA-recognition features as a set of mononucleotide and dinucleotide contributions (**Fig 1.5)**. Recently, it has been proposed that models encompassing a full biophysical description of protein-DNA interactions are an adequate alternative to more complex Deep Learning based classifiers [**Rastogi et al 2018**]. One of the main arguments for this is that Deep Learning models fail to consider the full spectrum of low affinity binding and ultimately fail in the prediction of those. To date, extended comparisons between traditional, biophysical and Deep Learning models for the prediction of DNA-binding sites *in vivo* would be required. However, the adoption and development of biophysical methods to score TF binding sites are delayed with respect to the vast amount of available tools and web services to obtain TF motifs from sequence data with PWMs [Bailey et al 2009].

Simpler models describing protein-DNA interactions, on the other hand, are based on the collective grouping of sequence patterns enriched in *in vitro* experiments and their experimental readout as groups of *k*-mers [**Mariani et al 2017**]. These *k*-mers can be directly used to directly interrogate biological sequences with a higher signal-to-noise ratio due to the direct inclusion of patterns that are not necessarily robustly represented in Position Weight Matrices.

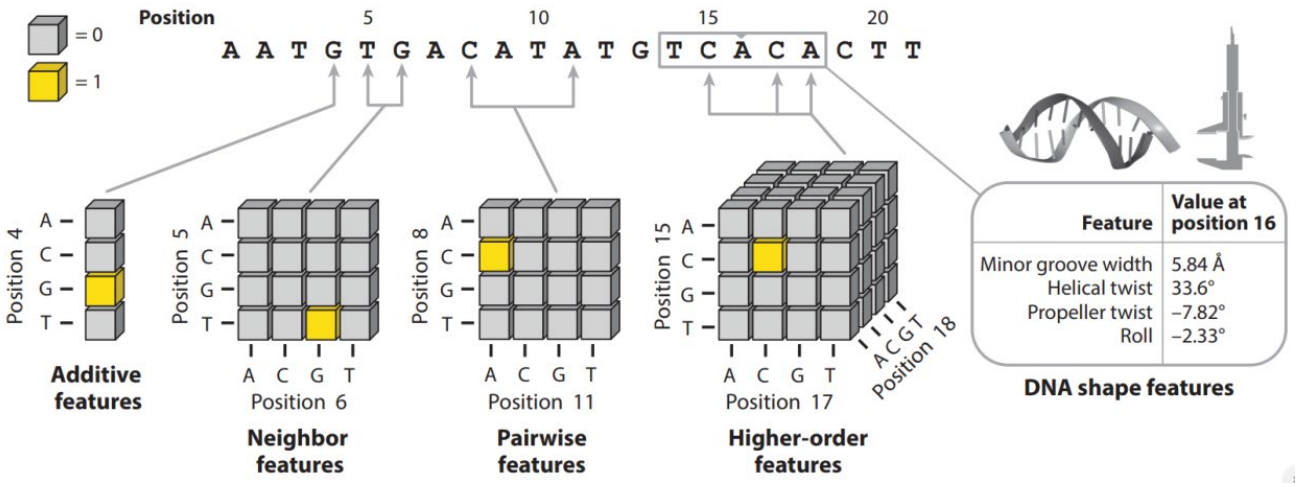

Figure 1.5. Commonly used features to describe TF-DNA interactions

From a target sequence considered to be a putative binding site for a TF, the weights for features that are additively used to score the binding site relevance are scored using mononucleotides (additive features), proximal dinucleotides (neighbor features), non-local dinucleotides (pairwise features), and trinucleotide combinations (higher-order features). Additionally, DNA pentamers can be assessed by their underlying DNA geometry (DNA shape features). *Adapted from* [**Kinney et al 2019**].

Altogether, the modeling of TF-DNA features allows the understanding of readout mechanisms contributing to TF binding. However, TFs do not act alone *in vivo*, and multiple biological features can affect their binding to DNA. The chromatin environment, methylation, coding and non-coding variation and interactions with other TFs are examples of such confounders (**Fig 1.6**) [**Inukai et al 2017**].

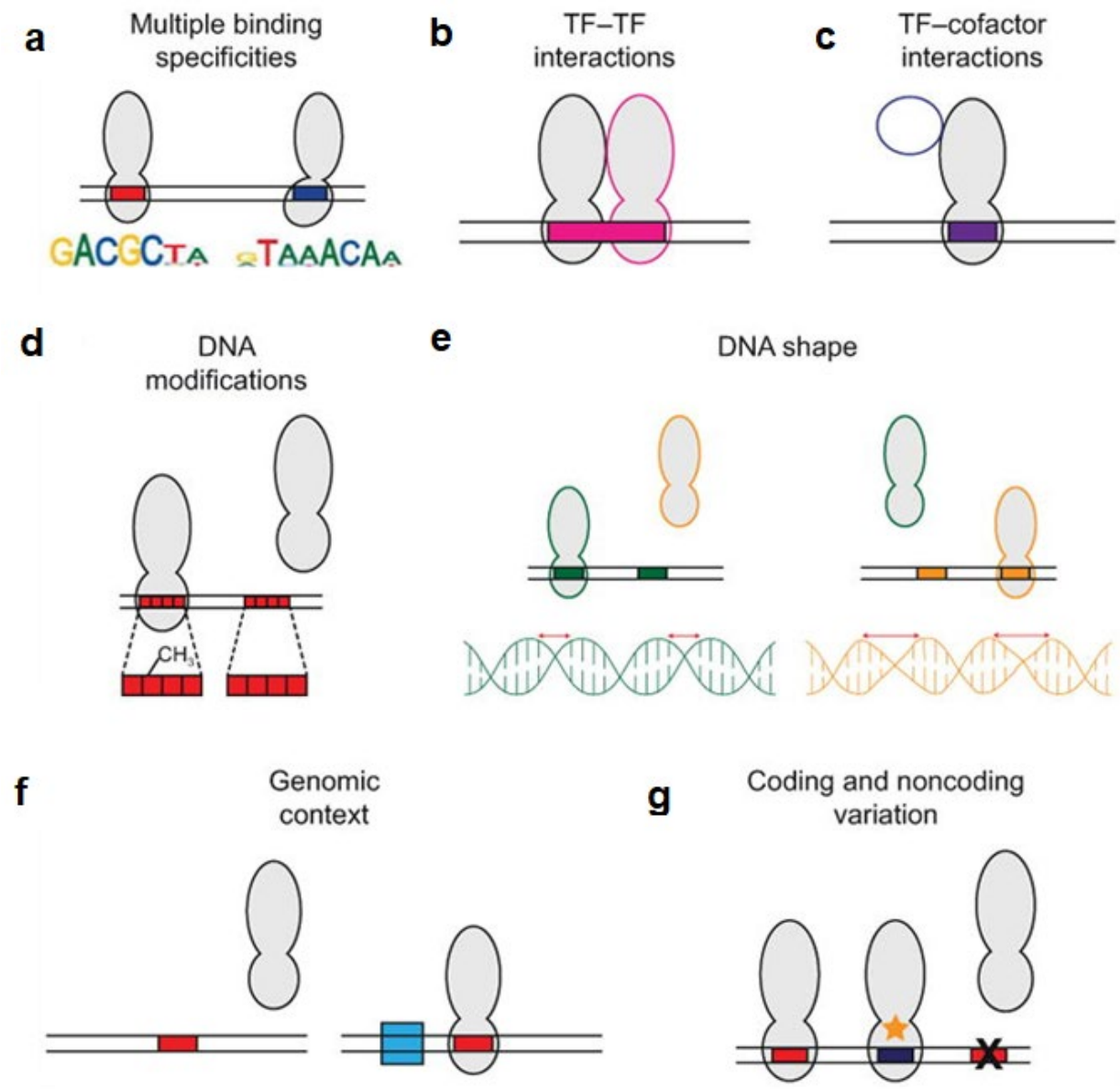

Figure 1.6. Features beyond binding primary TF motifs can modulate TF binding recognition.

(a) Various TFs can be in different binding modes, requiring additional models to describe all possible configurations. **(b)** Interactions with other TFs can confer additional cooperativity through protein-protein or DNA-mediated allostery. **(c)** Interactions with non-DNA interacting cofactors can modulate latent specificity TF-binding properties not activated in their presence. **(d)** Some TFs are specifically sensitive to methylation, and can increase or decrease their binding potential upon DNA-methylation. **(e)** The local DNA shape can determine TF binding specificity. (**f**) The genomic context can determine that some TFs will be preferably recruited to certain sites according to the chromatin state or nucleosome placement. **(g)** Coding (star) and non-coding (shown as X on binding site) variation adds additional complexity in TF binding. Adapted *from* [**Inukai et al 2017**].

1.4. TF interactions and cooperativity as an additional regulatory layer of biological function

As TFs can also form complexes with other TFs the spectrum of possible interactions and regulatory switches increases dramatically [**Morgunova et al 2017]** (**Fig 1.7)**. Biophysically, most of these TF-TF complexes with DNA require few or no protein-protein interactions at all [**Jolma et al 2015**].

Several efforts have been made to formulate the systematic prediction of composite TF-TF binding sites *in vivo* [**Guturu et al 2013; Jankowski et al 2014**]. Recent experimental surveys have described these interactions to be promiscuous and widespread across TF families, and estimate the amount of interactions to be around 25000. In fact, 1 out of 100 possible TF-TF pairs in the human genome are expected to form a cooperative pair. The implications of this kind of TF-TF interactions remain elusive but undoubtedly define yet another regulatory layer with functional consequences, as there are multiple cases in which TF-TF binding has been shown to be associated to downstream biological function. Examples are the homeodomain dimers in development [**Slattery et al 2011**], or the olfactory receptor regulation through Lhx2-Ebf binding [**Monahan et al 2017**]. Genetically, it has also been shown that Genome-Wide Association Studies (GWAS) variants affect immune function through disruption of IRF4-BATF complex binding [**Iwata et al 2017**].

As these interactions seem to be prevalent in particular configurations [**Jolma et al 2015**], they are limited to specific TF-family cooperative binding events. Interestingly, these complexes have been shown to bind similarly across members of the same TF-families, and to recognize low-affinity binding sites collectively across the whole family, such as in the SOX-PAX family pair [**Narasimhan et al 2015**]. This highlights the need of models that

take those families into account as well to provide better cooperative binding predictions *in vivo*.

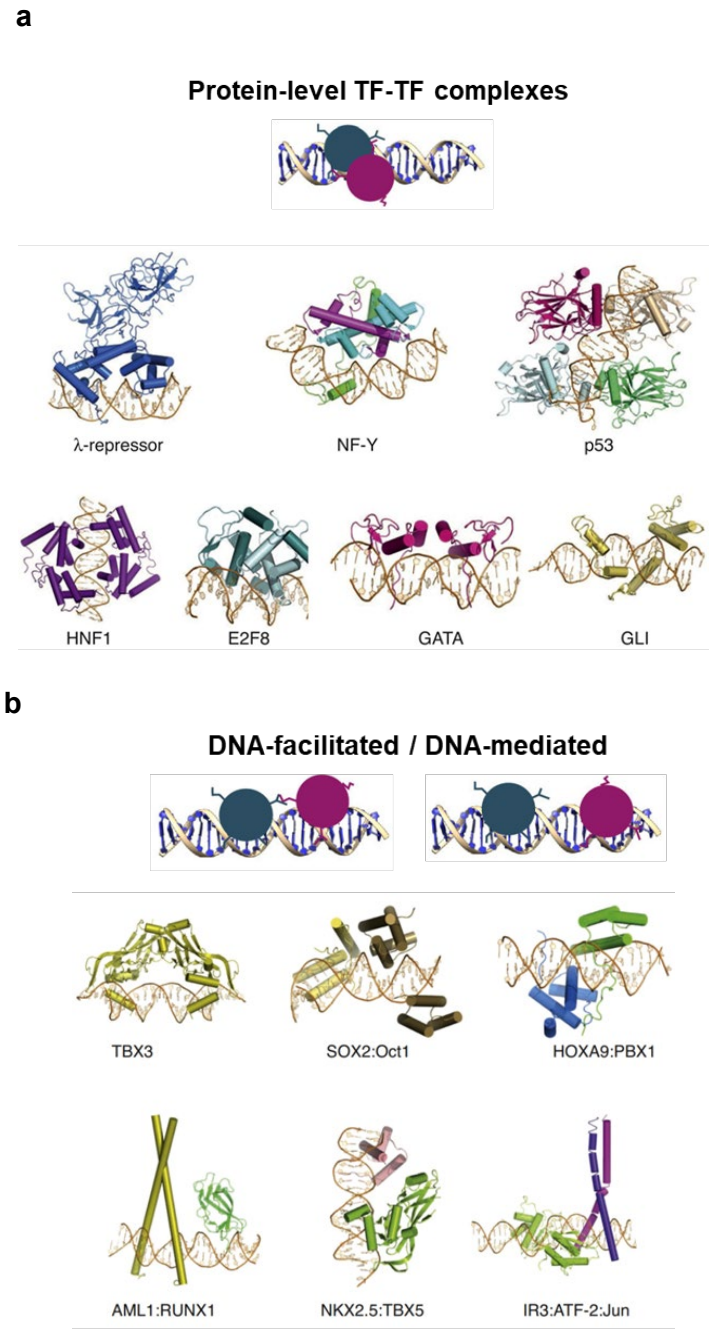

Figure 1.7. Types of interactions between TFs when bound to DNA.

(**a**) (*top*) Depiction of protein-level TF-TF complex with DNA. (*bottom*) From left to right in both rows: lambda-repressor (PDB: 3BDN); NF-Y trimer (alpha/beta/gamma subunits in green, cyan and magenta, respectively) (PDB: 1CF7); p53 tetramer (PDB: 2AC0); HNF1 homodimer (PDB: 1IC8); E2F8 domains (PDB: 4YO2); two GATA zinc-finger domains (PDB: 3DFV); three Zinc-finger domains of GLI (PDB: 2GLI). (**b**) (*top*) Depiction of DNA-facilitated or DNA-mediated TF-TF complex. (*bottom*) left to right in both rows: TBX3 bound to palindromic site (PDB: 1H6F); SOX2:Oct1 complex (light yellow and dark yellow, respectively) (PDB: 1O4X); HOXA9:PBX1 (green and dark blue, respectively); AML1:RUNX1 (green and yellow) (PDB: 1HJB); NKX2.5:TBX5 (pink and green) (PDB: 5FLV); IRF3:ATF-2:c-JUN (green, violet and magenta) (PDB: 1T2K). Adapted *from* [**Morgunova et al 2017**].

1.5. Profiling the accessible genome

To study such TF-DNA interactions it is necessary to profile genome-wide chromatin accessibility. Recently, a new methodology to quantify genome-wide chromatin accessibility called Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) has been introduced as a powerful and lower cost alternative to previous approaches [**Buenrostro et al 2013**] (**Fig 1.8**). The technique relies on the usage of a mutant Tn5 transposase that is hyperactive and inserts adapters in the open regions of the genome, by tagmentation. These fragments are prepared for sequencing by DNA purification and PCR amplification. Processing of reads via mapping and comparison between treatments allow the overall annotation of regions with gained, closed or unchanged chromatin accessibility. Specific analyses of the DNA sequences that are cover regions with changed accessibility allows mapping TF motifs and footprints that describe overall physical properties of potential TF binding in such loci [Schep et al 2017].

1.6. TF combinations orchestrating the chromatin environment and gene expression

The combinatorial role of TFs when binding to the accessible genome is understood as a collective recruitment of several factors in order to activate or repress a signal [**Spitz et al 2012**]. As such, the very minimum amount of TFs is required to be recruited in order to modify the chromatin environment, increase or decrease gene expression of local genes and ultimately initiate a regulatory response.

However, not all TFs can bind to nucleosome-occluded DNA without the prior binding of other factors that would open the chromatin structure. TFs with this capacity are known as "pioneers", based on their ability to displace nucleosome and open chromatin regions where they become bound [**Mayran et al 2018**]. As the binding of TFs without this property

at nucleosome-occluded DNA relies on such TFs, the binding of pioneer TFs is usually followed by co-regulators that bind motifs at those sites but do not actively open the surrounding chromatin. Given that not all TFs have an associated pioneering activity, the interactions between pioneer and non-pioneer factors is pivotal for our understanding of recruitment events in the first place [**Mayran et al 2019**].

TF binding is highly dependent on these chromatin accessibility changes, with accessible regions being enriched for pioneer and co-regulator TF motifs in a positional way [**Su et al 2017**] **(Fig 1.9)**. Relating chromatin accessibility changes to regulatory mechanisms through the recruitment and interactions of those factors is therefore undoubtedly a contemporary challenge. These relationships are the main topic of **Chapter 3**, where connections between neuronal activity and chromatin accessibility are studied.

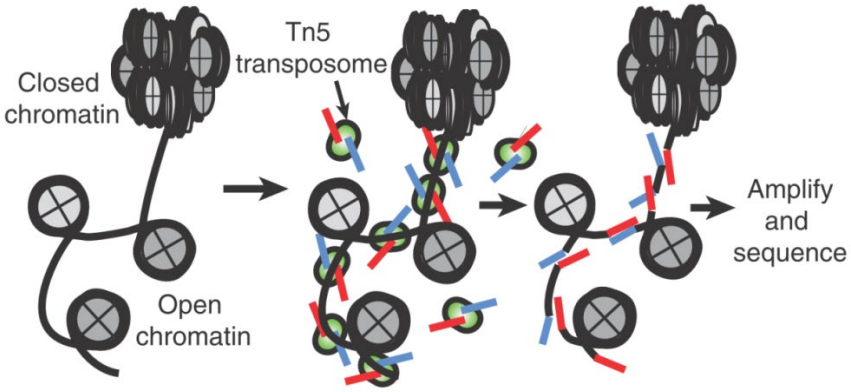

Figure 1.8. Schematic illustration of the ATAC-seq protocol.

(left) Visualization of chromatin regions with closed or open conformations (*middle*) addition of hyperactive Tn5 transposase tags open regions (*right*) purification, PCR amplification and deep-sequencing of tagged fragments and mapping indicates regions of high accessibility. Adapted from [**Buenrostro et al 2017**].

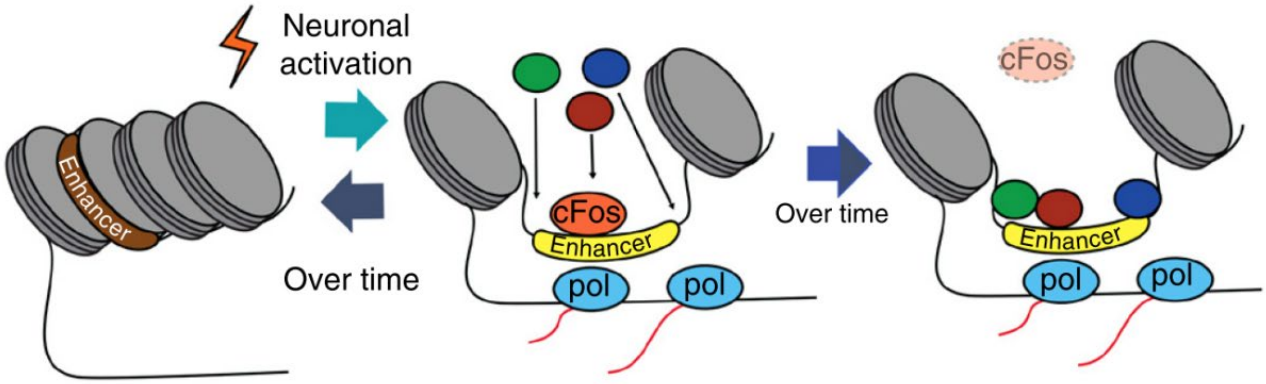

Figure 1.9. Working model of neuronal activity based on pioneering and co-regulators recruitment to regulatory DNA sequences.

Upon neuronal stimulation (e.g. with KCl), specific TFs with pioneering activity such as the ones with bZIP domains (cFos), bind to nucleosome-occluded regions and displace nucleosomes, increasing the overall DNA accessibility in those regions. Next, co-regulator TFs without pioneering activity (green, blue, brown) can bind to DNA and mediate gene response by enhancer remodeling and enhancer-promoter interactions. Finally, the pioneer factor is decreased in expression and its binding is not required anymore, but co-regulators are maintained in activated regions. Adapted from [**Su et al 2017**].

1.7. Cooperative binding of TFs involved in differentiation and reprogramming

It has generally been understood that TF combinations play a relevant role in determining cell differentiation, and conversion of cell types. One of the most important discoveries in the field of regenerative medicine has undoubtedly been the reprogramming of fibroblasts into induced pluripotent stem cells (iPSCs), using a specific combination of TFs known to be over-expressed in this type of cells (Sox2/Klf4/Oct4/Myc, or the Yamanaka Factors) [**Takahashi et al 2006**].

Most reprogramming experiments have been done in fibroblasts, and it is still unclear whether it is possible to trans-differentiate every single cell type into any other target cell type [**Fu et al 2017**]. Apart from identifying the activating TFs, other mechanistic rules need to be dissected. The epigenetic memory involves tagging specific chromatin regions, which would need to be chemically removed to allow robust differentiation [**Ng et al 2008**]. Dissecting such interactions between TFs and epigenetic modifications has proven

useful to further understand how a cellular fate is maintained after reprogramming [**Holmberg et al 2012**; **Hörmanseder et al 2017**].

At the same time TFs with roles in repression of non-cell fate genes have also been suggested to be important. The first of its kind, described in neurons, is the Myelin transcription factor 1 like (Myt1l) [**Vierbuchen et al 2010**]. This factor is able to increase the reprogramming efficiency of Mouse Embryonic Fibroblast into induced neurons [**Vierbuchen et al 2010**], and has been linked to the active repression of genes related to non-neuronal pathways [**Mall et al 2017**]. Given that Myt1l is overexpressed in most neuronal subtypes but it is absent in almost all other tissues, this factor is deemed a "terminal repressor" of non-neuronal cell fates relevant for neuronal fate maintenance. This concept allows speculation about the possible existence of such terminal repressors in other cell types opening an exciting avenue to be explored further. Indeed, such unknown factors with terminal repressor potential in other cell types can prove valuable for redesigning current reprogramming protocols for the purpose of increasing specificity and efficiency. This topic is certainly of relevance given that most tools used for predicting reprogramming do not include this feature to date [Rackman et al 2016]. Ultimately, annotating such terminal repressors would give better insight into how cells require them for robust differentiation. Moreover, many neuronal diseases are related to mutations in Myt1l [**Blanchet et al 2017**], and it can thus be expected that potential terminal repressors are also related to disease in many other cell types.

1.8. Aims of the Thesis.

Given the current state and recent efforts in the TF recognition field, several questions remain unanswered. In the following Chapters, we seek to precisely answer the following questions (**Fig 1.10)**.

Question 1: Can we identify TF-cooperativity and features allowing its prediction in published TF-binding data?

Based on the idea that TF **cooperativity** is able to leverage low-affinity binding sites and confer additional specificity to certain TF-motifs, this dissertation explores the inference of TF cooperativity from *in vitro* data. Specifically, the first section of **Chapter 2** describes the integration of publicly available CAP-SELEX, as well as biophysical and biochemical experiments that are used to explore this question. The data is effectively summarized into a framework for the unbiased assessment of DNA sites that are preferably bound by TFs in a cooperative manner.

Question 2: What are the consequences of cooperative TF-binding in function and disease?

Given that multiple TF-TF complexes are expected to bind DNA cooperatively, there is the possibility of numerous novel associations between TF interactions and downstream function to be identified and investigated. Approaches to address this question are introduced and applied in the second part of **Chapter 2.** Our framework associates specific TF-cooperative binding *k*-mers present in ChIP-seq data with downstream functional consequences. Our argument on deriving ontology connections leveraging cooperative TF-binding was substantiated by recovery of functional ontologies linked to the individual TFs, and the validation of interesting cases in development and disease.

Question 3: What is the interplay between TFs and chromatin accessibility and how does it confer specific neuronal activity?

This question is addressed in **Chapter 3**, which presents an integrative study of multi-omics data in mouse cortical neurons. Chromatin accessibility profiled by ATAC-seq is used to assess immediate response by TFs and their associations to gene expression programs. The data demonstrates that particular TFs and their interactions drive specific responses in each condition, precisely modulating neuron function.

Question 4: Can we systematically predict terminal repressors in different cell types?

Based on the idea that the TF Myt1l improves reprogramming efficiency in the conversion from mouse embryonic fibroblasts to neurons through repression of non-neuronal genes, we addressed the question of classifying and predicting TFs with a similar role through integration of available gene expression and TF-binding data. In Chapter 4, we present a framework to score TFs as activators of cell fate genes or repressors of non-fate genes that can be used in the prioritization of TFs for specific reprogramming protocols. Two cases were predicted as terminal repressors in this approach and were validated in reprogramming experiments.

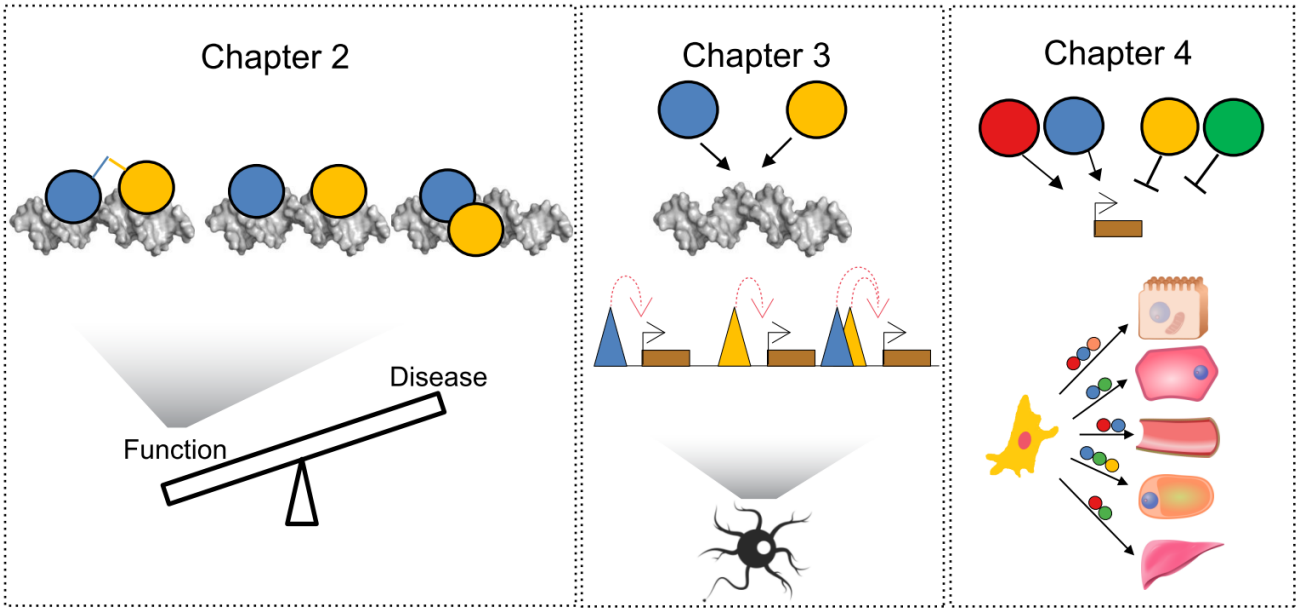

Figure 1.10. Visual overview of the main chapters of the dissertation.

Relationships between TF-TF cooperativity and function are studied in Chapter 2. Functional coordination between TFs binding and gene expression in the context of neuronal activity is presented in Chapter 3. Finally, in Chapter 4 the relationship between combinations of activator TFs (blue, red) with Terminal repressor TFs (yellow, green) is explored and discussed is the context of cell reprogramming

Chapter 2

Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions

In this chapter, I describe the analyses and results of a project exploring prediction of TF cooperative binding and their consequences. The methodology behind this work has been conceived by me, and I carried out all computational analyses, with support from other authors. The data underlying this analysis was obtained from published articles, as specified throughout this chapter. Additionally, biophysical and biochemical validations were generated by Nele M. Hollmann, Sandra A. Augsten, and Janosch Hennig. The work has been described in the following manuscript:

Ignacio L. Ibarra, Nele Merret Hollmann, Bernd Klaus, Sandra Augsten, Britta Velten, Janosch Hennig & Judith B. Zaugg (2019). Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Submitted.*

2.1. Introduction

Transcription factors (TFs) are essential for regulating cellular functions. This regulation is based on very specific protein-DNA interactions. For comprehending the regulation of biological processes it is therefore crucial to understand how TFs recognize their specific DNA binding sites [**Spitz *et al* 2012; Stormo *et al* 2013**].

The major determinants conferring TF binding specificity are the DNA sequence and DNA shape readouts [**Rohs *et al.* 2010**]. The former is guided by interactions between amino acids and DNA bases, whereas the latter is driven by the DNA structure preference of proteins mediated through DNA-backbone and DNA minor groove contacts. While sequence readout is a major driver of specific TF-DNA interactions, DNA-shape features improve the binding predictions for certain TF families both *in vitro* [**Zhou *et al* 2015; Yang *et al* 2017**] and *in vivo* [**Mathelier *et al* 2016**].

To date, over 1600 human TFs are annotated in the human genome [**Lambert *et al.* 2017**]. For many of them their DNA-binding preferences, summarized as TF motifs, have been determined either through *in vitro* assays [**Badis *et al.* 2009; Jolma *et al.* 2013; Jolma *et al.* 2015; Weirauch *et al.* 2014; Mariani *et al.* 2017**] or *in vivo* through chromatin immunoprecipitation followed by sequencing (ChIP-seq). However, despite the wealth of data and a good agreement between *in vivo* and *in vitro* derived TF motifs [**Orenstein et al 2014**] one of the long-standing challenges in the field is the high number of TF binding events that cannot be explained by the primary motif of the assayed TF. One of the proposed explanations for this phenomenon is that TFs can bind cooperatively and thereby strengthen their DNA binding affinity [**Morgunova *et al.* 2017**].

Recent studies leveraged high-throughput Systematic Evolution of Ligands by Exponential Enrichment coupled to Consecutive Affinity Purification (CAP-SELEX) [**Jolma *et al.* 2015**] to identify composite sites where cooperative TF-binding may occur. However, despite

these experimental advances, for most TFs the molecular mechanisms of cooperative binding and its distinction from co-binding remain elusive. Furthermore, even though it has been demonstrated that specific TF-TF interactions can alter sequence recognition through the formation of homo- or heterodimers, and are important for driving specific biological processes [**Slattery et al 2011; Monahan *et al* 2017; Huang et al 2015**], we lack a global understanding of the consequences of cooperative TF binding. This is mainly because we are missing the appropriate computational tools to systematically interrogate their functional associations.

Here, we implemented a framework to determine cooperative TF-binding preferences from *in vitro* SELEX data. We identified DNA shape as an important feature to predict cooperative TF binding, in particular for pairs between Forkhead and E26 transformation specific (Ets) members. This particular prediction was validated using nuclear magnetic resonance (NMR) spectroscopy and isothermal titration calorimetry (ITC). Through site-specific amino-acid mutagenesis we further showed that DNA shape readout likely contributes to the cooperativity mechanism. *In vivo* enrichment of these cooperative sequences indicates different prevalence across Forkhead-Ets pairs, suggesting an additional layer of regulatory complexity. Finally, through an extensive assessment of the biological consequences of TF-cooperativity *in vivo* we found that leveraging the knowledge of cooperative TF binding increases the power to discover functions regulated by TF pairs. Specifically, for the Forkhead-Ets families we showed that a joint upregulation of FOXO1-ETV6 in Chronic Lymphocytic Leukemia (CLL) patients was associated with significantly higher survival rates.

2.2. Results

2.2.1. Quantitative modeling reveals contribution of higher order sequence features to TF-cooperativity

In a recent study Jolma *et al* have reported hundreds of cooperatively bound TF pairs through CAP-SELEX experiments [**Jolma *et al.* 2015**]. Their study demonstrated that TF cooperativity is highly prevalent and proposed that a majority of TF pairs do not directly interact, but form complexes mediated by DNA. Here we wanted to gain more insight into the mechanisms and uncover general rules that drive cooperativity among the identified TF pairs. Specifically, we hypothesized that features encoded in the DNA may contribute to the observed cooperativity. To test this, we devised a framework to predict the relative affinity of TF pairs based on DNA features using CAP-SELEX data. By ranking the importance of each DNA feature we could then identify those that potentially drive cooperativity.

CAP-SELEX data was obtained from Jolma *et al* [**Jolma *et al.* 2015**]. After reprocessing and quality control (**Appendix A**) we built models to predict the relative affinity of *k*-mers (DNA sequences of length *k*) bound to TF pairs in a procedure adapted from Riley *et a*l [**Riley *et al.* 2014**], which was previously employed to identify DNA features that determine binding of mainly single TFs [**Zhou *et al* 2015, Mathelier *et al.* 2016; Yang *et al.* 2017; Rube *et al.* 2018**]. Relative affinity was defined as the enrichment of a *k*-mer in the last cycle of the SELEX experiment relative to its input abundance. We then compared the performance of a basic model, which predicted the relative TF affinities from the mononucleotide sequence (*1mer*) with models that included more complex features, such as dinucleotides, trinucleotides (*2mer* and *3mer)* or DNA-shape (*shape*) [**Zhou *et al* 2015**]. The latter models may capture DNA stacking interactions, local-structure elements, and the overall DNA

structure, respectively (**Fig 2.1a**) [**Rohs *et al* 2009; Zhou *et al* 2015**]. The models were implemented as L2-regularized multiple linear regression (L2-MLR), and the impact of the features on TF-binding was assessed by calculating the relative improvements measured as R^2 differences on testing data (ΔR^2) between the full model (*1mer+shape/ 1mer+2mer/ 1mer+2mer+3mer*; 12, 20 and 84 features per position, respectively) and the reduced model (*1mer* - 4 features per position) using cross validation (**Appendix A**).

For each TF pair, we used the reported consensus sequences [**Jolma *et al* 2015**] as references for *k*-mer selection, considering all sequences up to a defined number of mismatches, along with their relative affinity values (**Appendix A**). One challenge that arises when working with composite TF binding models, is that their DNA binding regions are often very long and require high *k* values, which leads to low coverage of *k*-mers and hampers relative affinity estimates. Therefore, we developed a "trim-and-summarize" approach where we generated sets of *tiled k*-mers for each original *k*-mer of lengths no shorter than ten nucleotides, and summarized their effect on the prediction as a median R^2 (**Fig 2.1a, Fig S2.1a-b; Appendix A**). This resulted in 507 composite motifs with relative affinity estimates, comprising 77 unique TFs in 280 unique TF pairs. We found that models including higher-order features (*1mer+2mer, 1mer+2mer+3mer, or 1mer+shape*) performed consistently better than sequence-only models (*1mer*) (mean $P < 1.0 \times 10^{-6}$; Wilcoxon rank-sum test).

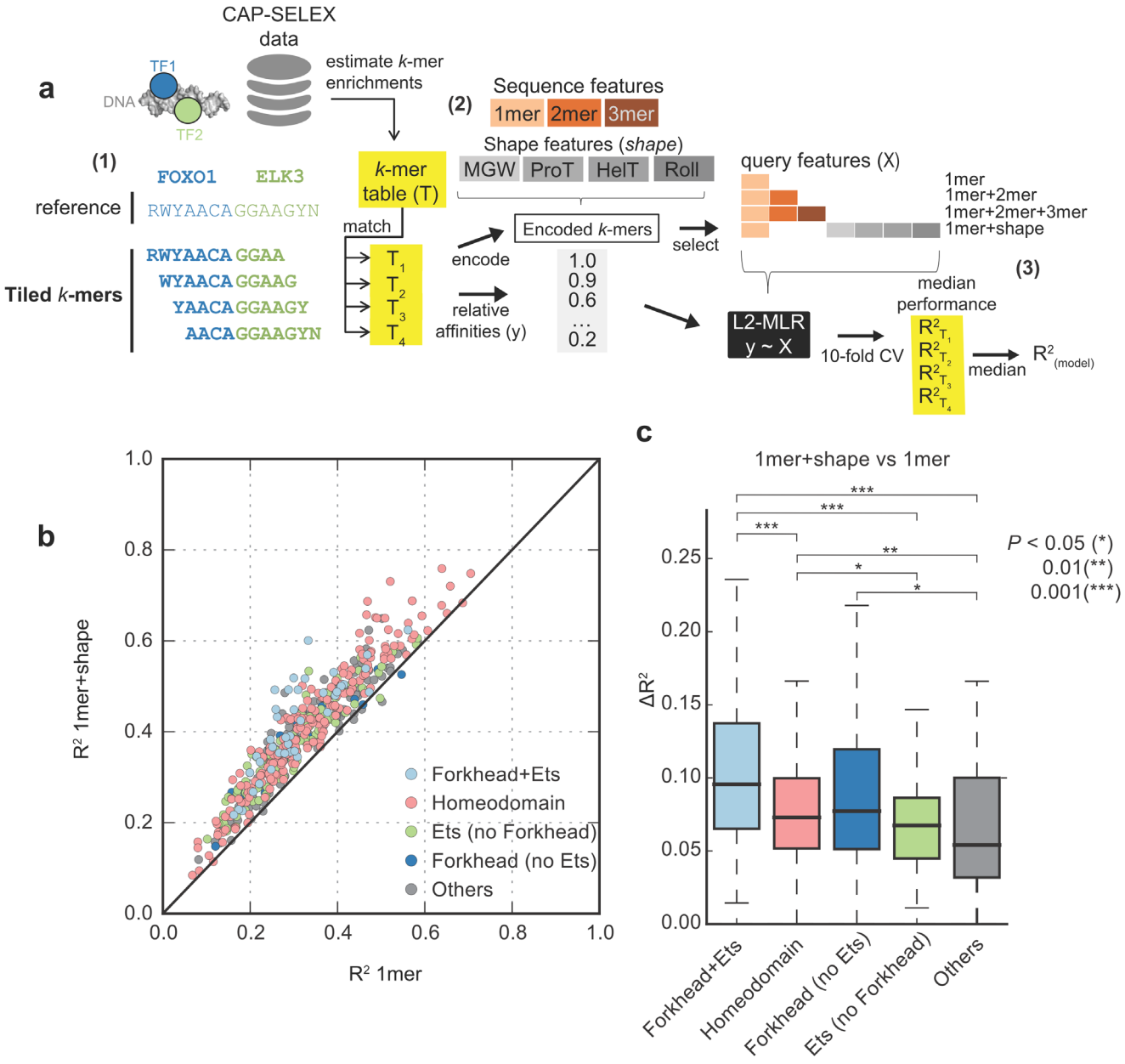

Figure 2.1. Addition of DNA-shape features improves combinatorial binding predictions in CAP-SELEX data.

(**a**) (1) Description of "trim-and-summarize" approach to obtain relative affinities for composite motifs (*k*-mers): a reference *k*-mer from CAP-SELEX data is trimmed from either side into multiple tiled *k*-mers with lengths no shorter than ten (blue regions assigned to consensus sequence for FOXO1 in reported *k*-mer, green regions assigned to ELK3 consensus sequence in reported *k*-mer) (2) L2-regularized multiple regression model (L2-MLR) are generated using DNA features as predictors and relative affinities as response variables (**Appendix A**). DNA sequence features (*1mer, 2mer, 3mer*) and DNA shape features (*MGW, ProT, HelT, Roll*), are tested in different combinations to assess their prediction contributions (3) A consensus improvement for each reference *k*-mer and model is obtained by cross-validation in each tiled *k*-mer table (10-fold CV) and calculation of the median tiled *k*-mer R^2 improvement for all cases. (**b**) Trim-and-summarize testing R^2 values are shown for each CAP-SELEX and reference *k*-mer combination, using tiled *k*-mers. Values above the diagonal indicate improvements in the testing set prediction performance when using mononucleotide and shape features together (*1mer+shape*, y-axis), with respect to models with only mononucleotide features (*1mer*, x-axis). Relevant TF family and TF family pairs are labeled by colors (Others = non-labeled families). (**c**) Trim-and-summarize ΔR^2 differences between 1mer+shape versus *1mer*, stratified by family, (*P* indicates Wilcoxon test adjusted *P*-values, corrected by Benjamini Hochberg's procedure).

This indicates that higher-order features are important for predicting TF cooperative binding (**Fig 2.1b**). Notably, since the predictions were done on held-out data, the positive ΔR^2 is not due to overfitting.

Overall, regardless of whether high-order sequence features are interpreted as "DNA shape" or as "dinucleotide dependencies", our results point towards their important role in guiding co-operative TF binding. In the following, we will only use *1mer+shape* models, which reasonably capture the improvements observed for the other models (**Fig S2.1c-d**).

2.2.2. Forkhead-Ets cooperativity is driven by DNA shape features

We next sought to assess whether DNA shape was important in driving cooperativity between particular TF families. To do so, we compared the ΔR^2 between full (*1mer+shape*) and basic (*1mer*) models across all TF pairs stratified by family. In agreement with previous studies we observed a moderate but significant increase in ΔR^2 for TF pairs involving homeodomain members [**Slattery *et al* 2011; Abe *et al* 2015; Yang *et al* 2017**]. (**Fig 2.1c**, median ΔR^2 = 0.07; P = 1.4 x 10^{-3}, one-sided Wilcoxon rank-sum test; (**Appendix A**). The strongest effect of DNA shape, however, was observed for pairs between Forkhead and Ets members (median ΔR^2 = 0.09, P = 1.8 x 10^{-5}). This shape-dependency was more pronounced than the ones obtained for each family alone (Forkhead and Ets median, both ΔR^2=0.07) thus highlighting its specificity. This is particularly interesting because crystallographic studies have demonstrated that DNA shape varies across Forkhead members [**Li *et al* 2017**]. In addition, Ets binding predictions have shown improvements by DNA-shape features [**Yang *et al* 2017**].

Due to the known bi-specificity of Forkhead TFs [**Nagakawa et al 2011**], we performed the same analysis after discarding DNA sequences containing the bi-specific Forkhead motif,

and obtained comparable results (**Fig S2.1d**). Overall, these observations highlight that DNA shape (or high-order features captured by shape) are important for predicting composite binding in a subset of TF-families, and particularly so for Forkhead and Ets members.

2.2.3. Prediction and validation of cooperativity between Forkhead and Ets

We next wanted to gain more mechanistic insight into the specific sequences presumably driving cooperativity between members of the Forkhead and Ets families and that could potentially explain TF binding to non-canonical sites. For that purpose, we used FOXO1 and ETS1 as a prototype Forkhead-Ets pair, and classified DNA sequences based on their level of cooperativity. Specifically, for each *k*-mer we compared the relative affinities for ETS1 and FOXO1 obtained from their respective High-Throughput SELEX (HT-SELEX) datasets (**Fig 2.2a**), and defined their cooperativity-potential as the ratio of predicted relative affinities between FOXO1:ELK3 (ETS1 paralogue) and the mean predicted relative affinity for FOXO1 and ETS1 (**Appendix A**). We found that the cooperativity potential dropped with increasing FOXO1 binding affinity, while the relative affinity of ETS1 had little effect on it (**Fig 2.2b**). This indicated that the FOXO1-binding strength determined the level of cooperativity, conclusion further corroborated by comparing representative DNA-sequences classified as non-cooperative, cooperative and highly cooperative (*ω*-none, *ω* and *ω*-high, respectively **Fig 2.2c**), which only differed in their Forkhead binding region. Similarly, Universal Protein-Binding Microarray data [**Berger *et al.* 2006**] revealed higher affinity of Forkhead members for *ω*-none than for *ω* sequences, while weak binding was observed for *ω*-high (**Fig S2.4a; Appendix B**). These results suggest that Forkhead TFs can bind to *ω*-none sequences on their own by recognizing a strong Forkhead binding site

while they rely on allosteric interactions with their Ets partner for recognizing *ω* (and possibly *ω*-high) sequences by forming a cooperatively bound ternary complex.

To validate the cooperativity predictions experimentally, we used isothermal titration calorimetry (ITC) to monitor changes in the dissociation constants (K_d) for the three DNA sequences with FOXO1 alone and in the presence of ETS1 (**Appendix B**). For FOXO1 alone we observed a 10-fold stronger binding for *ω*-none than for *ω* (K_d= 24 ± 3 nM and 352 ± 22 nM, respectively; $P < 0.01$, two-sided *t*-test **Fig 2.2d; Fig S2.2a**) while no interpretable results were obtained for *ω*-high of ITC. To test the effect of cooperativity on FOXO1 binding we titrated FOXO1 into a mixture of each DNA sequence and ETS1 and indeed observed a significant reduction in the K_d for *ω* (44 ± 11 nM; $P < 0.01$), but not for *ω*-none (26 ± 2 nM). This indicates cooperative binding between FOXO1 and ETS1 for *ω* but not for *ω*-none and thus validates our predictions. Since we were not able to measure *ω*-high using ITC, we resorted to measuring NMR chemical shift perturbations (**Fig 2.2e; S2.3a**), interpreted as weak, moderate or strong binding depending on the exchange regime (fast, medium, slow) to assess cooperativity between FOXO1 and ETS1. The results for *ω* and *ω*-none were corroborated qualitatively by NMR, as chemical shift perturbations switched from intermediate- to slow-exchange regimes, indicating an increase in binding affinity for *ω* in presence of ETS1 (**Fig 2.2e**). Additional peaks also show a similar behavior (**Fig S2.3b-c**). For *ω*-none, we observed slow-exchange (stronger binding) for both FOXO1 alone and in the presence of ETS1. Importantly, for FOXO1 on *ω*-high we observed chemical shift perturbations in the fast exchange regime, which enabled fitting and affinity determination. FOXO1 bound to *ω*-high three orders of magnitude weaker than *ω*-none (K_d = 21 ± 40 μM; **Fig 2.2f**) and changed to the intermediate exchange regime in presence of ETS1, indicating stronger binding and consistent with a cooperative interaction.

To generalize our cooperativity prediction to other FOXO1-Ets pairs binding in an equivalent conformation, we assessed the relative affinities of FOXO1 for sequences containing the binding patterns present in *ω*-none and *ω* (5'-GTAAACA-3' vs 5'-AACAACA-3') in single (HT-SELEX) and paired (CAP-SELEX) data. As expected, FOXO1 showed significantly higher relative affinities for *ω*-none versus *ω* in HT-SELEX late rounds (**Fig S2.3b, Appendix A**). In CAP-SELEX, however, relative affinities for *ω*-none and *ω* were similar for the majority of datasets comprising FOXO1 paired with Ets, Homeodomain and GCM members.

In summary, our framework to predict cooperativity for FOXO1-ETS1 pairs based on combining single-TF HT-SELEX and paired-TF CAP-SELEX data was experimentally validated by ITC and NMR. Our findings suggest a widespread mechanism whereby Forkhead TFs recognize non-optimal binding sites through cooperative interaction with specific partner TFs.

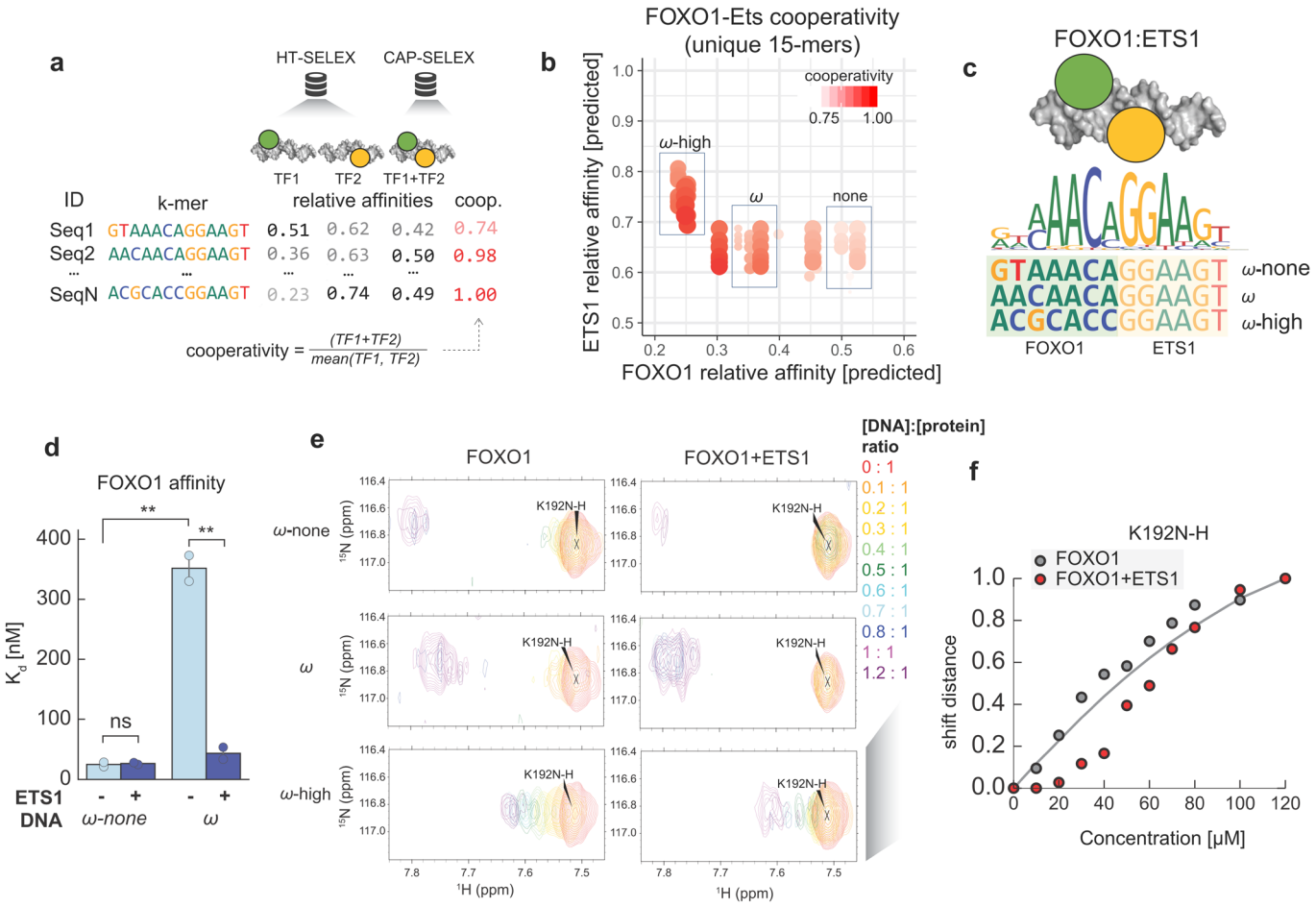

Figure 2.2. Prediction and validation of cooperative binding sites from SELEX data

(**a**) Workflow describing the calculation of the cooperativity potential between *k*-mers from CAP-SELEX and matched HT-SELEX data. Predicted relative affinity values of *k*-mers from CAP-SELEX data are scaled by the mean value observed in HT-SELEX for the matched single TF datasets. (**b**) Comparison relative affinity predictions for ETS1 and FOXO1 for 15-mers using HT-SELEX. K-mers are weighted by their estimated cooperativity potential using CAP-SELEX of FOXO1:ELK1 and HT-SELEX datasets for FOXO1 and ETS1 (ELK1 paralogue). Y- and X-axes indicate predicted relative affinities for ETS1 and FOXO1 datasets, respectively, using 1mer+shape models. (**c**) (*top*) Cartoon description of binding mode for the FOXO1-ETS1 ternary complex. (*bottom*) Sequences chosen for validation from regions of none (ω-none), moderate (ω) and high (ω-high) cooperativity are shown and aligned with Forkhead-Ets composite motif. Green and yellow highlighted regions indicate Forkhead and Ets binding regions, respectively. (**d**) Dissociation constant ITC measurements for FOXO1 with ω-none and ω DNA sequences in the absence and presence of ETS1 (ns = non-significant, ** = t-test $P < 0.01$). (**e**) $^{1}H-^{15}N$ HSQC spectra focused on K192N-H NMR titration peak for FOXO1 with increasing DNA concentration of ω-none, ω and ω-high. Colors indicate DNA to protein concentration ratios. Fast to intermediate regime change in titration peak for ω-high highlights cooperativity. (**f**) Titration curve of FOXO1 binding to ω-high using K192N-H peak, without (gray dots), and with ETS1 (red dots). Gray line indicates titration fit without ETS1.

2.2.4. Quantitative modeling reveals structural insights into DNA-ternary complexes

Based on the above observations we wanted to gain more structural insights into the cooperative Forkhead-Ets interaction as well as other TF pairs. So far, we showed that DNA shape features are important to predict binding of TF pairs (**Fig 2.1c**), and that, in the case of FOXO1:ETS1, differences between high- and non-cooperative DNA sequences were locally restricted (**Fig 2.2c**). We therefore hypothesized that position-specific DNA shape features may determine the cooperativity potential of DNA sequences. To test this, we used our quantitative modeling framework to calculate the importance of DNA shape features at each position along the composite motifs for all TF pairs in the CAP-SELEX data. Specifically, we compared models with and without all DNA shape features at a given position, and reported the maximum ΔR^2 per position, adapting an approach developed by [**Yang *et al.* 2017**] (**Appendix A**). For each composite motif, we thereby obtained a "shape profile" that captures the importance of DNA shape at each position for predicting the relative affinities of a TF pair (**Fig 2.3a**; benchmark with HT-SELEX data in **Fig S2.5**).

To globally explore the positional effects of shape profiles, we scaled them to the same length and grouped them into five groups using unsupervised clustering (**Fig 2.3b**; **Fig S2.5a**; **Appendix A**). All clusters are characterized by a single peak in the shape profile, indicating that the effect of DNA shape is localized to a specific region along the protein-DNA interface. We observed a significant enrichment of TF pairs containing Forkhead members in cluster 1 (odds ratio=3.1, adjusted *P* value < 0.1, Fisher's exact test). Other TF families and TFs were specifically enriched in other clusters (**Fig S2.5b**). This result indicates that some TF families have a conserved preference for shape or high-order features when interacting with other TFs. Interestingly, shape profiles for pairs that include Forkhead members in cluster 1 peaked at the Forkhead binding site (**Fig 2.3c**).

Together with the results from the previous section, this suggested that shape readout at the Forkhead region might guide the cooperative interaction between Forkhead and partner TFs. This mechanism was further supported by comparing the shape profiles of exemplary Forkhead-Ets TF pairs with the profiles of single Forkhead and Ets TFs, obtained from HT-SELEX: While FOXO1 by itself still showed a higher shape preference than its Ets partner, the maximum value of the profile was shifted by at least three positions relative to the one of the TF pair. Similarly discrepant patterns between single and composite profiles were observed for other Forkhead members such as FOXI1 with Ets TFs (**Fig S2.3d-e**). These results suggest that the shape profiles are likely related to Forkhead-Ets cooperative binding for many members of the respective families, and unlikely due to individual TF binding.

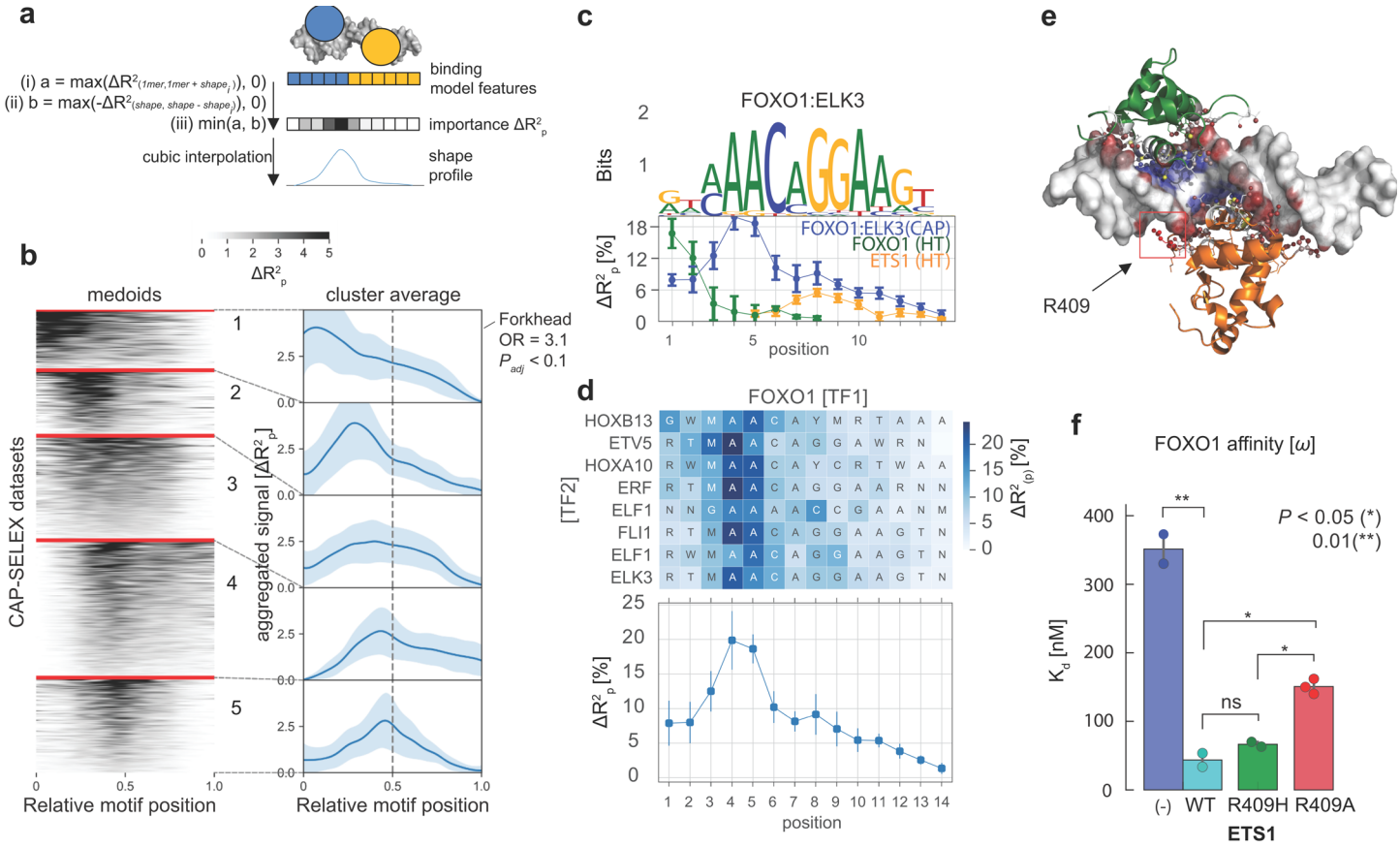

Figure 2.3. Clustering of shape improvements by position in CAP-SELEX data reveal combinatorial TF binding shape-recognition biases.

(**a**) (*top*) Scheme describing shape profiles calculation. For a homo- or heterodimeric protein-DNA complex, shape features are individually added to assess their relative contribution to the global increase in performance. Averaged contributions are transformed by interpolation to a curve representation. (**Appendix A**) (**b**) (*left*) PAM clustering of shape profiles across all CAP-SELEX models analyzed. Each row shows a composite motif (N=438). Five representative clusters are separated by red lines. (*right*) Average shape profiles for each cluster. Blue shades indicate one standard deviation. Enrichments for TF families within clusters are labeled (OR=odds ratio). (**c**) (*top*) Forkhead-Ets FOXO1:ELK3 motif, (*bottom*) ΔR^2_p changes (as percentages) in FOXO1:ELK3 CAP-SELEX data (blue line). Additional lines indicate equivalent values for FOXO1, and ETS1, an ELK3 paralogue (green and orange, respectively) calculated from HT-SELEX [Jolma et al. 2013]. Error bars indicate windowed average standard deviation, with window value of 4. (d) (top) ΔR^2 per position changes are shown for selected CAP-SELEX data that contain FOXO1 in combination with other binding partners of similar binding topology. IUPAC DNA symbols in heatmap indicate aligned k-mers. (bottom) Column averages for heatmap values. Error bars indicate standard deviations in each position. (**e**) FOXO1-ETS1 ternary complex. ETS1 residue R409 interacting with DNA minor groove is highlighted in a red box (PDB ID: 4LG0). Visualization was enhanced by PDIviz [Ribeiro et al 2015]. (**f**) Dissociation constant measurements using ITC for FOXO1 binding to ω DNA sequence upon addition of ETS1 wild type and selected mutants (* indicate adjusted P values obtained using two sided t-test)

2.2.5. Site-directed mutagenesis reveals ETS residue R409 as driver of cooperativity

Given that the shape profiles were highly conserved across many members of the Forkhead and Ets families (**Fig 2.3d**), we wanted to investigate whether protein-DNA interface properties at the peak of the profile may confer cooperativity. Through assessment of the available crystal structure for FOXO1:ETS1 bound to DNA, we observed an arginine residue (R409) of ETS1 interacting with the minor groove at the position with highest shape relevance of the FOXO1 binding site (**Fig 2.3e**; PDB ID: 4LG0) [**Choy et al. 2014**]. This agrees with the high relevance of Minor Groove Width features for binding prediction in our models (**Fig S2.5c**). Given the strong conservation of positively charged residues in this position across Ets family members (94%; **Appendix A**), we hypothesized that the DNA-cooperativity between Forkhead and Ets is mediated by this residue.

To test this, we performed site-directed mutagenesis of the ETS1-residue in question (R409) and monitored the changes in the dissociation constants of FOXO1 for one of our previously validated cooperative DNA sequences (ω) using ITC (**Fig 2.3f; Appendix B**). Replacement with alanine (R409A) significantly reduced the cooperative effect between ETS1 and FOXO1 with a significant drop in binding affinity of FOXO1 to ω relative to wild type ETS1 (K_d = 151 ± 11 nM in R409A vs 44 ± 11 nM in WT; P = 2.4 x 10^{-3}). In contrast, replacing the arginine with another positively charged residue (Histidine; R409H), resulted in a FOXO1 binding affinity similar to wild type ETS1 (K_d = 67 ± 4 nM) thus retaining the cooperative interaction. To study whether this effect solely depends on that specific residue, we tested a neighboring residue (Y410A), and observed almost no changes in FOXO1 affinity (K_d = 313 nM; **Fig S2.5f**). We concluded from these analyses that the cooperativity between FOXO1 and ETS1 is indeed mediated by the interaction of R409 of ETS1 and the DNA minor groove opposite the FOXO1 binding site. As the affinity of FOXO1

in presence of the mutant ETS1 was still higher than for FOXO1 alone, we cannot exclude that other residues may also contribute to the cooperativity.

2.2.6. Cooperativity between Ets and Forkhead determined *in vitro* is relevant *in vivo*

Having demonstrated, mechanistically analyzed, and experimentally validated cooperativity between members of the Forkhead and Ets TF families *in vitro*, we next wanted to assess whether these findings can be translated to *in vivo* systems based on ChIP-Seq data and whether TF-TF interactions can aid in explaining TF binding events.

We first tested whether DNA shape was equally important for predicting co-occupied ChIP-Seq regions as it was for predicting cooperative binding based on CAP-SELEX data. To do so we used a classification framework similar to [**Mathelier *et al* 2016**], to compare models based on motif match (*PWM*) only and motif match plus DNA-shape features (*PWM+shape*) for predicting co-occupied ChIP-Seq regions between pairs of TFs (**Appendix A**). Overall, we obtained similar results as for the *in vitro* data with 105 peak sets showing improved classification performance after addition of *shape* features in mapped TF cooperative sites ($P < 0.0001$; one-sided Wilcoxon rank-sum test) (**Fig 2.4a**). In agreement with the *in vitro* data analyses, TF pairs that include a Forkhead family member particularly benefited from DNA shape ($P = 0.03$; one-sided Wilcoxon rank-sum test) (**Fig 2.4b**).

When comparing the shape profiles obtained from the *in vivo* and the *in vitro* data (**Appendix A**), we observed a strong agreement for 40% of them (FDR = 10%) (**Fig 2.4c**); median spearman correlation = 0.25). This suggests that DNA shape plays a similar role in driving cooperativity *in vivo* for specific TF pairs. Among the correlated profiles were several Ets-Forkhead pairs e.g. FOXO1:ETV4 (**Fig 2.4d**).

We next wanted to assess to which extent Forkhead:Ets members bind cooperatively versus non-cooperatively *in vivo*. To do so, we calculated the enrichment of the ω and ω-none motifs in co-occupied ChIP-Seq data for members of both Forkhead and Ets families - assuming that the FOXO1:ETS1 ω and ω-none *k*-mers are representative of other Forkhead-Ets members (**Appendix A**). We found both ω-none and ω enriched among the co-occupied regions of the 126 TF pairs using the single occupied regions as background (**Fig 2.4e**). 18 TF pairs showed enrichment for both the non-cooperative as well as cooperative sequences (ω-none and ω) confirming the co-existence of cooperative and non-cooperative binding patterns between the same pairs of TFs *in vivo.* Another 29 pairs were only enriched for either cooperative or non-cooperative sequences (5 and 24 respectively). These results suggest variable degrees of cooperativity between Forkhead-Ets TF pairs, thus hinting at a TF-pair specific cooperativity, which adds an additional layer of regulatory complexity *in vivo.*

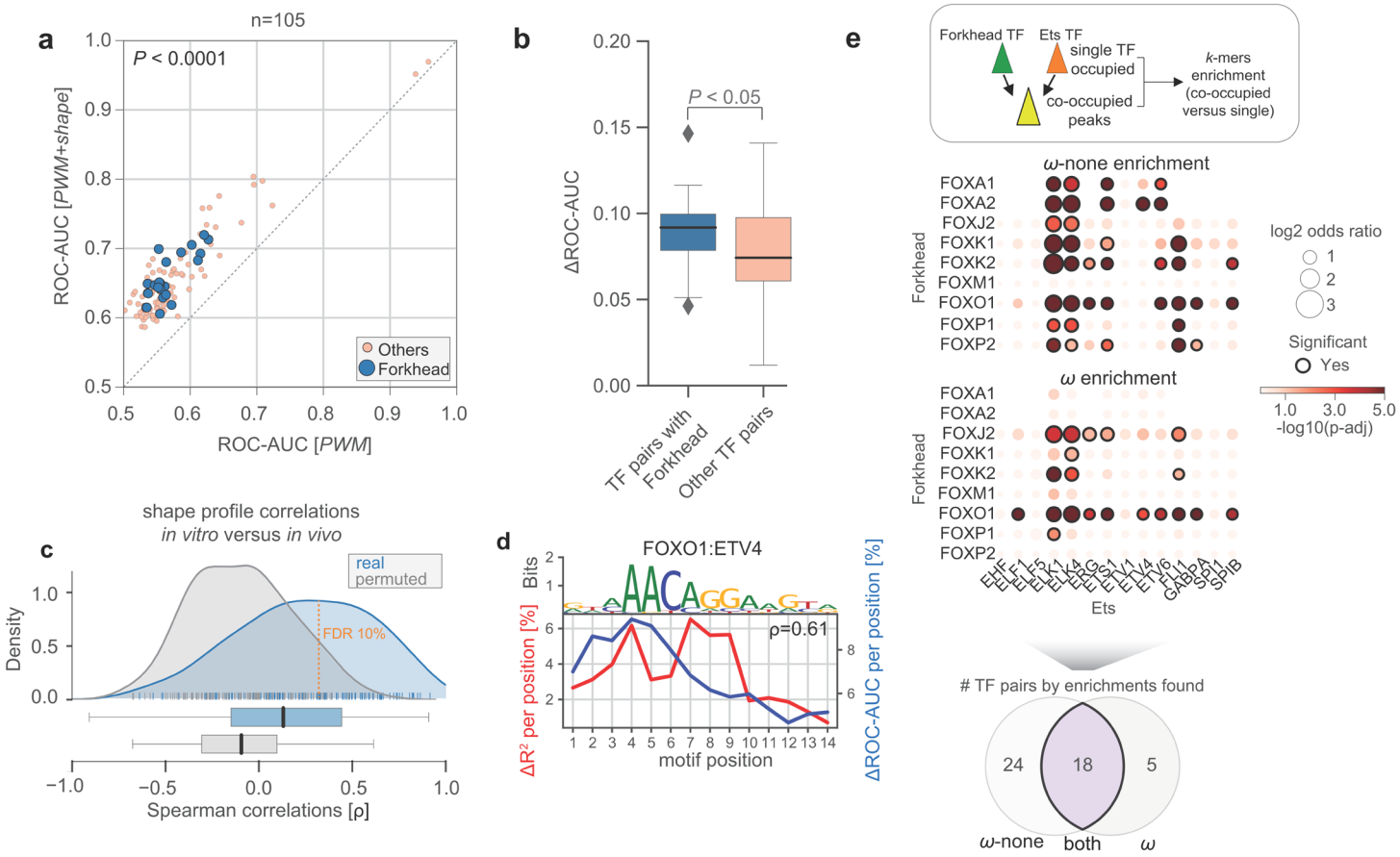

Figure 2.4. Cooperative TF binding agreement between SELEX and in vivo data.

(**a**) Classification performance comparison between PWM+shape (y-axis) versus PWM models (x-axis) in regions that were selected for being co-bound by ChIP-seq for TF pairs present in CAP-SELEX data (*N*=105). Classification performance is measured by the Area Under The Receiver Operating Characteristic Curve (ROC-AUC). Blue points indicate TF pairs with Forkhead as one of its members. (**b**) ROC-AUC differences between classification models for datasets containing at least one Forkhead member (blue) and all other TF pairs (pink). *P* value obtained using Wilcoxon rank-sum test. (**c**) Spearman correlation distribution of performance changes per position in in vitro (R^2), and in vivo (ROC-AUC) between matched TF pairs. Orange line indicates FDR 10% cutoff for positive correlations (**d**) (*top*) Forkhead-Ets composite motif model between FOXO1 and ETV4, (*bottom*) aligned performance changes per position observed in in vitro (CAP-SELEX; red line) and in vivo (ChIP-seq; blue line). ρ indicates effect size. (**e**) (*top*) Scheme illustrating co-enrichment calculations for ChIP-seq regions co-occupied between Forkhead and Ets versus single TF occupied. (*middle*) Dot plot showing ω-none and ω *k*-mer enrichments between co-occupied and single TF peaks (Adjusted *P* value obtained from a Fisher's exact test between fraction of regions with motif in co-occupied peaks versus fraction of region with motif in single TF occupied peaks) (*bottom*) Venn diagram indicating significant observation for tested *k*-mers, and number of datasets with enrichments for both.

2.2.7. Inference of TF-phenotype associations using TF-cooperativity *k*-mers

Having shown that cooperativity is both prevalent and specific *in vivo* we further investigated its potential functional impact. We first wanted to assess whether certain biological processes are specifically regulated by cooperative TF binding. To do so, we assumed that genes regulated by a pair of TFs should reflect biological functions common to both TFs and that these functions should be captured by gene ontology terms. Further, we defined TF-pair-to-gene links by mapping regions co-occupied by both TFs (using ChIP-Seq from ReMap [**Cheneby et al 2018**]) to target genes (using GREAT [**McLean *et al* 2010**]; **Appendix A**).

With this, we devised an "Ontology Association Probability" that quantifies relationships between each TF pair and an ontology term using logistic regression. Briefly, for each TF pair we modeled their membership in a given ontology term based on the number of their target genes and regulatory elements (normalized as z-scores) mapped to it (**Fig 2.5a; Appendix A**). To test the effect of cooperativity on the ontology association probability we compared models with only ChIP-Seq data as features ("*peaks*") to models with only cooperativity k-mers *("k-mers")* and models using both ("*peaks+k-mers*"). For all models, we observed higher association probabilities for ontology terms annotated to the TF pair ("TF1 and TF2") than for random background terms ($P < 0.001$; Wilcoxon one-sided test) (**Fig 2.5b**). The highest associations were obtained for models considering genes regulated by cooperatively bound peaks (*peaks+k-mers*; $P < 0.01$), emphasizing the role of TF cooperativity in regulating specific processes.

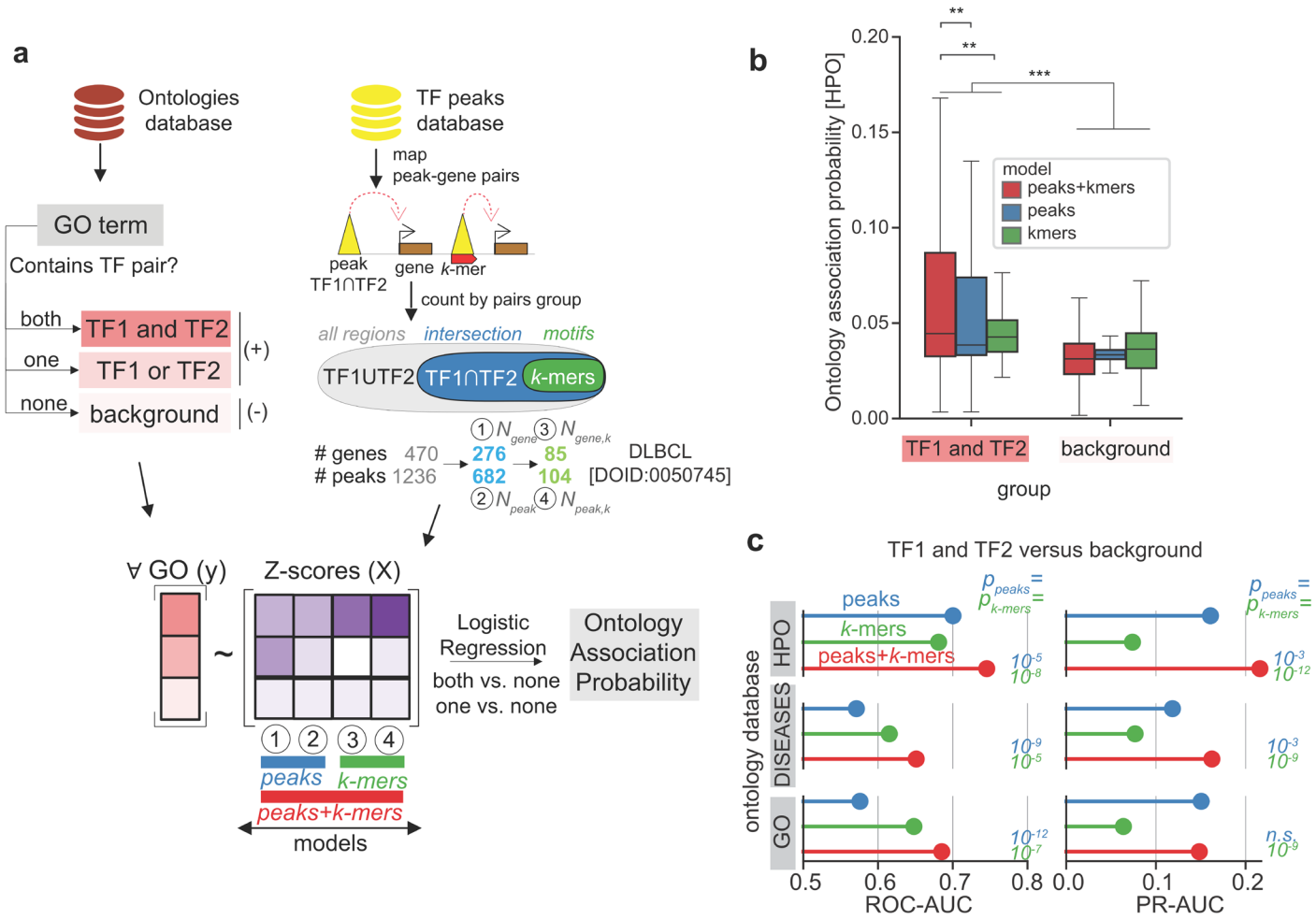

Figure 2.5. Inference of TF-phenotype associations using TF-cooperative k-mers.

(**a**) Scheme illustrating calculation of Ontology Association Probability values using TF-TF *k*-mers, ChIP-seq data and ontology databases. For each TF-pair and ontology combination, four metrics describing the numbers of genes and peaks proximal to the ontology-related genes are converted into Z-scores (**Appendix A**). From peaks assigned to for both TFs (TF1UTF2, gray oval), the sub-selection using the co-occupied peaks (TF1∩TF2, blue oval; features 1 and 2) allows calculating N_{gene} and N_{peak}, and co-occupied peaks with TF-TF *k*-mers allow calculating $N_{gene,k}$ and $N_{peak,k}$ (green oval; Features 3 and 4). Ontologies are labeled by the presence of both TFs (TF1 and TF2), one (TF1 or TF2), or none (background) in the ontology. Models with different combinations of features are tested (peaks = 1 and 2. (Blue); *k*-mers = 3 and 4. (Green); peaks+*k*-mers = 1, 2, 3 and 4. (Red)). (**b**) Distributions of association probabilities for Human Phenotypes Ontology (HPO) terms for terms labeled as "TF and TF2" versus background terms are shown (* indicate Wilcoxon rank-sum test *P* values). (**c**) Classification task performances in the assessment of "TF1 and TF2" versus "background" terms in three ontology databases. ROC-AUC and PR-AUC indicate areas under the ROC and Precision-Recall curves. *P* values are derived from 10-fold cross validation metrics comparisons between *peak+k-mers* and reference approaches using an independent *t*-test, after Benjamini-Hochberg correction.

We next sought to use the derived metric for discovering biological functions of cooperatively bound TFs. In particular, the association probability can be used to predict ontology terms of all TF pair combinations (**Appendix A**). Using defined ontology terms common to both TFs as a gold standard, performance metrics indicated that predictions were on average better when using *peaks+k*-mers versus *peaks* or *k*-mers. This was the case in three tested ontology databases (GO; DISEASES; HPO) [**Ashburner et al 2000; The Gene Ontology Consortium 2019; Köhler et al 2018; Pletscher-Frankild et al 2015**] and irrespective of the performance metric (mean ROC-AUC = 0.70 (*peaks+k-mers*), 0.61 (*peaks*), and 0.64 (*k-mers*) (P = 6.7 x 10^{-8}); mean Area under the Precision-Recall Curve (PR-AUC)) = 0.18, 0.15 and 0.06 (P = 2.3 x 10^{-9})) (**Fig 2.5c**). Interestingly, the classification performance of ontologies related to both TFs is higher than the one where only one TF of the pair is associated to the ontology ("TF1 or TF2"; mean ROC-AUC = 0.66; mean PR-AUC = 0.14; **Fig S16a**). These results indicate that processes cooperatively regulated by two TFs can be distinguished from those regulated by each TF individually.

To capture the strongest associations between TF pairs and terms across all used ontology databases, we defined a model-dependent signal-to-noise threshold on the association probabilities (**Appendix A**); this recovered 6600 strong associations with high probabilities and both TFs as members of the ontology (**Fig S2.5b**). We considered this number an underestimate limited by the availability of ChIP-Seq data since applying a variation of our model using only TSS *k*-mers identified cooperative TFs interacting with partner TFs in cell differentiation and disease (**Fig S2.5c**) that for which no ChIP-seq exists. Following up on our previous results we focused on Forkhead-Ets pairs and recovered strong associations between specific partners and ontology terms for 20 of them (**Fig 2.6a**). FOXO1 showed the highest number of associations with different TFs (nine),

suggesting a multifunctional role for this TF through cooperative interaction with multiple partners. FOXO1 was most strongly associated with sarcoma (DOID:1115, with ELF1), and squamous cell carcinoma (DOID:5520, with ELF3) (*k*-mer = WAAACAGGAAG for both terms; average *k*-mers z-score > 5). This is in agreement with previous reports proposing FOXO1 as a prognostic marker in sarcoma [**Zhang *et al* 2009**]. Moreover, ELF3 has been proposed as a marker in squamous cell carcinoma [**AbdulMajeed *et al* 2013**] and ELF members have been generally recognized to play a role in sarcomas [**Ando *et al* 2016**]. In light of these results and their strong agreement with the literature, we hypothesized that expression levels of FOXO1 together with predicted TF partners could be a potential readout to interrogate clinical cancer data.

We examined this concept using available data on lymphoid leukemia patients to examine the effect of predicted associations with cooperative binding of FOXO1 and ETV6 (DOID:0050745, *k*-mer GAAAACCGAANM; mean *k*-mers z-score = 3.2). Specifically, we stratified patients in a Chronic Lymphocytic Lymphomas (CLL) cohort [**Dietrich *et al.* 2018**] into high/low expression levels for both TFs (**Appendix A**), to explore their usage as prognostic markers. Strikingly, we obtained a significant increase in overall survival when both TFs were highly expressed (Hazard Ratio (HR)=0.21, 95% CI 0.10-0.45; $P = 6.5 \times 10^{-5}$) (**Fig 2.6b**). This association was not found when considering each factor separately, and it was not confounded by p53 and IGHV mutation statuses (HR=0.19, 95% CI 0.07-0.48, $P = 5.0 \times 10^{-4}$, **Fig S2.5d**). Importantly, this is the strongest association to survival among all FOXO1-Ets combinations, for which ChIP-seq data was available. Together with reports of FOXO1 and ETV6 as putative tumor suppressors in lymphomas [**Xie et al. 2012; Peker et al. 2013**] this suggests an important role of this TF pair in lymphoid leukemia.

Overall, our results demonstrate the increased power of cooperative TF-binding models applied to *in vivo* data for an unbiased screening of novel TF pairs as potential drivers of function and disease.

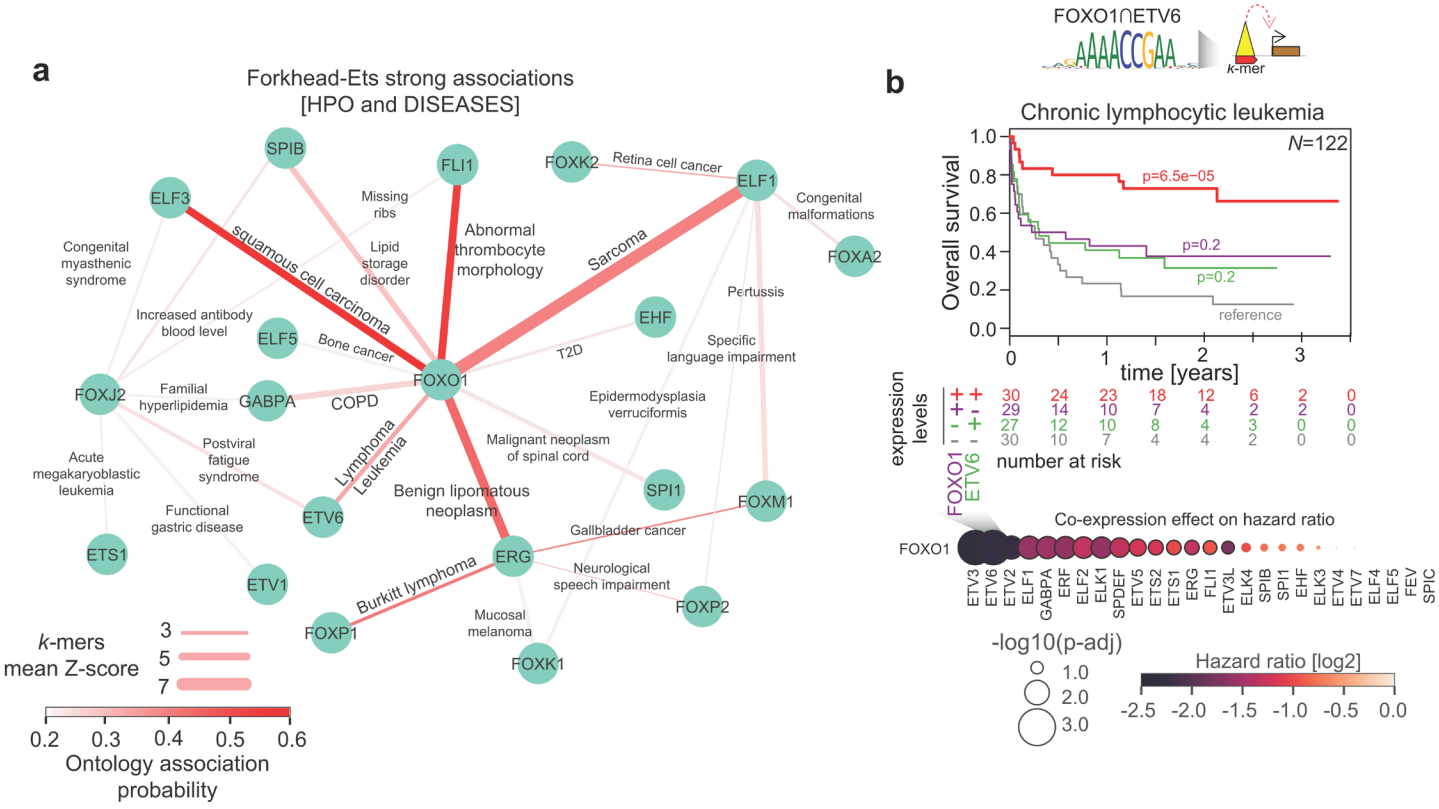

Figure 2.6. Forkhead-Ets cooperativity related association to function and disease

(**a**) Strong associations networks between Forkhead-Ets families using HPO and DISEASES ontologies (**Appendix A**). Nodes indicate TFs and edges indicate ontologies names. Edge width indicates relevance of features 3 and 4 the final association probability value (COPD=Chronic obstructive pulmonary disease; T2D=Type 2 Diabetes) (**b**) (*top*) Cartoon describing association between Forkhead+Ets *k*-mer GAAAACCGAANM and lymphoma associated genes through intersected peaks. (*middle*) Kaplan-Meier plot of overall survival in CLL patients when using FOXO1 and ETV6 expression medians ("high/low" defined as above/below median, and labeled as + and -, respectively). *P*-values are derived from two-sided log-rank comparison with respect to -/- expression levels for both FOXO1 and ETV6 (data from [**Dietrich et al. 2017**]).

2.3. Discussion

Here we provide a novel framework to study different types of TF binding data for single and co-binding TFs in *vitro* and *in vivo*, allowing to systematically gain structural insights into TF cooperative binding, and revealing their functional and disease-related relevance. Statistical learning proved to be an integral part to understand the contributions of DNA features to TF binding, such as in approximating positional relevance of nucleotide interactions and DNA-shape features in TF binding models. Thereby our models provide a platform for generating hypotheses about the possible consequences of disruptions in DNA-shape readout [**Slattery *et al.* 2011; Yang *et al* 2017, Kribelbauer *et al.* 2017, Rube *et al.* 2018**]. Importantly, applying those concepts to cooperative TF binding data, we derived specific and conserved binding preferences across TF families. Using FOXO1 and ETS1 as representative members of the Forkhead and Ets families, we demonstrated that such conserved TF-interactions are clearly linked to DNA-shape readout with stronger effect sizes than the ones for homeodomain pairs [**Slattery et al 2011**]. We reinforced this argument by identifying a conserved residue that mediates cooperativity in Ets family members. This particular arginine residue happens to harbor multiple DNA-binding domain polymorphisms [**Barrera *et al.* 2016**], suggesting that the extent of this particular cooperativity between Forkhead-Ets members can be prone to variation across healthy individuals.

Our work presents a major methodological advance over recent studies on the quantitative assessment of DNA-shape readout and its contribution to TF binding, which are limited by data sparsity due to long binding (composite) motifs. To estimate feature preferences for such motifs, we introduced a "trim-and-summarize" approach allowing the reliable quantification and comparison between models considering motifs spanning a mean of 18 base pairs in CAP-SELEX data from Jolma et al. (**Appendix A**). Given the reasonable

agreement of our results with *in vivo* data, this approach could prove useful in integrating low-coverage SELEX data with other studies of higher data quality [**Zhang *et al* 2018; Rastogi *et al* 2018**], as well as screening for novel cooperative TF binding sites *in vivo*.

TF binding has been associated to chromatin regulation [**Grubert *et al* 2016**] and disease [**Deplancke *et al* 2016**], yet cooperative binding has not been systematically analyzed in such contexts. The knowledge of TF-TF allostery can be used to predict co-occupied TF regions and annotate cryptic binding sites [**Narasimhan *et al* 2015**]. As genetic disruptions in such TF-cooperativity regions are important to understand failures in developmental programs [**Slattery *et al* 2011**] and disease [**Iwata *et al* 2017**], there is a requirement for models that predict preferences for TFs acting in combination and the functional consequences of such events. Here, the integration of cooperative TF *k*-mers with ontology associations of TFs allowed us to thoroughly examine potential functional consequences stemming from genome loci targeted by cooperative TF-binding. Although other studies have associated composite motifs to specific cell types using *in vivo* data before [**Jankowski *et al.* 2014; Guturu *et al.* 2013**], we successfully demonstrate that incorporating a new layer of knowledge on the degree of cooperative binding gives a significant leverage in identifying biological processes specific to TF pair binding. In fact, the knowledge of cooperative *k*-mers translates into better TF-ontology predictions and could thus increase the extent of our functional knowledge on cooperative TF binding and its underlying biology (**Fig 2.7**). Importantly, we release our current predictions for community examination of new mechanistic interactions between TF pairs.

Given the considerable amount of strong associations between TF pairs and disease, the clinical power of revealing such functional connections in a systematic manner is not to be underestimated. Our investigation of the TF pair FOXO1:ETV6 and its cooperativity-driven association with overall survival in CLL exemplifies this clearly and is reinforced by the

observation of a FOXO1/ETV6 gene fusion in leukemia patients [**Stengel et al 2018**]. Both FOXO1 and ETV6 have been described as putative tumor suppressors in lymphomas [**Xie et al. 2012; Peker et al. 2013**], yet the extent of the cooperativity-driven functional impact in Leukemia relative to other FOXO1-Ets combinations has not been understood nor quantified. Future work will be required to understand whether this particular mechanistic relationship occurs prior or after FOXO1 mutations [**Trihn et al 2013**] or whether it represents an independent event in cancer progression in the first place. Systematic modeling of such associations and their network interdependencies remains, however, an indispensable component in leveraging TF cooperativity for functional interrogation and prioritization of disease-related TF combinations.

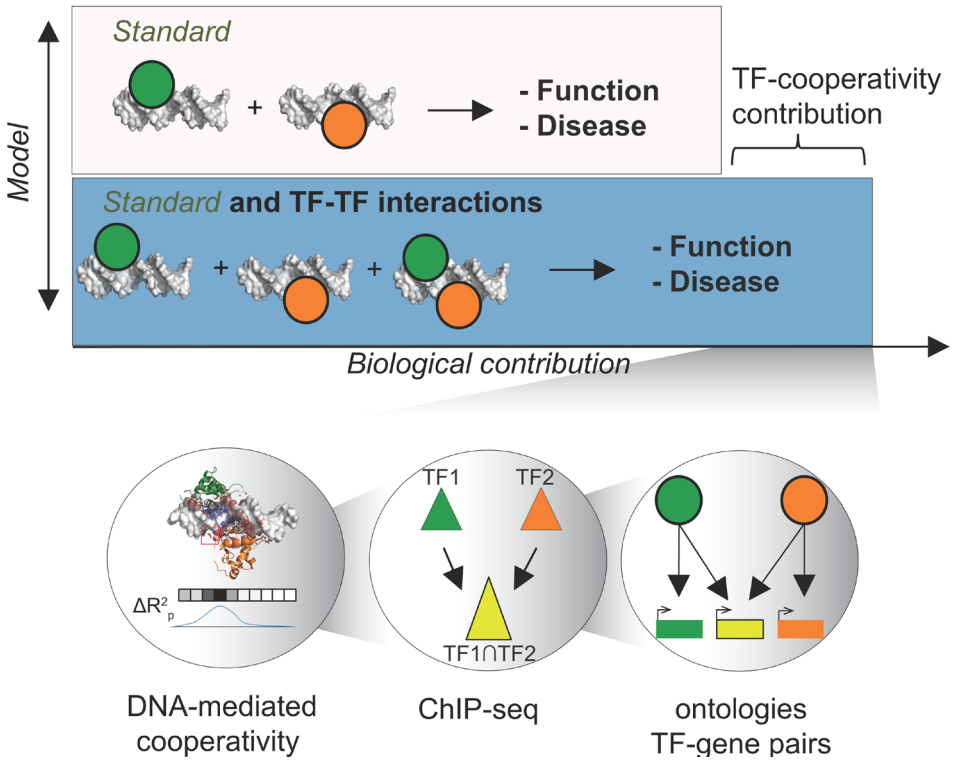

Figure 2.7. Model to estimate cooperative TF-binding contribution to TF-ontology associations.

(*top*) Illustration of different types of models that describe associations of TF with function and disease. The Standard model describes the contribution of single TF-DNA binding, which can be improved by the addition of TF-TF interactions for an enhanced understanding of function and disease. This is translated into an overall increased discovery of strong phenotypes associated to TFs when acting in combination. (*bottom*) Information used in this work to describe TF cooperativity and reveal TF-cooperativity linked processes (DNA-mediated cooperativity = TF-TF *k*-mers and prioritization of important binding modes; ChIP-seq = co-occupied peaks for TF pairs; ontologies and TF-gene pairs = associations between co-occupied regions by TF pairs and their associated genes, linked to function through ontology data).

Chapter 3

BDNF promotes enhancer accessibility essential for gene activation and exon usage in neurons

In this chapter, I describe the analyses and results of a collaboration project that explores the molecular response of mouse cortical neurons by multi-omics data integration. Methodologies and experimental data generation behind this work have been initially conceived by Vikram Ratnu, in a collaboration between the groups of Kyung Min Noh and Judith Zaugg at EMBL Heidelberg. I carried out all computational analyses and suggested further validation experiments that were part of the results sections. The work has been described in the following manuscript:

Ignacio L. Ibarra*, **Vikram S. Ratnu***. Lucia Gordillo, Luca Mariani, Katy Weinand, Martha L. Bulyk, Judith B. Zaugg & Kyung-Min Noh (2019). BDNF promotes enhancer accessibility essential for gene activation and exon usage in neurons. *In preparation.*

3.1. Introduction

How neuronal activity is linked to function of the nervous system is a question of paramount importance in neuroscience. Molecular mechanisms regulating synaptic transmission and plasticity have been intensely studied for decades, and their relevance is acknowledged as their functional deficits can lead to neurogenetic disorders, linked to learning and behavior on many levels [**de la Torre-Ubieta et al 2018**].

It has been understood that neuronal activity is mainly determined by changes in gene expression patterns, which is tightly regulated at the genome level through epigenetic and accessibility markers [**Peixoto *et al* 2012**]. Multiple studies have assessed the impact of such markers on neuronal activity and their species-wide conservation as a way to prioritize regions related to developmental and intelligence disorders [**Reilly et al 2015**]. However, studies focusing on the link between genome-wide accessibility changes and gene expression programs at intermediate and late time points after neuronal stimulation are lacking. Nor has it been shown how different stimuli confer specific neuronal activity by differentially impacting the accessibility and gene expression landscape.

Here we present a study that profiles both accessibility and gene expression changes in a genome-wide manner to identify features that determine mouse cortical neuron response to stimuli. Through joint temporal profiling of chromatin accessibility and gene expression upon three stimuli (BDNF, KCl and Forskolin) in mouse primary cortical neurons, we delineated molecular rules determining chromatin-to-expression programs. Our genome wide analysis pointed at regulatory factors mediating neuronal response to our stimuli in a shared and treatment-specific way, allowing us to identify the underlying mechanisms as well. Specifically, we found and validated an axis between co-regulators and co-repressors controlling the expression response in BDNF, whereas the neuronal response upon KCl

stimulation was mostly determined by an interplay between CTCF and accessibility-mediated transcriptional changes.

3.2. Results

3.2.1. Stimulus specific biphasic transcription in response to neuronal activity

We investigated how different neuronal stimuli (BDNF, KCl, Forskolin) impact gene expression programs in mouse primary cortical neurons. BDNF activates p75NTR and Trk receptor tyrosine kinases which trigger signaling pathways involved in neuronal plasticity [**Chao et al 1995; Poo et al 1991**]. KCl (55 mM) induces membrane depolarization, calcium influx [**Greer et al 2008**], and calcium-dependent signaling pathways leading to changes in gene expression. Forskolin increases secondary messenger cAMP by activation of adenylate cyclase [**Seamon et al 1981**]. As KCl is well characterized for neuronal activity *in vitro* [**Bading et al 1993, Macias et al 2001**], we compared it with concentrations of BDNF (5, 10 and 20 ng/mL) and Forskolin (5, 10 and 20 µM) by immunoblotting serine 10 phosphorylation of histone H3 (H3S10P) (**Appendix B**), a marker for neuronal activity [**Wittmann et al 2009**]. H3S10P levels were higher in Forskolin than BDNF and KCl (**Fig. 3.1a**), but concentrations of BDNF and Forskolin have a similar impact. Thus, we used an intermediate dose of Forskolin (10 µM) and BDNF (10 ng/mL) for further experiments.

We analyzed the gene expression from RNA-sequencing (RNA-seq) at three time points (1, 6, and 10 hours) and at each time the three stimuli (BDNF, KCl, Forskolin) were compared to matched controls (**Fig. 3.1a; Appendix A and B**). Hierarchical clustering of log2 fold-changes of all differentially expressed genes (DE-genes) in all the conditions (FDR=10%) revealed that after 1h similar DE-genes are induced upon BDNF and Forskolin treatment. At later time points (6 and 10h), however, DE-genes are clustered by individual treatment

(**Fig 3.1b; Appendix A**). The total number of early DE-genes varied across treatments with a majority of them induced by KCl, followed by BDNF and Forskolin (**Fig 3.1c**). Treatment differences emerged when DE-genes were annotated by their first appearance ("new DE-genes"). We observed a decreasing number of new DE-genes at each time point for BDNF (3201, 2597, 722) and KCl (5352, 4344, 905), whereas Forskolin showed a peak of new DE-genes at 6h (458, 1570, 548). These results indicate that major transcriptional changes for all treatments occur at 1h and 6h, and BDNF and KCl share common transcriptional dynamics despite the higher similarity between BDNF and Forskolin at 1h (**Fig 3.1b**). As lower levels of H3S10P are observed in BDNF and KCl compared to Forskolin, these results also indicate that H3S10P levels alone do not fully capture the transcription response.

Given the dynamics of gene expression at 1 and 6 h, we sought to assess the biphasic transcription, a key feature of neuronal activity-induced transcription which comprises immediate early genes (IEGs) e.g., transcription factors and delayed response genes (DRGs) involved in synaptic plasticity and neuronal function [**Flavell et al 2008**]. Using unsupervised clustering of the topmost 5000 significant DE-genes across treatments and time points (**Appendix A**), we observed that BDNF and KCl show distinct early and late gene clusters (**Fig. 3.1d**). For example, IEGs expression is divided for BDNF (cluster 1; *Arc*, *Egr2*) and KCl (cluster 3; *Npas4*, *Fosb*). At late time points (6h and 10h), upregulated genes are separated for BDNF (cluster 2) and KCl (cluster 4, 5 and 6). Cluster 2 includes known neuronal function related genes (*Bdnf*, *Cebpb*). Clusters 4, 5 and 6 contain many solute transporters and ion channel related genes (*Slc43a2*, *Cacna1d*, *Slc25a25*, *Kcne4* etc.) [**Tyssowski et al 2018**]. In contrast to upregulation, early downregulated genes are prevalent in KCl (cluster 8), whereas several clusters of late down regulated genes appear in BDNF (cluster 7) and KCl (clusters 9 and 10). Gene ontology (GO) analysis for individual clusters (**Fig. 3.1e**) shows that early induced gene clusters for both up (1, 3) and down (8)

regulated genes are enriched with TFs, regulation of transcription, and DNA binding. Late clusters (2, 4, 5 and 6) showed enrichment for different types of ion channels and transporters for BDNF and KCl. The expression of different transporters may contribute to the electrical diversity in neurotransmission between treatments [**O'Rourke et al 2012**]. For BDNF, late downregulated genes (cluster 7) are also enriched for TF activity and DNA binding terms, whereas for KCl (clusters 9 and 10) are enriched for neurological and cell division related terms.

Altogether, our results reveal cortical neuronal activity differences across stimuli, at the level of gene expression, both in dynamics and in terms of activated and repressed functions.

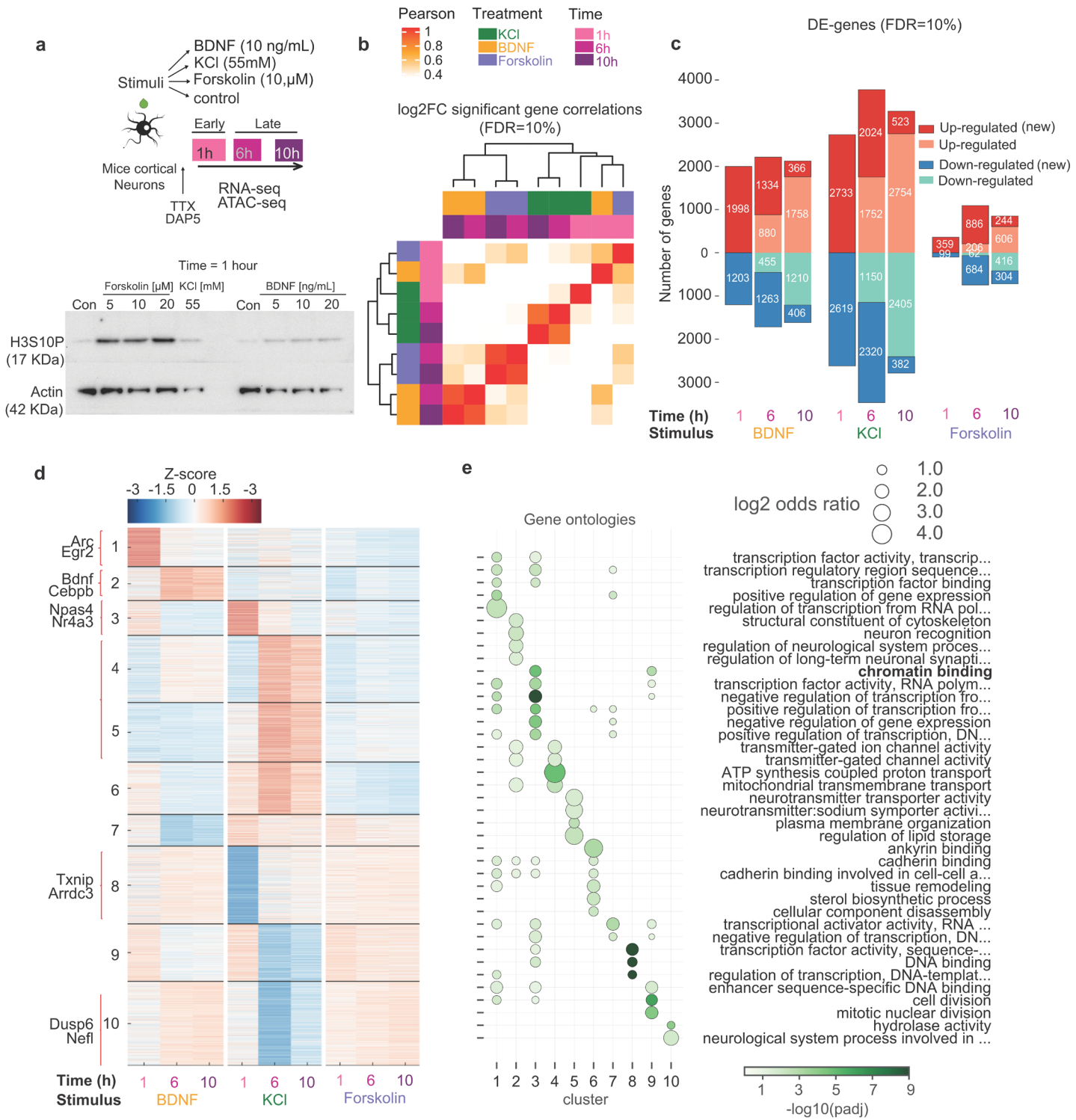

Figure 3.1. (legend on next page)

Figure 3.1 Differential expression dynamics in mouse cortical neurons upon neuronal activity.

(a) *(top)* Experimental setup. Cultured cortical neurons are stimulated with different treatments and prepared at three specific time points for joint RNA-seq and ATAC-seq. *(bottom)* Histone mark phosphorylation H3S10P for Forskolin, KCl and BDNF (Con = no stimulation). Actin is shown as internal control. **(b)** Clustering of correlations calculated from log2 fold changes (versus control samples) of all DE-genes (differentially expressed genes). **(c)** Bar plots indicate the number of DE-genes at each time point and treatment combination (above *X*-axis = up-regulated; below *X*-axis = down-regulated). For time points 6 and 10h the lined box indicates the number of newly acquired DE-genes that did not appear in a previous time point. **(d)** Unsupervised clustering of differential expression changes (FDR=10%, n=5000). 10 clusters resulted from the row mean *Z*-scores calculated from expression values. **(e)** Ontology term enrichments for clusters shown in *(d)*.

3.2.2. Stimulus specific chromatin accessibility upon neuronal activity

Chromatin remodeling tightly controls gene expression [**Gallegos et al 2018**]. Using the Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq), we quantified the chromatin accessibility differences between stimuli at all the time-points (**Fig.3.1a; Appendix A**). Hierarchical clustering of the log2 fold changes of 15566 differentially accessible peaks obtained in all conditions with control comparisons (DA-peaks, FDR=10%) (**Fig. 3.2a; Appendix B**) revealed that, unlike DE-genes where samples did not cluster by treatment at early time, DA-peaks always clustered in a treatment specific manner (**Fig. 3.2a**). When we further classified DA-peaks into gained and closed peaks (log2 fold changes greater and lower than zero, respectively), we observed for BDNF the maximum number of DA-peaks at 1h, followed by KCl and Forskolin (9012, 3701 and 373 respectively) (**Fig. 3.2b**) which is different to the RNA-seq result that shows a maximum of DE-genes at 1h KCl. This result implies that the genome response through accessibility changes is stronger in BDNF than KCl. Similar to DE-genes we annotated DA-peaks by being firstly observed as DA-peaks in specific time points ("new DA-peaks") and found them consistently decreased for BDNF, but not for KCl. The high number of late DA-peaks in KCl is explained by a higher fraction of newly gained DA-peaks at both 6h (1551, 71.5%) and 10h (1025, 47.3%) relative to 1h (2168). In comparison, BDNF response displays a much smaller fraction of newly gained DA-peaks at 6h (446, 6.9%) and 10h (418, 6.5%) versus 1h (6379). For closed DA-peaks, we observed a similar pattern for both BDNF and KCl, with

decreasing numbers from early to late time points, and comparable fractions of new closed DA-peaks (**Fig. 3.2b**). These results show that chromatin changes induced by BDNF are an early event, while KCl exhibits similar levels of chromatin response at late time points. As the low number of DA-peaks obtained for Forskolin limited analysis, we henceforth focused analyses on BDNF and KCl.

To define the genome wide distribution of activity-induced accessible chromatin regions we annotated gained and closed DA-peaks to their genomic features using Homer [**Heinz et al 2010**] (**Appendix A**). The DA-peaks were distributed in the three topmost categories e.g., intergenic regions, introns, and gene promoters (**Fig 3.2c; Fig S3.3a**), but more intergenic DA-peaks observed in BDNF (45%) than KCl (40%), and more promoter DA-peaks for KCl (20%) than BDNF (5%). Widespread intergenic DA-peaks for BDNF suggest that its main chromatin accessibility changes occur at distal regulatory elements (DREs). Increased promoter DA-peaks for KCl imply a rapid gene expression response, which is in agreement with the greater number of DE-genes found in KCl at 1 h (**Fig 3.1c**). We further defined the epigenomic states of the DA-peaks using a chromatin states model from adult mouse neurons generated with Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq) data (**Fig 3.2d; Fig S3.3b**) [**Su et al 2017**]. In KCl, gained and-closed regions often appeared at active TSS and bivalent promoters and were co-marked by active histone marks H3K4me1, H3K27Ac and H3K4me3. Early closed DA-peaks showed a specific enrichment for CTCF (Fold Enrichment = 8.1) (**Fig 3.2d**). In BDNF, gained and closed DA-peaks showed enrichment for active TSS, downstream of TSS and gained enhancers, co-marked by active histone marks H3K27Ac and H3K4me3. Moreover, moderate enhancers and enhancers within a gene are enriched for BDNF closed and gained DA-peaks, respectively (**Fig. 3.2d**). Thus, accessibility changes in DREs are more prevalent in BDNF than KCl. Ontology analysis revealed a strong association to abnormal associative learning

in both BDNF and KCl DA-peaks. Additionally, peripheral nervous system and potassium channel activity terms were enriched more for BDNF gained DA-peaks, while abnormal peripheral nervous system synaptic transmission was enriched for KCl gained DA-peaks at 1h (**Fig. 3.2e**). Altogether, our analyses show wide differences in chromatin response between BDNF and KCl, which connects activity-dependent gene expression to their signature responses.

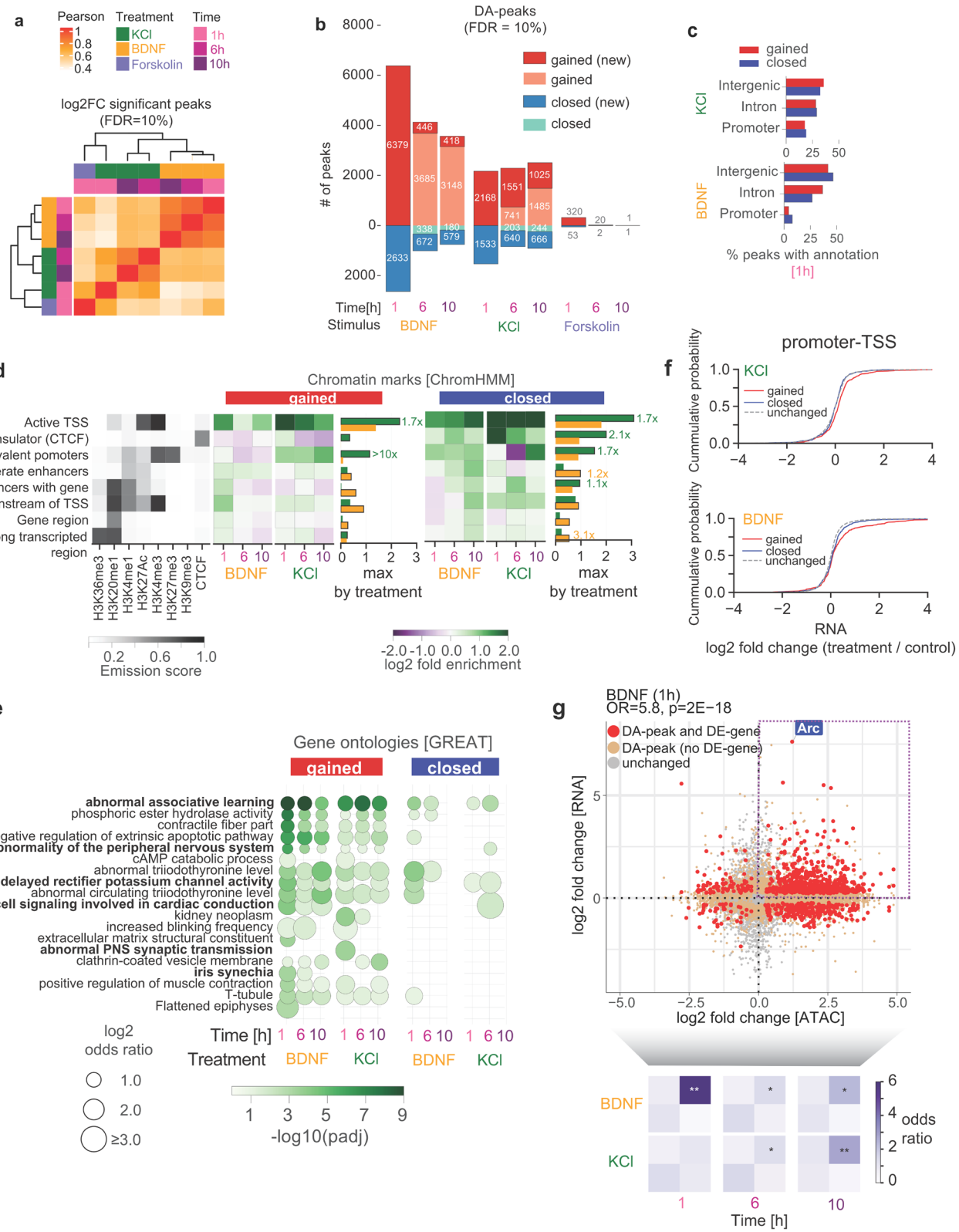

Figure 3.2. (legend on next page)

Figure 3.2 Variation in chromatin accessibility shows early neuronal activity specificity in BDNF- and KCl-treated samples.

(a) Clustering of correlations from log2 fold changes of all DA-peaks (differentially accessible peaks) versus control samples. **(b)** Bar plots indicate the number of DA-peaks at each time point and treatment combination (above *X*-axis = gained; below *X*-axis = closed). Newly gained and closed peaks at 6 and 10 h are highlighted by lined boxes. **(c)** Percentage of DA-peaks with HOMER genomic annotations at 1h **(d)** Log2 fold enrichments for ChromHMM neuron chromatin states. *(left)* Emission scores and state names; *(middle)* log2 fold enrichments of gained DA-peaks. Lined boxes indicate time point with greatest fold enrichment value. Bar plot compares BDNF versus KCl maximum values. *(right*) Log2 fold enrichments of closed DA-peaks. **(e)** Log2 odd ratios for gained and closed DA-peaks related to ontology terms scored using GREAT. **(f***)* Cumulative distributions for log2 fold changes next to TSS related genes. **(g)** (*top*) Association between DREs and gene expression at BDNF 1h. Each point indicates the log2 fold change of an ATAC-seq derived peak and its linked gene expression as log2 fold changes. Colors indicate whether none, only the peak, or both peak and gene (red) show significant changes (*bottom*). Enrichment for DA-peaks and DE-genes in the four quadrants are summarized for BDNF and KCl. Asterisks indicate adjusted *P* value obtained from Fisher's exact test, and corrected by Benjamini Hochberg procedure.

3.2.3. Coordination between expression and accessibility between proximal distal regulatory elements and their target genes

We investigated the dependencies in changes between chromatin accessibility and gene expression upon neuronal activity by determining the global correlation between putative DRE associated peaks and their proximal genes for BDNF and KCl at all time-points (**Appendix A**). The highest coordination was observed between gained DA-peaks and upregulated DE-genes in BDNF at 1h (odds ratio = 5.8; $P < 0.0001$), without any significance in other comparisons and in KCl at 1h (odds ratio = 1.1; $P > 0.05$). At 6 and 10h we observed significant coordination between gained DA-peaks and gained DE-genes for both BDNF and KCl (**Fig. 3.2f**). These results indicate that for BDNF the chromatin DREs affect gene expression starting at 1h, while for KCl the coordination occurs later.

Co-variation of DREs such as enhancers and their target promoters can occur due to their physical proximity and the formation of physical contacts [**de la Torre-Ubieta 2018**]. Using our ATAC-seq data, we calculated the correlations between accessible DREs and accessible promoter pairs located within 50 Kbp and obtained positive correlation distributions indicating peak co-variation (**Fig S3.3c**). These values were further increased when considering only pairs with at least one DA-peak. When we considered only peak

pairs that are annotated as part of Hi-C contacts [**Bonev et al 2017**], a significant increase in the correlation values distributions was observed in links that are part of Neural Progenitors Cells (NPCs) and Cortical Neurons (CNs) versus Embryonic Stem Cells (ESCs) (**Fig S3.3c**). These cell type differences are the highest when either none or one of the peaks is a DA-peak, implying that DA-peak co-variability can be used to describe contacts that are not necessarily captured by Hi-C. Remarkably, IEGs related to neuron function are associated to significantly co-varying ATAC-seq peaks (**Fig S3.3d**).

Collectively, these results reveal a complex landscape of chromatin accessibility changes during neuronal activity, and a coordinated interplay between chromatin accessibility and gene expression across stimuli, associated through co-variation between distal and proximal accessible regions.

3.2.4. Subset of transcription factors underlies stimulus-specific accessibility responses

To determine whether changes in chromatin accessibility after stimuli are related to transcription factors (TF), we searched TF motifs within the DA-peaks using *8*-mers of 108 TF specificity groups [**Mariani et al 2017**] and a database of Position Weight Matrices (PWMs) [**Weirauch et al 2011**] (both are henceforth referred to as "motifs"). For gained and closed DA-peaks in each treatment, we quantified the relative frequencies in comparison with a control set of negative sequences, and ranked TF motifs according to Receiver Operating Characteristic Area Under The Curve (ROC-AUC) values (**Fig 3.3a; Appendix A**). The basic region leucine zipper (bZIP) domain was the most enriched motif in DA-gained peaks for both BDNF and KCl (mean ROC-AUC = 0.65; $P < 0.0001$; Wilcoxon rank sum test) consistent with activity-dependent changes in bZIP expression playing a role in synaptic plasticity, learning and memory [**Kandel et al 2012**]. The positional

centrality of bZIP motif in DA-gained peaks suggests a role of these TFs as pioneers [**Su et al 2017]** (**Fig S3.4a**). *8*-mers related to Homeobox (Hbox-III), and POU domain (POU; POU-HMG) were also enriched in DA-gained peaks for both treatments (ROC-AUC > 0.55). Two distinctive Homeobox subgroups, Hbox and Hbox-II, were only enriched in BDNF, suggesting a role for a subset of Homeodomain TFs in this chromatin response. Furthermore, gained DA-peaks in BDNF exhibited ETS, TALE-zfC2H2 and EGR motifs while in KCl we observed E2F-zfC2H2 and KLF motifs. The Early Growth Response (EGR) motifs, a class of IEGs is related to regulation downstream target genes involved in neurobiological processes such as synaptic plasticity and memory formation [**Beckmann et al 1997, Gallitano-Mendel et al 2007**]. Closed DA-peaks in BDNF contained HIC1 and RFX motifs, while in KCl included CCCTC-binding factor (CTCF), E2F, KLF, and zfCXXC-SAND motifs (**Fig 3.3b**).

As TFs of the same family can have similar DNA target sequences due to the shared recognition specificities [**Mariani et al 2017**], we examined whether individual TF expression levels can further define the observed motif enrichments. The bZIP group encompasses ten members and among them *Fos*, *Fosb*, *Fosl2* and *Atf3* showed significant up-regulated expression in both BDNF and KCl across time points (**Fig. 3.3c**). For BDNF increased expression of most bZIP members except *Fos* and *Xbp1* disappeared at 6h, but for KCl expression of many bZIP members were maintained up to 10h. The EGR module contains eight members and four (*Egr-1/2/3/4*) of them are strongly induced by both BDNF and KCl. Importantly, up-regulated *Egr-1/2* levels were only observed in BDNF but not in KCl which showed late reduction. On the other hand, HIC1 motif enrichment in closed DA-peaks is consistent with higher expression levels of *Hic1* in BDNF compared to KCl. As HIC1 has been described to act as a repressor [**Pinte et al 2004; Ubaid et al 2018, Boulay et al 2012**], this result suggests a link between HIC1 upregulation and stimulus-induced

chromatin accessibility decrease in BDNF. Unlike *Hic1*, *CTCF* did not show significant expression changes, yet its motif is highly enriched in 1h KCl closed DA-peaks, suggesting a specific layer of regulatory control linked to CTCF. Overall, we revealed coordinated binding and expression of TFs explaining the stimulus-specific changes in chromatin accessibility (mean fraction of DA-peaks explained by enriched TF motif = 68.8%) (**Fig S3.4b**).

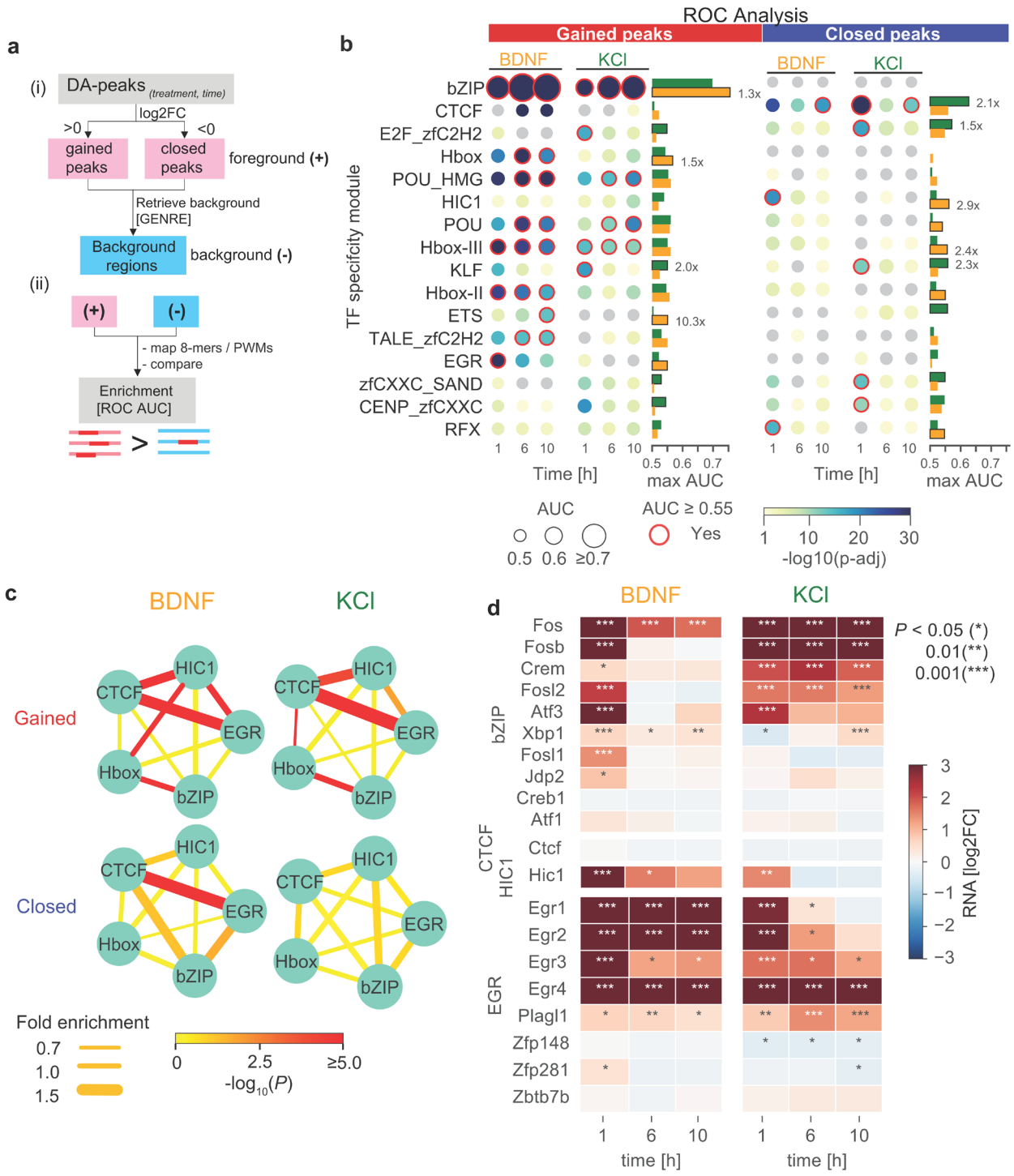

Figure 3.3 Transcription factors linked to gained and closed DA-peaks reveal stimulus specific regulation.

(a) Scheme indicating how gained and closed DA-peaks for a treatment-time combination are prepared for motif enrichment analyses (PWMs and 8-mers). *(i)* For each set of peaks, a background set of sequences matching genomic features is generated with GENRE. *(ii)* Each combination of foreground and background sequences is scanned in the motif databases, and ROC AUC values are generated for each combination. **(b)** Enrichment of main regulatory TF-modules enriched in gained and closed DA-peaks. Circle size indicates ROC AUC value and color indicates significance (Wilcoxon adjusted *P-value*). Bar plot compares highest value obtained between BDNF (green) versus KCl (orange). ROC AUC values lower 0.5 are depicted in gray. **(c)** Co-enrichment of motifs in DA-peaks for 1h in KCl and BDNF. Network edges indicate fold enrichment (edge thickness) and Benjamini–Hochberg adjusted *P*-value (color). **(d)** Expression values for genes related to bZIP, CTCF, HIC1 and EGR *k*-mer groups. Significant log2 fold changes versus control are displayed with asterisks (using DESeq2 [**Anders et al 2010**]).

3.2.5. TF combinations define stimulus-specific gene expression

Interaction between TFs can lead to more stable binding and function of the regulatory element [**Vandel et al 2019, Jolma et al 2015; Junion et al 2012**]. We therefore examined pairs of TF motifs in 1h BDNF and KCl carrying most DA-peaks to assess their collaborative role in chromatin accessibility changes. By significance, the strongest combination of three motifs DA-peaks for both treatments was for between EGR, HIC1 and CTCF (FE = 2.3 and 2.8 for BDNF and KCl respectively, adjusted $P < 0.0001$). Additionally, the combination of these factors with bZIP is also enriched, suggesting a coordination between bZIP pioneering activity and the interactions observed for these three factors (FE = 2.2 and 1.7 for BDNF and KCl respectively; adjusted $P < 0.01$). Co-occurrence of TF motifs can define cell-type specific enhancers and be used to understand their response [**van Bömmel et al 2018**]. Indeed, co-occurrence of EGR and bZIP motifs significantly increases chromatin accessibility compared to either bZIP or EGR alone in BDNF gained DA-peaks (**Fig. S3.5c**), suggesting an interaction between pioneer factor and co-regulator. Furthermore, co-occurrence of bZIP and EGR motifs in gained DA-peaks associated with Transcription Start Sites (TSSs) (<5kb) lead to a significant upregulation of genes for 1h BDNF, but not for KCl, compared with either bZIP or EGR alone (**Fig. 3.4b**). Thus specific increase in chromatin accessibility consisting of the two TFs (e.g., bZIP and EGR) upon BDNF treatment contributes to gene expression synergy.

Among the genes showing high correlation between accessibility in putative DREs and expression (**Fig 3.2g**) we found the Activity Regulated Cytoskeleton associated protein (Arc), which is a well-known IEG pivotal for learning and memory formation [**Tzingounis et al 2006, Plath et al 2006**]. Arc is induced by both BDNF and KCl but higher at 1h BDNF (**Fig. 3.4c**) consistent with a coinciding increased accessibility at both the promoter and putative DRE region. We considered this region an enhancer for Arc because of its

enhancer-related histone modifications present in neuronal epigenomes data (H3K27ac and H3K4me1) [**Malek et al 2014**]. Additionally, CTCF tracks [**Sams et al 2016**; **Ren et al 2017**], and a Hi-C contact between Arc gene and this DRE region are observed [**Bonev et al 2017**].

The Arc gene enhancer region contains the DA-peak specific to BDNF which carries four important TF motifs (bZIP, EGR, HIC1 and Hbox-II) (**Fig 3.4c**). As co-occurrence of bZIP and EGR motifs in DRE showed higher expression of the linked gene, we hypothesized that the accessibility increases at DRE could explain a higher Arc expression in BDNF compared to KCl. Also, HIC1 is present in the DRE. Like other IEGs, expression of Arc goes down after 1h, therefore, HIC1 might contribute as a repressor in this downregulation by binding to the Arc enhancer region in a BDNF-specific manner.

To validate the role of the Arc gene DRE in BDNF-mediated gene expression, we tested variants that remove sections of the Arc enhancer containing EGR and HIC1 motifs by means of CRISPR-Cas9. Indeed, significant reduction of BDNF-mediated Arc gene expression occurred in DRE deleted clones, but not in control (**Fig. 3.4d**) indicating that this DRE is involved in BDNF-specific Arc up-regulation. Altogether, our results show a complex interaction between TF combinations and DREs to control the expression of neuronal activity related genes in a stimulus-specific way.

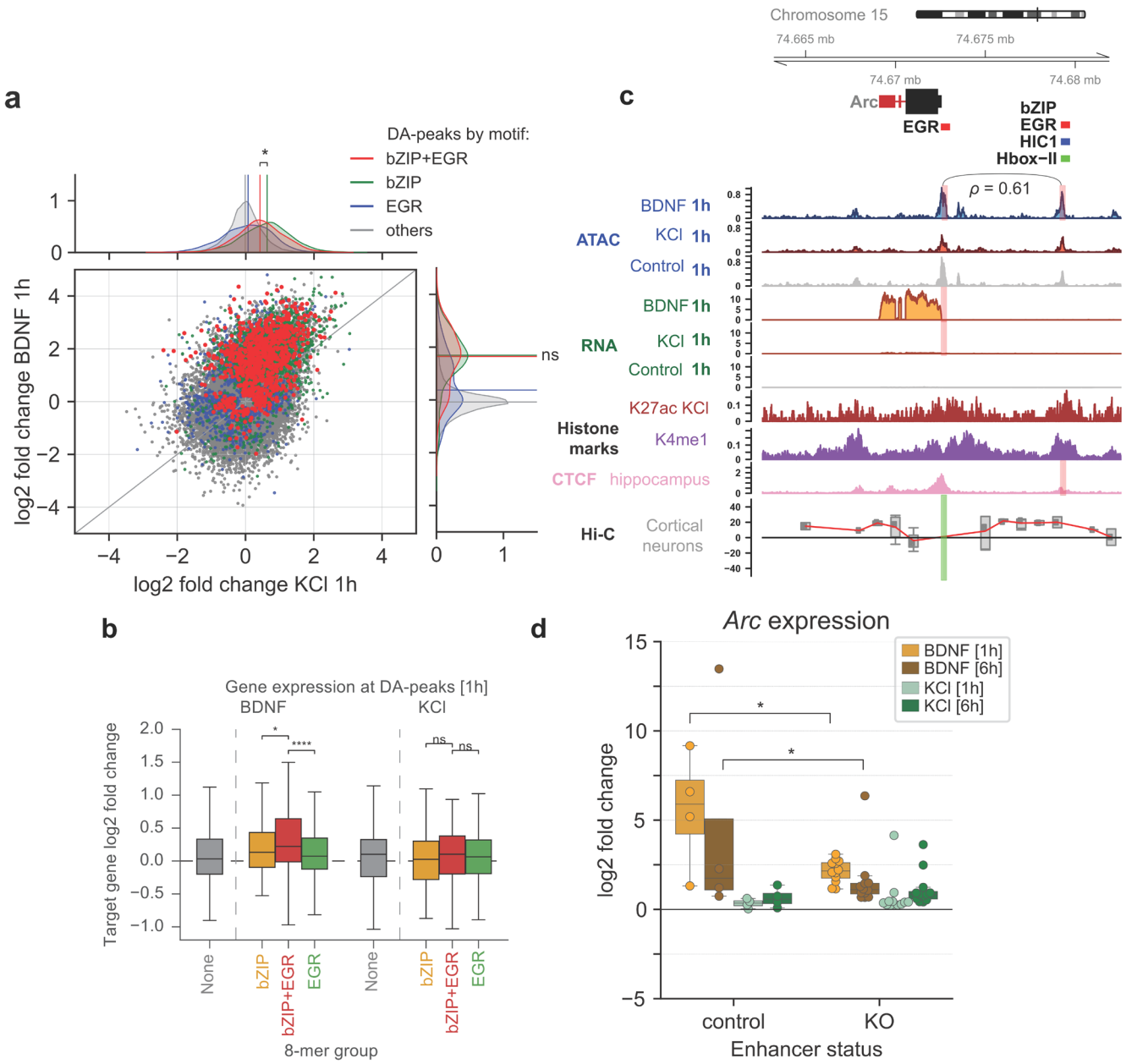

Figure 3.4 Variability in BDNF gene expression is linked to bZIP+EGR combinations acting in promoters and enhancers.

(a) Accessibility changes for BDNF and KCl peaks grouped by presence of *k*-mers for bZIP (green), EGR (blue), or their combination (bZIP+EGR, red). Lines in density plots indicate median value for distributions. **(b)** Expression of genes with TSS proximal to DA-peaks, subset by *k*-mers for bZIP (orange), EGR (green), or both (red). Asterisks indicate adjusted *p*-values derived from Wilcoxon test. **(c)** Chromatin tracks of *Arc* gene. Blue = 1h ATAC (counts per million); Green = 1h RNA-seq (counts per million); Brown = H3K27ac; Purple = H3K4me1 [**Malek et al 2014**]; Pink = CTCF [**Sams et al 2016; Ren et al 2017**]; Gray = Cortical neurons Hi-C data [**Bonev et al 2017**]. Red bars in ATAC-seq tracks indicate gained DA-peaks in BDNF, and red bars in RNA-seq tracks indicate differential expression in BDNF and KCl. The green block in Hi-C tracks represents the anchor point for calculation of scores, using Shaman. The curved line between ATAC-seq peaks indicates the Spearman correlation of the normalized counts. Motif names indicate presence of *k*-mers for enriched specificity groups in those peaks. **(d)** Arc expression in BDNF is reduced upon deletion of selected regions in enhancer (* = $P < 0.05$ using two-sided *t*-test).

3.2.6. CTCF at promoter-exon loops is linked to differential exon usage in neuronal genes.

CTCF enrichment in KCl closed DA-peaks suggests a regulatory role of this TF in neuronal-activity. CTCF controls genome organization by forming TADs [**Hansen et al 2018**], and intra-TAD contacts can direct enhancers to its target promoters by CTCF looping [**Heintzmann et al 2007**]. CTCF has also been associated through genetic variation [**Li et al 2016**] and promoter-exon contacts [**Ruiz-Velasco et al 2016**] to splicing. As CTCF motifs are enriched in closed DA-peaks in 1h KCl (**Fig. 3.3b**), we sought to study further associations between CTCF and gene regulation (**Appendix A**). Interestingly, we found and enrichment of 1h KCl DA-closed peaks for CTCF promoter-exon loops (odds ratio= 3.5; adjusted $P < 0.001$; **Fig. 3.5a**), an enrichment of convergent CTCF motifs for those peaks, and a stronger enrichment in KCl than in BDNF for DA-peaks with CTCF binding sites in both intronic and exonic regions **(Fig S3.6a-b)**. This result suggests that CTCF might regulate transcriptional events after transcription initiation in a treatment specific way [**Stadhouders et al 2012**] by hindering Pol II elongation and alternative mRNA splicing [**Paredes et al 2013; Shukla et al 2012; Ruiz-Velasco et al 2017**]. Thus, we assessed the levels of differential usage of exons (DUEs) between BDNF and KCl. Globally we found 7188 exons differentially used within their genes between BDNF versus KCl (FDR = 10%). When filtering for DUEs with at least ≥2 fold change; the expression levels of 307 exons were repressed and 1246 exons showed increase in expression between BDNF and KCl (**Fig 3.5b**). To validate this, we selected three genes with important roles in neuronal function and activity, containing a significant differentially used exon and a CTCF loop between the DUE and promoter: Trio Rho Guanine Nucleotide Exchange Factor (*Trio*) [**Fujita et al 1998**], Syntaxin Binding Protein 5 (*Stxbp-5*) and Carboxypeptidase E (*Cpe-201*) [**Woronowicz et al 2010**]. We used RT-qPCR assay to quantify relative exon usage - one exon was

differentially used based on our analysis and the other exon was from the same gene but remained unchanged between treatments, and was used for within-gene correction. We normalized the expression changes with *Rpl13* as a reference gene and compared the exon ratio (fold change ratio between positive and control exon) for BDNF, KCl and untreated neurons.

Haploinsufficiency in *Trio* causes severe deficits in behavior and neuronal structure and function [**Goebbels et al 2006; Katrancha et al 2019**]. Expression of *Trio* DUE exon 29 located in the last exon of a transcript variant that carries an additional 3'UTR sequence showed a significantly higher exon ratio relative to the regular exon 29 (without 3'UTR) in BDNF versus KCl at 1 h (mean exon ratio = 1.4 and 0.9 for BDNF and KCl respectively; adjusted $P < 0.001$, two-sided *t*-test). *Stxbp5* functions to regulate synaptic capturing and recycling of secretory vesicles with the presynaptic plasma membrane [**Geerts et al 2017**]. Murine *Stxbp5* has at least 15 transcript variants and we revealed that DUE exon 1 compared to exon 5 showed a higher expression in BDNF than KCl at 1 h (mean ratio = 1.15 versus 1.1; adjusted $P < 0.01$, two-sided *t*-test). *Cpe-201* acts as neurotrophic factor to promote neuronal survival [**Cheng et al 2014**] and can also function as a sorting receptor that can bind to BDNF [**Lou et al 2005**]. DUE number 9 of *Cpe-201* versus exon 6 showed a higher exon ratio increase at 1h for neurons treated with BDNF in comparison to KCl (mean ratio = 1.13 versus 1.03; adjusted $P < 0.01$, two-sided *t*-test). Taken together, our results demonstrate a stimulus-specific regulatory layer associated to alternative transcription in neuronal activity, likely mediated by CTCF.

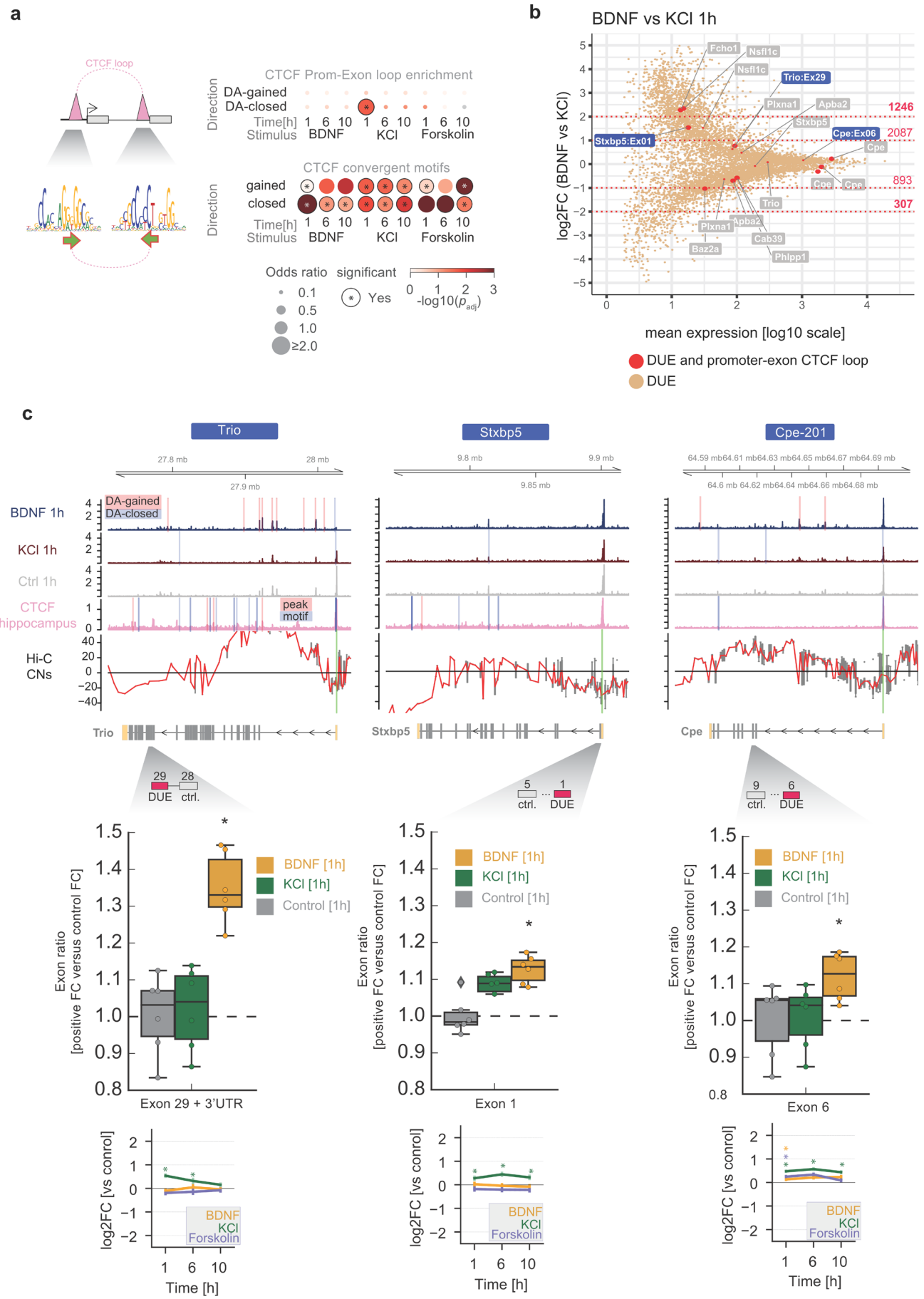

Figure 3.5. (legend on next page)

Figure 3.5 Association between CTCF in DA-peaks and differential exon usage.

(**a**) (*left*) Depiction of promoter-exon CTCF loops. CTCF peaks (pink) pairs contain one peak close to the gene promoter region, and another peak close to exons. These peaks contain CTCF motifs that can be in convergent orientation (green arrows). (*top heatmap*) Odd ratios for enrichment of promoter-exon CTCF loops in gained and closed DA-peaks. Promoter-exon loops are defined in [**Ruiz-Velasco et al 2017**]. (*bottom heatmap*) enrichment of convergent CTCF motifs in promoter-exon regions overlapping gained or closed ATAC-seq peaks **(b)** Exon log2 fold changes between BDNF and KCl in 1h as quantified by DEXSeq [**Anders et al 2012**]. Orange dots indicate Differentially used exons (DUEs), and red dots indicate DUEs with promoter-exon CTCF loops in their genes. Genes highlighted in blue are selection for validation. **(c)** *(top)* Genome tracks for *Trio*, *Stxbp5* and *Cpe* genes. (*middle*) Depiction of reference DUE exon (ref) and control exon (gray) used for comparison using RT-qPCR. *(bottom)* Fold change ratios between reference and control exons at 1h after treatment with BDNF (orange), KCl (green), and control (gray). Asterisk indicates significant change versus control (two sided *t*-test). Bottom plot shows log2 fold changes for gene expression values versus control (*= $P < 0.1$ in treatment versus control comparison of normalized counts using DESeq2 (**Appendix A**)).

3.3. Discussion

In this work we have performed a comprehensive temporal analysis of gene expression and chromatin accessibility changes to compare neuronal activity dynamics across multiple stimuli. The integration and dissection of involved regulatory elements allowed us to predict and validate principles that determine specificity among these stimuli and are especially relevant for neuron function. We identified functional expression profiles explaining early and late waves of gene expression separating KCl, BDNF and Forskolin responses. Despite the global agreement between early expression profiles, our results suggest a differential outcome in expression programs, likely mediated by specific TFs modulating the common early response into targeted functional outcomes. Further work would be required to identify TFs or additional factors affecting the expression of these induced early genes.

The integration of our time-course chromatin accessibility data with other epigenomes and HiC-contacts data allowed us to pinpoint features that distinguish BDNF and KCl responses. Importantly, we found a strong coordination between distal regulatory elements and target genes as an early event in BDNF-induced neuronal activity, whereas for KCl the early expression response was mainly defined at the level of promoter regions, indicating that early neuronal expression events are associated to active TSS elements.

Additionally, the independent clustering of accessibility and expression data (**Fig 3.1a, 3.2a**) and the observation of late coordination for accessibility and expression for both treatments (**Fig 3.2f**) suggest major rearrangements in the chromatin landscape affecting enhancers at late time points.

Through an extensive analysis of TF motif signatures at differentially changing accessible regions we predicted TFs involved in regulating the chromatin landscape, and combinatorically controlling gene expression. The AP-1 complex acts a classical pioneer factor (bZIP module in our work) [**Biddie SC et al 2011**] and explains a major part of the gained chromatin regions across tested stimuli. Importantly, we identified multiple secondary TFs enriched along bZIP motifs such as Hbox/EGR/HIC1 and CTCF, and in the case of HIC1 and EGR found them to be specifically associated to the BDNF response through their expression levels, which is also in agreement with the prediction of bZIP-related TFs such as cFos recruiting co-regulators that ultimately determine response specificity [**Su et al 2017**]. EGR motifs predominantly act as co-regulators and specifically increase gene expression of target genes upon BDNF treatment. HIC1 motifs, on the other hand, are usually associated with co-repressing specific subsets of binding regions opened by BDNF, and overall closing of early accessible regions. The co-enrichment of HIC1 and EGR motifs hinted at a further regulatory role through interaction of these factors, where HIC1 presumably acted as an early repressor of regions activated by bZIP and EGR. We mapped a distal regulatory element active in BDNF and identified the differentially increasing Arc gene expression to be controlled by a combination of these factors (bZIP-EGR-HIC1). The reduced expression of Arc, concomitant to disruption of enhancer EGR and HIC1 motifs in the associated distal regulatory element highlights a precise activation of gene expression which is likely mediated by interactions between bZIP and EGR/HIC1. The role of these factors thus constitutes a functional triad that modulates the response

specificity in BDNF, and its additional functional roles and interactions with the repressor TF HIC1 remain to be understood.

The enrichment of CTCF motifs on early closed KCl peaks and the recently reported CTCF role in alternative splicing to KCl response in closed peaks [**Ruiz-Velasco et al 2017**], suggests an association between CTCF and differential exon usage. We validated our hypothesis using three genes shown to be affected and involved in neuron function. These results highlight a formerly unexplored response mechanism in classical neuronal activity. Behavioral processes, including learning, have been related to genes and variants affecting the Trio gene in regions close to our reported DUE [**Pengelly et al 2016**]. Chromatin accessibility and promoter-exon loop contacts could therefore potentially mediate responses through treatment-specific exon usage. Finally, as Trio contains several Single Nucleotide Polymorphism associated to learning in targeted domains of this protein close to our studied exon [**Pengelly et al 2016; Sadybekov et al 2017]**, the alternative splicing of these genes could potentially mediate functional outcomes. Further work requires finding how these CTCF associated chromatin changes are triggered and understanding their specific functional consequences.

Chapter 4

Prediction and validation of terminal repressors of non-lineage gene expression programs for cell reprogramming

In this chapter, I describe the analyses and results of a project aiming at systematically predicting transcription factor combinations for reprogramming using the concept of Terminal Repressors. This project was conceived as a collaboration between the groups of Moritz Mall (DKFZ) and Judith Zaugg (EMBL Heidelberg). All computational analyses were carried out by me. Experimental validations have been carried out by Dr. Juan Segarra, a postdoc in the group of Moritz Mall. The analyses in this work are described in the following manuscript:

Ignacio L. Ibarra*, **Segarra Juan***, Judith Zaugg and Moritz Mall. Systematic classification of terminal repressors in multiple cell lineages. *In preparation.*

4.1. Introduction

The understanding of how cell identities are defined is of relevance in developmental biology and regenerative medicine. Starting with the Yamanaka factors [**Takahashi et al 2006**], progress in this field currently allows reprogramming cell types from fibroblasts to pluripotent cells, as well as other cell types, including macrophages, cardiomyocytes, neurons and dendritic cells [**Feng et al 2008, Ieda et al 2006, Vierbuchen et al 2010, Rosa et al 2018**].

The main determinants of cell types are transcription factors (TFs) [**The Tabula Muris Consortium et al 2018**]. Historically, TFs used in reprogramming experiments have been selected based on enrichment of putative TF binding sites along promoters of cell fate genes. Their overexpression or removal was concurrently associated to changes in expression of determinant genes related to a specific cell fate [**Vierbuchen et al 2010**]. Upregulation of marker genes is used as a predictor of cell type commitment and thus serves as a proxy for reprogramming efficiency. Nowadays, multiple caveats affecting reprogramming efficiency need to be taken into account as well, such as reversing epigenetic memory in pre-marked genes [**Hörmanseder et al 2017**] and maintaining low expression levels of non-cell fate genes [**Battaglioli et al 2002**].

The influence of a TF on its target genes is associated with its overall role as an activator or repressor [**Wang et al 2013**]. To date, classifying TFs as global activators or repressors is limited for many TFs given the lack of conclusive data [**Han et al 2018**]. Additionally, cofactors can be determinant in the final response, implying that TFs can possess a dual role according to their molecular context and partnering with co-factors [**Remedy et al 2004**].

In recent years, the reprogramming of cell types has been studied by monitoring the consequences of the addition of TFs on gene expression and cell physiology. In Mouse

Embryonic Fibroblasts (MEFs), addition of the myelin transcription factor 1-like (Myt1l) improves reprogramming efficiency and cell commitment to induced Neurons (iNs) [**Vierbuchen et al 2010**]. An integration of ChIP-seq and gene expression profiling showed that this particular TF represses genes of negative neurogenesis pathways such as Wnt and Notch, as well as genes involved in mouse embryonic fibroblasts maintenance. Additionally, Myt1l motifs are over-represented in non-neuronal genes such as the ones in keratinocytes, the pancreas and hepatocytes, suggesting a specific repression of all these programs [**Mall et al 2017**]. From these observations and the almost exclusive up-regulation of Myt1l in neuronal cell types [**The GTEx Consortium 2013; Tabula Muris Consortium et al 2018**], Myt1l has been proposed as a 'terminal repressor' required to maintain the neuronal state by repression of other cell type-programs.

The role of Myt1l in neurons allows for speculation on the existence of additional TFs in other cell types that could be acting as terminal repressors during and after differentiation or reprogramming. Based on the idea that genomic features such as the ones observed for Myt1l can be recovered and systematically interrogated for other TFs in other cell types as well, terminal repressors can be predicted and benchmarked accordingly. In this Chapter, we describe an approach to obtain terminal repression signatures such as the ones obtained for Myt1l in neurons. Curation and selection of specific cases put forth novel terminal repressors in multiple cell types, such as Prospero homeobox protein 1 (Prox1) in liver reprogramming and T-box transcription factor (Tbx5) in cardiac cell reprogramming. Bioinformatics analyses indicate that these factors are related to the repression of non-cell fate genes. These two factors, similarly to Myt1l, specifically increased the expression of reprogramming markers and showed positive immunofluorescence patterns in reprogramming experiments.

4.2. Results

4.2.1. Classification of terminal repressors across cell types.

One of the signature features for Myt1l in neurons is that its TF binding model -summarized as a Position Weight Matrix (PWM)- reveals an enrichment for motifs in promoters related to non-neuronal genes versus neuronal genes (**Fig S4.1a**) [adapted from **Mall et al 2017**]. This result is opposite to transcriptional repressors such as REST that represses neuronal genes in non-neuronal cell types. We therefore hypothesized that comparing the enrichment of motifs in non-cell fate versus cell-fate genes across cell types could serve as an adequate feature to detect such terminal repressors. To explore this concept, we annotated expression data from the Genotype-Tissue Expression Project [**The GTEx Consortium 2013**] and the Tabula Muris [**Tabula Muris Consortium et al 2018 Tabula Muris Consortium et al 2018**] for several human tissues and cell types to obtain a set of signature genes able to summarize each, and compare their promoters based on TF motifs (**Appendix A**). Mapping motifs for each TF using reference PWM models [**Lambert et al 2018**], allowed us to calculate enrichments and depletion estimates for each tissue by means of log2 fold changes between events in non-cell fate genes versus cell-fate genes. After normalizing these values to Z-scores, we obtained an estimate of how enriched a given motif is in promoters of non-cell fate genes (**Fig 4.1a**). As the majority of non-cell type genes are not expressed in the cell type of interest, we suggest that motif biases summarized by such Z-scores allow correlating gene repression in these genes to the TFs of interest.

Additionally, another property of a putative terminal repressor is its exclusive up-regulation in the cell type of interest as this suggests cell-type specific functional relevance. This behavior is different from TFs that are required during differentiation, as

their expression levels increase and decrease concomitantly with cell differentiation [**Singh et al 2016**]. The ubiquitous expression of Myt1l in all neuronal cell types, on the other hand, is in agreement with its relevance to maintain cell state regardless of the differentiation status. To explore this idea in other tissues as well, we used expression counts reported as Transcripts Per Million (TPMs) and normalized those as Z-scores to explore expression biases for each TF in every single tissue versus all other tissues (**Fig 4.1a**). Together, both Z-scores derived from expression and motif bias provide a bidimensional score of activation/repression potential that is successful at classifying Myt1l as a repressor (**Fig 4.1.b**). This analysis is also able to recover Ascl1, a TF related to activation of neuron related genes [**Raposo et al 2015**], and relevant for neuronal reprogramming [**Vierbuchen et al 2010**]

From this result, we sought to integrate these scores from across tissues and cell types in an unbiased way, to discover potential new terminal repressors. Both metrics for expression and motif bias are added into a score (S_g) that describe the classification of TFs into activators of fate genes and repressors of non-fate genes in each tissue (**Appendix A**). When applied to GTEx data, this metric validates main activator TFs in multiple tissues, such as Hnfa1 in hepatocyte (**Fig 4.2a**), Myod1 in myocytes [**Tapscott et al 2005**] (**Fig S4.2b**), and Nkx2-5 in cardiac muscle cells (**Fig 4.2b**). Additionally, putative TF activators such as Meox1 in adipocytes and Arnt2 in oligodendrocytes are suggestive of new roles of these factors as main activators of fate-genes (**Fig S4.2c-d**). Globally, we observe an enrichment of TFs used as reprogramming factors in the top S_g quantiles [odds ratio > 10; mean Fisher's exact test adjusted $P < 0.0001$] (**Fig S4.1**). Altogether, our results indicate that a simple annotation based on the two Z-score metrics is able to recover differentiation related factors of reprogramming relevance. Given data for activators and

Myt1l in neurons, we can therefore attempt to classify TFs based on their global role as activators or repressors.

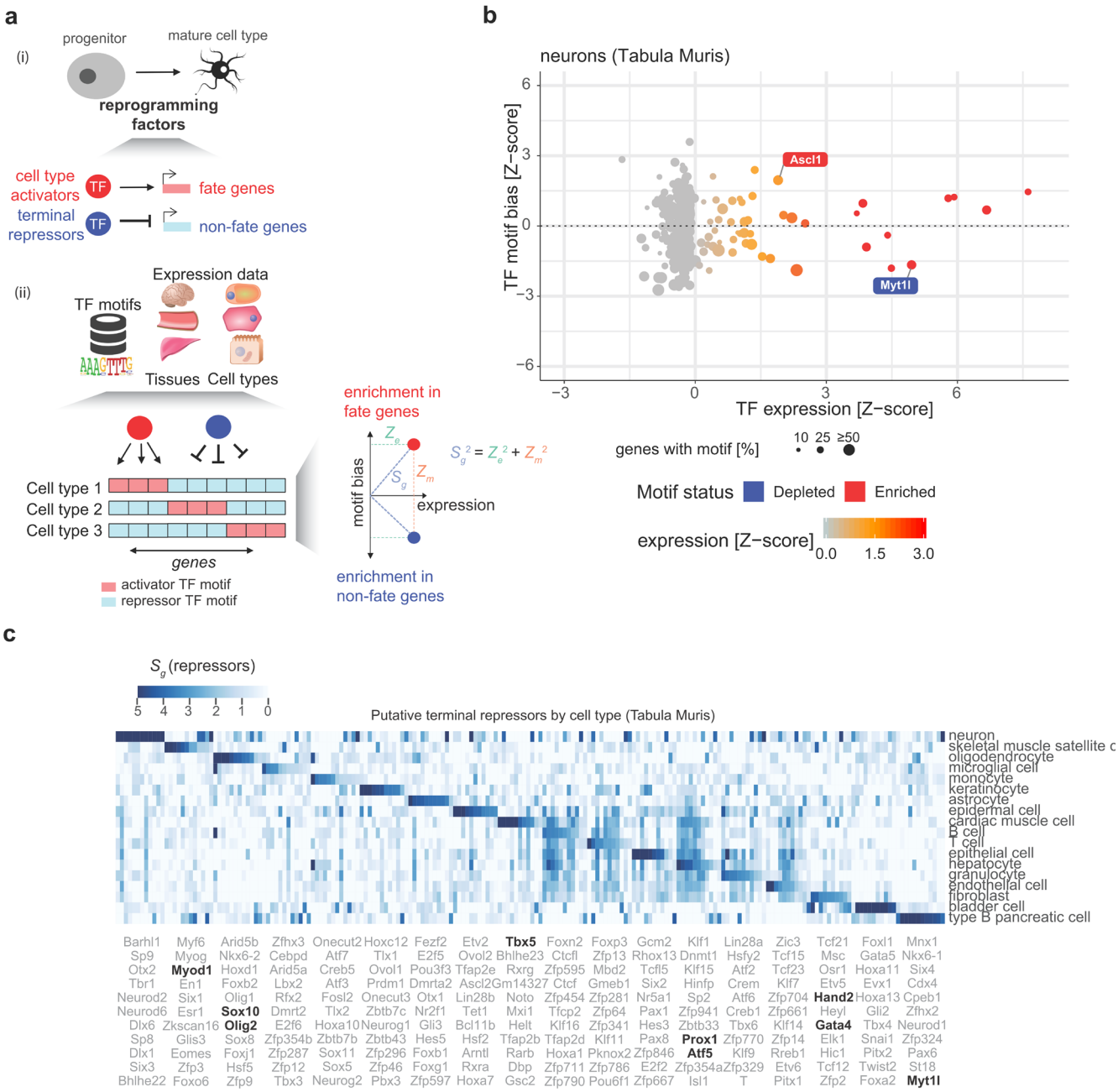

Figure 4.1 Classification of terminal repressors through integration of expression and motif data.

(a) (i) Scheme describing terminal repressors (blue) role in downregulating non-fate genes, whereas activators (red) are related to up-regulation of target genes. Specific TFs in cell types are predicted to repress a set of genes unrelated to the cell fate, thereby increasing cell fate maintenance. (ii) Through integration of TF motifs and expression data from public cohorts, TFs are classified according to their enrichment in fate versus non-fate genes and their expression values into activators and putative terminal repressors. **(b)** Motif enrichment in target genes versus expression for neuronal genes reveals Myt1l in Tabula Muris (database) as a terminal repressor (blue label). Additionally, related reprogramming factors such as Ascl1 (red) can also be observed **(c)** Heatmap of selected putative terminal repressors ($Z_m < 0$) selected using Tabula Muris, based on S_g score (**Appendix A**). Among highlighted TFs, the ones used as reprogramming TFs are highlighted in bold.

4.2.2. Annotation of new terminal repressors in liver and cardiac muscle cells.

Annotation of TFs based on our score revealed a shortlist of three candidates per tissue, on average, showing a putative terminal repressor-like signature (**Fig S4.1b**). Interestingly, for certain tissues such as muscle or esophagus, we are not able to select under a common threshold putative repressor TFs, suggesting that the concept of terminal repressors concept might not apply to all cell types. To narrow down the range of factors for validation, we focused only on those that have been suggested previously to be relevant for the cell reprogramming in those specific cell types (**Table S4-I**) [**Tabula Muris Consortium et al 2018**], or to have indications of a repressor potential [**Han et al 2018**]. This gave us a shortlist of 21 factors out of which we selected two for validation.

A first case for validation was the Prospero homeobox protein 1 (Prox1) in the context of liver differentiation (**Fig 4.2a**). This factor has been suggested to be relevant for differentiation of this tissue, and has been linked to liver cancer due to the impact of tumor progression on its expression levels [**Dudas et al 2008**]. We therefore hypothesized that this factor plays a relevant role in cell fate maintenance for hepatocytes through repression of non-hepatocyte genes. Another candidate predicted to act as a terminal repressor is the T-box transcription factor (Tbx5) in cardiac cells (**Fig 4.2b**). This factor has been reported to be important for cell fate commitment in mouse cardiac cells, and its functional repression has been related to dual effects in terms of gene expression [**Waldron et al 2016**], where joint analysis of ChIP-seq and RNA-seq expression upon knock-out of Tbx5 indicates an enrichment of occupied Tbx5 peaks in cardiac repressed genes. Further, its role in cardiac differentiation has been shown to be conditioned through interactions with the NuRD repressor complex.

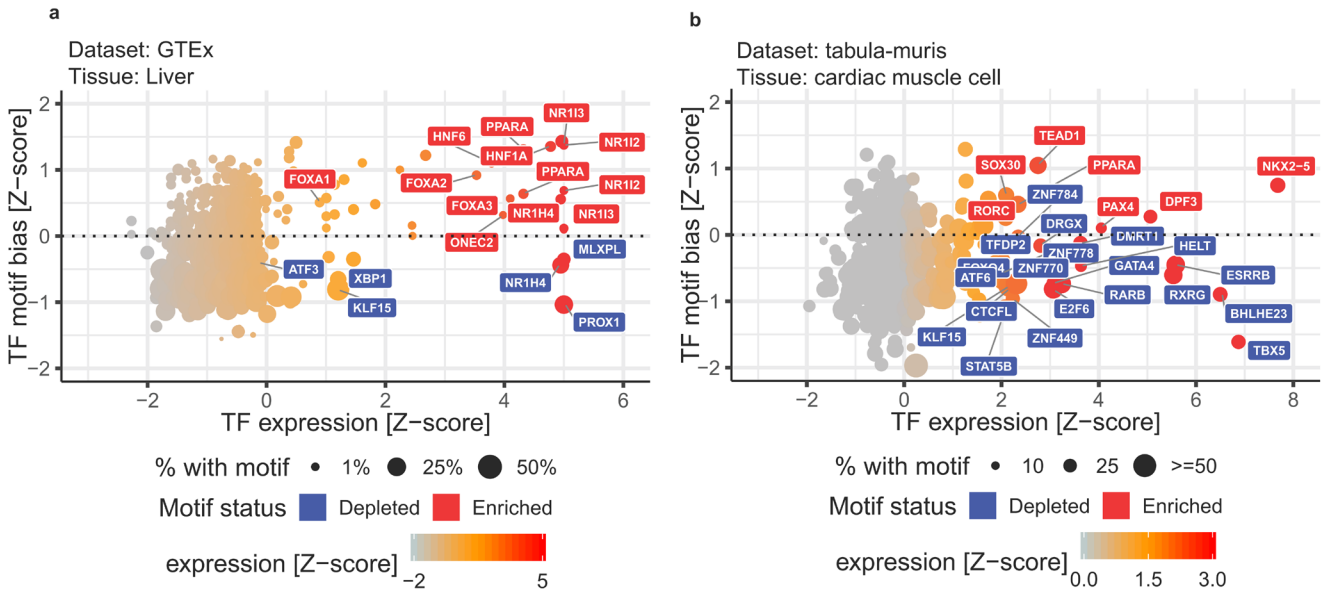

Figure 4.2 Putative terminal repressors in liver and cardiac muscle cells.

(a) TF motif and expression biases suggest Prox1 as a terminal repressor of non-hepatocytes genes. **(b)** TF motif versus expression bias based on scRNA-seq data from Tabula Muris suggests Tbx5 as a terminal repressor of non-cardiac genes in cardiomyocytes.

4.2.3. Validation experiments for Prox1 and Tbx5 indicate specific reprogramming potential

To validate the reprogramming potential of Prox1 and Tbx5 we set up a reprogramming scheme in which mouse embryonic fibroblasts (MEFs) cells were reprogrammed into neurons, hepatocytes and myocytes with the addition of differentiation-specific reprogramming TFs. Independent addition of Myt1l and each putative terminal repressor then provide a consistent way to assess differentiation improvements in a given reprogramming protocol, such as by assessing the extent of the activation or repression of marker genes.

In agreement with previous data, the reprogramming of MEFs into neurons upon addition of Myt1l in a differentiation protocol with Ascl1 was increased, as deduced from the elevated expression levels of the neuronal marker gene Neuron-specific Class III β-tubulin (TuJ1) [**von Bohlen and Halbach 2007**], and decreased levels of muscle marker

Desmin (**Fig 4.3b**). This effect also occurs in hepatocyte and myocyte reprogramming for both markers, suggesting that Myt1l exerts its influence towards repressing non-neuronal fates regardless of neuronal induction. In contrast, addition of Prox1 reduced Tuj1 and Desmin expression in neuron reprogramming, suggesting that this TF is acting as a non-hepatocyte fate repressor. Additionally, Prox1 increased E-cadherin and Albumin expression in hepatocyte reprogramming. Tbx5, on the other hand, increased expression of Desmin, a muscle marker, specifically when added to neuron and hepatocyte reprogramming, but not in during myocyte reprogramming. We hypothesize the lack of Tbx5 to induce Desmin might be due to Desmin not being a proper marker for tracking cardiomyocyte reprogramming [**Lindskog et al 2015**].

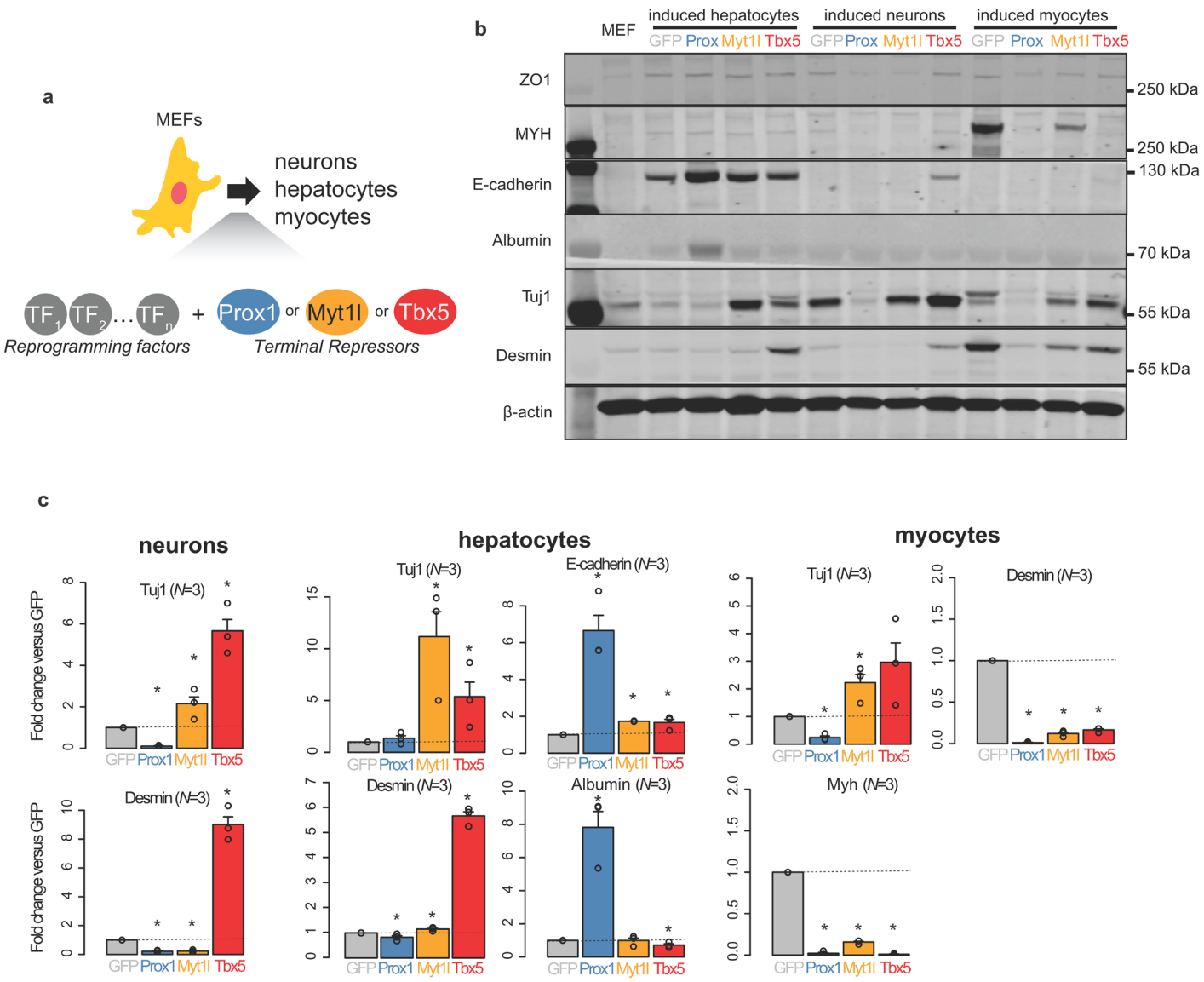

Figure 4.3 Reprogramming experiments to evaluate Prox1 and Tbx5 role as Terminal repressors.

(a) Depiction of reprogramming experiment. Known TFs to induce reprogramming to induced neurons (iNs), induced hepatocytes (iHep) and induced myocytes (iMyo) are added into MEFs, plus either of Prox1 (blue) Myt1l (orange), or Tbx5 (red). (b) Western blot for liver markers (ZO1, Albumin, E-cadherin), neurons (Tuj1), and myocytes (Desmin, MYH). Fold changes are obtained by relative comparison versus GFP abundances, corrected for β-actin (c) Fold changes for marker genes (N=3 replicates). Desmin quantification in hepatocytes conversion is shown as $\log_2$(FC + 1) units. *= adjusted P < 0.1, from two-sided t-test versus GFP, corrected by Benjamini-Hochberg procedure.

On the morphological level, immunofluorescence demonstrates an increase of neuronal cells, and a decrease of alternative hepatic or muscular fates upon addition of Myt1l in the respective reprogramming protocol (**Fig 4.4a**). Similarly, introducing Prox1 into a MEF-to-hepatocyte reprogramming setup decreased the expression on non-hepatocyte genes, and increased the amount of visible [**Li et al 1990**] hepatocyte positive cells in liver (**Fig 4.4b**). It should be noted that these effects upon Prox1 addition were not observed other combination of cell type conversions (myocytes or neurons). We can thus assume that the underlying mechanism of Prox1 that contributes to cell fate determination is highly specific for hepatocyte reprogramming. Finally, the addition of Tbx5 in a cardiac muscle reprogramming setup did not increase the presence of myocytes which could arguably be due to low efficiency in repression of non-cell-fate genes, or its specific role in heart and not in skeletal muscle differentiation (**Fig S4.2b**).

Taken together, our experimental results validate the *in silico* approach to discover repressor potential across TFs. The case-in-point is represented by Prox1 in hepatocyte conversion and the strong increase in reprogramming efficiency due to the addition of the repressor TF.

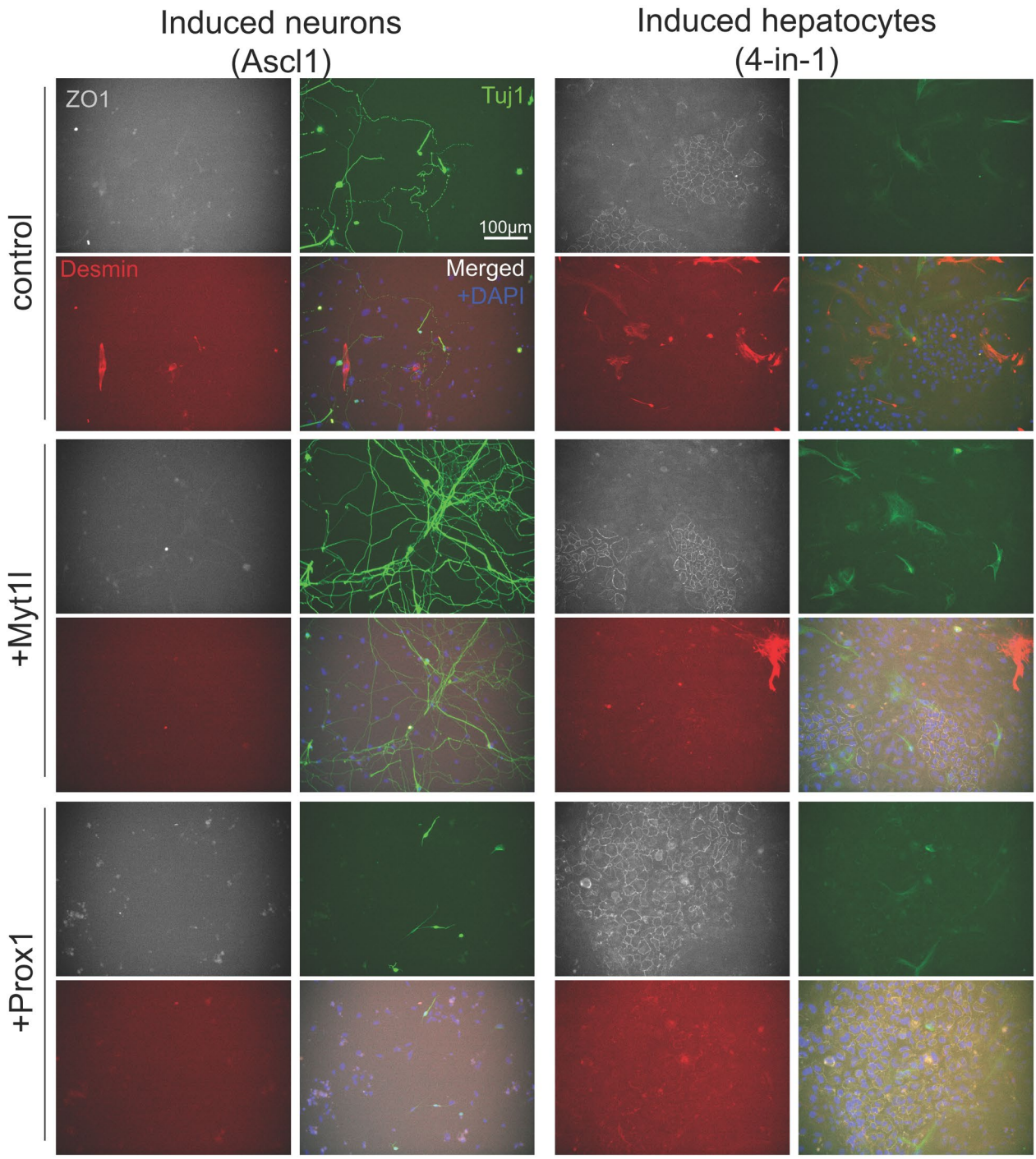

Figure 4.4. Prox1 improves reprogramming to hepatocytes and it does not induce neuronal reprogramming.

Representative visualization of converted cells seven days after the induction of reprogramming factors, showing liver marker ZO1 (upper left), neuronal marker Tuj1 (upper right), muscle marker Desmin (lower left), and merged visualization with nuclear DAPI staining (lower right) (scale bar indicates 100 microns). *(left panels*) Addition of Myt1l (+Myt1l) in a neuronal reprogramming protocol with Ascl1 increased neuron morphology, but not in induced hepatocytes (4-in-1). Addition of Prox1 (+Prox1) in a hepatocytes reprogramming experiment (*right panels)* induces positive ZO1 cells, and it does not contribute to increase in neuronal morphology.

4.3. Discussion

In this work we introduce a classification approach to quantify the enrichment or depletion of Terminal Repressors in multiple cell lines where expression data is available. This approach is based on learning features from Myt1l-containing expression sets and their extrapolation to other cell types for the identification of similar patterns. As new expression data from multiple consortia projects allow a fast annotation of cell-types and their subdivisions, we envision that this approach will be useful in predicting terminal repressors for those.

Our reprogramming validation and the exclusive repression of non-cell type genes in particular cell types of interest reveal some promising directions for future reprogramming trials. Further efforts will need to ensure efficient binding of such repressors during reprogramming using ChIP-seq, and robust quantification of the impact on target genes.

Consensus benchmarks in literature can be consulted to measure performance of reprogramming protocols. The authors of Mogrify [**Rackham et al 2016**] compare reprogramming TFs derived from integration of FANTOM expression [**FANTOM Consortia**] and GRNs datasets and report classification performance based on available benchmarks. A presented comparison of their method with other protocols indicate a higher recall and ranking of reprogramming TFs used to convert fibroblast into macrophages, heart, myocytes, iPSCs, and hepatocytes. Despite this, Mogrify (1) is not able to predict Myt1l as a reprogramming factor in neurons, and (2) use a limited set of available reprogramming protocols [**Tabula Muris Consortia 2018**]. With our approach, on the other hand, we have not only observed FOXQ1, one of their candidates predicted as a relevant TF for keratinocyte conversion, but specifically predict it to be a terminal repressor (**Fig S4.2c**). Further work and a comparison of these approaches in light of new

available data will clearly prove useful for the discovery and prediction of reprogramming TFs.

As multiple expression and reprogramming data resources become available, their integration and comparison will help identify TFs that contribute positively and negatively to cell type conversions [**Guerrero-Ramírez et al 2018**]. Enabling community-driven endeavors in defining common benchmarks and models will be relevant for the collective assessment of the usage of terminal repressors in cell reprogramming.

Chapter 5

Conclusions & Outlook

The rapid advent of sequencing technologies has allowed us in recent years to comprehensively describe the biological consequences of transcription factor (TF) binding at multiple levels. Various experimental techniques, their customized versions, and multi-omics data integration enable the tracking of changes in the regulatory genome concomitant to binding of TFs. Thus, it became possible to directly probe causal gene-regulatory networks and to connect them with downstream biological functions. This thesis has carefully explored the regulation of gene expression in this context and provided answers to the following questions:

Question 1: Can we predict TF-cooperativity and features allowing its prediction in published TF-binding data?

The work presented in **Chapter 2** describes the integration of publicly available SELEX data capturing TF-DNA binding in pairs and alone. The presented statistical framework for the assessment of features contributing to binding predictions allowed us to observe improvements for certain TF family combinations that had not been reported before, such as the one for Forkhead and Ets families. Our findings highlight the potential of describing these interactions across multiple TF pairs and their functional consequences more thoroughly. Importantly, the selection and experimental validation of specific cases of DNA sequences- classified as cooperative or non-cooperative for the FOXO1-ETS1 pairs- substantiated the potential of our computational approach to predict cooperativity.

Question 2: What are the consequences of cooperative TF-binding in function and disease?

The second part of the work described in **Chapter 2** integrates multiple ChIP-seq datasets and prior knowledge of protein-DNA interactions, allowing the interrogation of functional terms related to single TFs and their joint presence. Benchmarking across different databases indicates that these associations increase discovery of ontologies specifically related to TF-TF binding events where both members are part of the ontology. This proves more accurate than approaches where only co-occupied regions are considered without TF-TF binding information. We therefore conclude that knowledge of TF-TF binding combinations independently contributes to biological processes related to function and disease.

Question 3: What is the interplay between TFs and chromatin accessibility and how does it confer specific neuronal activity?

In **Chapter 3** the combined assessment of neuronal accessibility changes and gene expression in mouse cortical neurons for different treatments revealed common and specific features, with strong response-specific TF associations. Functional validation of specific cases through analysis of expression data allowed the dissection of specific response types and their consequences. Specifically, interactions between TFs from the groups bZIP, EGR, CTCF, HIC1 and Hox genes are related to chromatin accessibility and gene expression upon BDNF or KCl stimulations, with a predicted weaker role for HIC1 and EGR in KCl. Strikingly, we also observed differences in exon usage between these treatments that are potentially due to the interaction between CTCF promoter exon-loops, most likely triggered by EGR-related factors and HIC1. These results demonstrate that a well-established interplay of TF combinations have a direct functional impact on the onset of neuronal activity.

Question 4: Can we systematically predict terminal repressors in different cell types?

In **Chapter 4** the combined scanning of differential motifs and expression levels of differentiation-associated TFs enables the identification of TFs related to repression of non-cell-fate genes. Features of terminal repressors such as Myt1l could indeed be systematically recovered for a number of previously unknown terminal repressors in non-neural cell types. Validation experiments highlighted the role of Prox1 in Hepatocytes as well as the potential role of Tbx5 in cardiomyocytes as terminal repressors.

5.1. Relevance of TF cooperativity in the understanding of Biology.

In order to understand the biological consequences of a molecular readout, i.e. interpret the readout *per se*, we take the most explanatory variables in our models and try to predict the system output by perturbing those [**Lazebnik 2002**]. In the context of TF regulation *in vivo*, their contribution in maintaining a cell state seems to be hierarchical, as few TF motifs are able to explain the major changes in genomics data. This regulatory setup is in agreement with the observation of multiple TFs being expressed in multiple tissues, with just a few ones being differentially expressed or not at all in particular cell types [**Wei et al 2018**]. Systematically adding TF-TF interactions into these predictive models as a component allows then to assess the contribution of combinatorial TF-binding to biological processes in general. However, quantifying and assessing the improvement in these types of models is beyond the scope of this dissertation. Further adaptations of the work presented in **Chapter 2**, with incorporation of additional data, could be promising to discover TFs related to function or disease, through interactions with another TF and TF-TF cooperativity.

The discussion on whether such alternative binding models or TF-TF combinations are relevant, arguably ensues a quantitative statement on the additional variance explained with respect to the additional number of features that would be included. In **Chapter 2** and **Chapter 3** we deliberated on the possibility of clustering specific genome regions in an unsupervised way to systematically interrogate the contribution of TF-TF binding to TF-regulatory models. Our results indicate that such TF-cooperative binding events are involved in function and disease. However, the number of such TF-TF binding events is considerably lower than the number of events that involves primary TF-binding. Ultimately, cooperative TF-TF interactions together with single TF binding contribute to

the eventual biological output, and it is the quantitative estimation of the contribution for each TF-TF interaction, on top of single TF binding, that would require further work.

As explained in the Introduction, interactions that are not mediated by direct TF-TF binding (DNA-mediated or DNA-cooperative interactions), can also translate into cooperative interactions, i.e. when TFs bind to proximal or same regulatory *loci* of identical or complementary genes. In the case of neuronal activity, TFs with the pioneer bZIP domain and TFs of the groups Homeobox/EGR/HIC1 and CTCF functionally cooperate to bind similar regions upon stimulation. These interactions modes mediate very specific downstream responses in cortical neurons that cannot be observed for all neuronal activity stimuli. It is the target of further research to determine which and why EGR-related factors and HIC1 partners are acting specifically upon BDNF treatment and not upon KCl stimulation (**Chapter 3**).

Given that cell differentiation and fate maintenance heavily rely on a tight control by TFs, it is an open question whether TF combinations affect differentiation and to what extent those combinations can be identified (**Chapter 4**). From our results in **Chapter 2** we can conclude that specific TF pairs are linked to genes regulating neuron differentiation and keratinization (**Fig S2.7**). This sheds light on DNA-facilitated TF-binding determining the activation and repression of target pathways related to cell fate in a TF-TF specific manner. Ultimately, leveraging such additional knowledge on TF-TF binding and their functional consequences in prediction models could allow a rational prioritization of TFs for cell reprogramming.

5.2. Integration of TF binding data with other -omics datasets

The vast amounts of TF binding data, along with the remarkable improvements in the computational methods for analysis and integration will undoubtedly spur an upcoming progress in the active interpretation of the genome and its regulation. Combined models that take into account genomic data across studies is a priority in bioinformatics methods. Preliminary solutions such as the UniBind database [**Gheorghe et al 2019**] allow for a comparison of TF binding models, but lack a systematic integration with other omics layers. New tools such as Virtual ChIP-seq [**Karimzadeh et al 2019**], for example, attempt to predict TF binding from expression and ATAC-seq data, but have implemented their TF binding priors based on PWMs scores only (see **Introduction 1.2**). Considering all the caveats of PWM-based models discussed previously by others [**Ruan et al 2017**] and in this work, models for such resources need to be adjusted accordingly. It will take a community effort and large-scale coordination for methodologies to shift from mere PWMs to inclusion of *k*-mers [**Guo et al 2018**], biophysical models [**Rastogi et al 2019**], and TF-TF cooperativity. Such a change in the methodological paradigm in the TF-modeling promises to have a big impact in the upcoming years, allowing the description of wider binding affinity spectrum for protein-DNA interactions.

Moreover, the rules guiding cooperativity, as presented in this work, do not just apply to TF-DNA interactions, but to nucleic acids in general, including interactions between proteins and RNA. It is widely assumed that RNA-recognition properties of proteins are comparable to sequence-specific TFs [**Jolma et al 2019**], but with higher degrees of freedom of the RNA-molecule during the formation of tertiary structures. This work hence gives some directions on potential assessment of RNA-binding protein cooperativity from available data [**Ray et al 2017**], and their association to biological processes.

Consortia efforts have associated human genetic variation to molecular phenotypes at the level of Single Nucleotide Polymorphisms (SNPs). A fraction of those SNPs has been associated to TF binding, either at the level of allele-specific binding or at the level of chromatin marks affected by such [**Shi et al 2016**]. As 15-20% of all SNPs seem to be associated to a TF binding model [**Grubert et al 2015; 1000 Genomes Project Consortium**], an unanswered question is to what extent TF-TF interactions cooperativity binding can increase the unexplained fraction, and ultimately bridge the gap between genetic variation and phenotypes. As GWAS variants are linked to biologically relevant phenotypes through TF-cooperative binding [**Iwata et al 2017**], it is expected that more variants associated to low-affinity TF-binding sites and TF-TF interactions will be discovered in the next years. Approaches such as the one developed in **Chapter 2** will be helpful to readily mine public genetic data for this purpose.

Expression data has been proven useful to assess TF combinations and their enrichment in ChIP-seq data as co-regulators [**Mariani et al 2017**]. As TF-cooperative binding events are promiscuous but limited to certain TFs families, the results described in all chapters indicate novel ways of reducing the search space by using TF expression levels, and monitoring cooperativity events between TFs in context of certain molecular phenotypes. The future incorporation of known or predicted TF-interactions, their expression and any additional data will be useful for the assessment of how specific TF-pairs are linked to function or disease.

The findings and discussions presented in this dissertation should ultimately shed light on TF cooperativity as an important contributor to the precise regulation of a biological system. With TFs being prime determinants of cellular programs, the further study of the molecular interplay between TFs and their functional consequences is crucial to the overall understanding of biology.

Appendix A

Computational Materials and Methods

Related to Chapter 2

Transcription factor binding datasets

Transcription factor (TF) binding data analyzed in this work was collected from *in vitro* and *in vivo* studies. Specifically, CAP-SELEX and HT-SELEX sequencing reads are retrieved from the European Nucleotide Archive, under accession entries PRJEB7934, PRJEB7934, and PRJEB20112, respectively. Protein Binding Microarray (PBM) data was downloaded from the UniProbe database [**Newburger et al 2013**]. ChIP-seq peak datasets are collected from the ReMap2 database [**Chèneby *et al.* 2018**].

In vitro data preparation

The first step in the computational processing of SELEX data is the generation of count tables for *k*-mers (sequence patterns of length *k*), for each experiment where a TF or a TF pair was processed. CAP-SELEX sequencing data used to generate count tables always comes from a fixed selection round (positive) and is compared against an input library (background, or round zero). For each TF pair we select the positive round where a binding motif targeted by the two TFs is overrepresented in the reads for a given topology versus all other possible topologies [**Jolma *et al.* 2015**]. From this, the initial value of *k* is

the length of the reported reference *k*-mers, trimming out ambiguous nucleotides in the flanking (N). For example, the reference *k*-mer GAAAACCGAANM has a length of 12, and thus k = 12. If more than one reference *k*-mer is enriched in one dataset, those are processed independently.

Once *k*-mer tables are defined, relative affinity estimates for each *k*-mer can be obtained from the counts of each *k*-mer observed in the positive round, versus the amount estimated in the input data (round zero) using a fifth-order Markov Model. This correction takes into account sequencing biases **[Riley *et al.* 2014].** Given this information, for a *k*-mer k in selection round r, its relative affinity or $S(k,r)$ is calculated as

$$S(k,r) = \sqrt[1+r]{P_{obs}(k,r)/P_{exp}(k,r)}$$

Where $P_{obs}(k,r)$ is the fraction of counts for k in r, and $P_{exp}(k,r)$ is the expected fraction of counts for k in round r. The derivation of the formula has been extensively described in previous work **[Riley *et al* 2014].**

From the *k*-mer tables and their relative affinity estimates, we further subset this table for *k*-mers with high similarity between those and the reference *k*-mers indicated to be enriched, allowing up m mismatches **[Yang *et al.* 2017]**. The m value threshold is proportional to the consensus sequence length and the information content of each of its nucleotides, using the proposed formula

$$m = \lfloor (L - 4)/2 \rfloor + 1$$

where L is the length of the consensus sequence, corrected by the ambiguity of each nucleotide. E.g. GAGCA has an L value of 5, but RRGCA has an L value of 4, as R can represent either G or A.

Tiled k-mer tables

There is an exponential decrease in the counts recovered per *k*-mer and the value of *k*, which prevents the calculation of robust *k*-mer tables and robust relative affinity estimates for high *k* values. To overcome this, we trimmed nucleotides from both flanking regions of each consensus sequence in the list derived from the CAP-SELEX data. We thereby obtain tiled *k*-mer tables with sufficient counts for further analyses. To avoid lower complexity of DNA sequences, *tiled k*-mer tables of length lower than 10 are not considered for further analyses. Our trimming approach was benchmarked through a comparison of the effect of shorter *k*-mers in the final performance metrics (see section "Trim-and-summarize coefficient of determination" and **Fig S1a-b and S4**). To avoid relative affinity estimates with low support, minimum threshold of counts per *k*-mer are defined [**Riley *et al* 2014; Yang *et al* 2017**]. In this work, *k*-mers derived from CAP-SELEX data are discarded if the number of counts is lower than 20 counts.

Regression models and features describing SELEX relative affinities

To relate binding affinities with sequence and/or shape features we used L2-regularized Multiple Linear Regression (L2-MLR), in the form:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where y is the vector of relative affinities for each *k*-mer in the *k*-mer table, whereas X represents a concatenated set of features that encode their respective DNA sequences, β_i (i=1,...,n) represent the regression coefficients, and $\beta_o$represents the intercept. To prevent overfitting, L2-regularization employs an additional penalty term on the coefficients in the loss function $L(\beta)$, i.e. coefficients are obtained by minimizing

$$L(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} \beta^2{}_j$$

with λ set to one.

For regression models based on DNA sequence features, the baseline models are named *1mer* and are defined by mononucleotide representations of each k-mer. At any k-mer position i, four features $4i$ to $4i + j$ with $j < 4$ are defined based on the nucleotide identity

$$X_{N_j,4i+j} = \begin{cases} 1, if\ k_i = N_j \\ 0, if\ k_i \neq N_j \end{cases}$$

of k_i:

$$with\ N_0 = A, N_1 = C, N_2 = G, N_3 = T$$

In total, *1mer* models require $4k$ features for each sequence of length k to fully encode its sequence in numbers. For *2mer* or *3mer* models, dinucleotide or trinucleotides are also converted into coefficients, thus requiring more features per position. For *2mer* models features, 16 coefficients between $16i$ to $16i + j$ with $j < 15$ features are necessary to describe the dinucleotide identity of each k-mer position and its immediate right-nucleotide

$$X_{16i+j} = \begin{cases} 1, if\ k_i k_{i+1} = N_j \\ 0, if\ k_i k_{i+1} \neq N_j \end{cases}$$

$$with\ N_0 = AA, N_1 = AC, N_2 = AC, N_3 = AG, \ldots, N_{15} = TT$$

Similarly, for *3mer* models 64 features representing all the possibilities for trinucleotides are required. In general, for a *Nmer* model where $N \in Z^+$, 4^N would be required per k-mer position. Combinations of these models require the sum of features for each individual model, per position. For example, *1mer+2mer* models require $4^1 + 4^2 = 20$ coefficients per position. Equivalences between some of these models are further described in [**Yang *et al* 2017**].

Models that include DNA-shape features are labeled with the keyword *shape* (e.g. *1mer+shape*), and consider DNA structure estimated for each tested DNA-sequence in all datasets, defined as descriptors of the overall DNA structure for that sequence. These

values are listed in a DNA pentamers table, and are obtained from the DNAShapeR package [**Chiu *et al* 2016**] centering each feature value on the middle nucleotide. In this work, we considered the four main features provided in the original version of this table: Propeller twist (ProT), Roll, Helical Twist (HelT), and Minor Groove Width (MGW). In addition to these values, second order shape values are obtained by calculating the product of features in two consecutive positions, as a way to describe longer structure features. For that reason, 4 main shape and 4 second order shape features are required per position, allowing for 8 features per position to be described in *shape* models. Additively, *1mer+shape* models require 4 + 8 = 12 features per position where a centered DNA pentamers exists.

Flanking positions cannot be described by shape features as these miss one or two nucleotides to successfully map a DNA pentamers. Solutions such as describing the flanks as *3mer* features have been proposed (*1mer+3merE2,* where *E2* symbolizes using *3mer* features in the two end positions) [**Yang *et al* 2017**]. In this work, we extended the *shape* model features to include flanking regions as well by including the average feature value of all pentamers that contain a common tetramer or trimer as found in the flanking region. Briefly, whenever a shape feature in the flanking regions is required, we average pentamers shape features that contain a fixed trimer (16 options) or tetramer (4 options). This is done with similar rules and upstream or downstream of the *k*-mer flank, according to the 5' to 3' directionality (left flank = upstream trimming, right flank = downstream trimming), respectively. We calculated errors for each DNA-pentamers to estimate the amount of uncertainty for each calculation using all trimers and tetramers available in the dataset in comparison with all DNA-pentamers. Shape features based on averaging across trimers and tetramers are closer to real pentamers DNA-shape features than the global mean generated by using all 1024 DNA pentamers or scrambled versions

of it. In this work, we refer to *shape* features as models that include these flanking features.

Trim-and-summarize coefficient of determination

For each tiled *k*-mer table in each dataset, we use a 10-fold cross validation scheme to randomly separate the table into 10 fixed groups of equal size, iteratively fitting L2-MLR models with 9 out of ten groups, and then assessing coefficient of determination (R^2) in the held-out group. This is done using scikit-learn [**Pedregosa *et al.* 2011**]. As a summary statistic for each tiled *k*-mer table, we report the median R^2 of all held-out groups. As a quality control and to remove datasets with low variability and enrichment for mapped *k*-mers, at this stage we filter out datasets whose minimum testing R^2 value across all models for all tiled *k*-mer tables is lower than zero (i.e. fitting is worse than using the mean of all values as a single feature).

To generate a global R^2 for each dataset and model combination, we calculate the median of all median 10-fold CV R^2 values in each tiled *k*-mer table when using a reference *k*-mer. We refer to this as the 'trim-and-summarize' R^2 performance, and use this number for global performance comparison across models and datasets (**Fig S2.1a**). To validate that this metric is a robust approach to obtain global R^2 values without major information loss, we tested whether this approach provides similar R^2 statistics to the ones reported by Yang *et al* in HT-SELEX data [**Yang *et al* 2017**], comparing reference *k*-mers and *tiled* *k*-mers. Globally, R^2 statistics are in strong agreement, defined as a difference of less than 3 nucleotides between models for reference and tiled *k*-mers. Hence we conclude that R^2 values obtained through trim-and-summarize are indicative of longer *k*-mer R^2 values as long as the length difference with respect to used *tiled* *k*-mers is three or less (**Fig S2b**).

TF family shape-specific improvements in combinatorial binding

We assessed the relationship between our trim-and-summarize R^2 improvements by DNA-shape features and specific TF family membership of each studied pair of TFs. Briefly, each annotation for a TF to a particular protein structure family is retrieved by the JASPAR database [**Mathelier *et al.* 2014**]. Significant increases in R^2 are assessed using a Wilcoxon rank sum test (`wilcox.test` in `R`), with p-values being corrected using a Benjamini Hochberg procedure (`p.adjust` in `R`) [**Benjamini *et al.* 1995**].

To discard bispecificity in the Forkhead+Ets datasets as a feature explaining the R^2 improvements, we repeated the calculation for these datasets discarding all k-mers containing the pattern GACGC up to one mismatch (**Fig S1d**).

Cooperativity estimations in matched CAP-SELEX and HT-SELEX data

To estimate TF cooperativity we used relative affinities obtained from CAP-SELEX and HT-SELEX data, and their predicted scores from *1mer+shape* models. We defined the ratio between predicted relative affinities for a TF pair and matched single TF datasets, to estimate how close a CAP-SELEX score is to the average score in matched HT-SELEX data that would be expected for non-cooperative binding. Hence, the cooperativity for a k-mer k is defined as

$$cooperativity = S_{ab}(k) \,/\, mean(S_a(k_a), S_b(k_b))$$

where $S_{ab}(k)$ is the predicted relative affinity in CAP-SELEX for TFs *a and b*, and $S_a(k_a)$ and $S_b(k_b)$ are the predicted relative affinity estimates obtained for TFs a and b in HT-SELEX data for subsequences k_a and k_b that are contained in k. Lengths for k_a and k_b are selected based on [**Yang *et al* 2017**]. This score can be used to calculate the relative cooperativity for specific DNA sequences within a TF pair given the three experiments that are available. In this study we limited calculations to DNA sequences that contain motifs

associated to at least one TF, to prevent calculation of cooperativity estimates on DNA sequences that are linked to amplification and sequencing biases rather than TF binding specificity.

Since no HT-SELEX data was available for ETS1 we used the FOXO1:ELK3 CAP-SELEX dataset to estimate cooperativity factors for the TF pair FOXO1:ETS1, as it contains one common member, FOXO1, and ELK3 is a paralog to ETS1. To generate *k*-mer tables, we used *k*=13 for FOXO1:ELK3 (reference *k*-mer: RTMAACAGGAAGT), *k*=12 for ETS1 (NNNNGGAANNNN), and *k*=8 for FOXO1 (RTAAACAW). This setup allows to measure cooperativity estimates using *1mer+shape* models that contain the Forkhead-Ets 13-mer binding pattern, plus two 3' flanking positions to align the ETS1 binding model (minimum *k* = 15).

PBM data analysis

To examine the DNA binding affinity of Forkhead TFs for ω-none, ω and ω-high, we used PBM data from the UniProbe database to compare E-scores for all *8*-mers containing the patterns GTAAACA, AACAACA, and ACGCACC across all available Forkhead family members. The E-score threshold of 0.35 is used to define high-affinity sites.

Shape profiles calculation

To quantify the contribution of DNA-shape in each TF binding position, we adapted an approach based on a conservative estimation of the performance change in R^2 after adding or removing a given shape feature in a given position [**Yang *et al* 2017**]. Briefly, for each position *i* in the TF binding model based on a consensus *k*-mer of length *k*, we calculated the minimum absolute change in the R^2 value (ΔR^2) between two schemes: (i) increase after adding shape features in a sequence-only model (*1mer+shape*$_i$) and (ii) decrease after removing a shape feature in a shape-only model (*shape-shape*$_i$)

$$\text{(i) } a = \max(\Delta R^2(1mer, 1mer + shape_i), 0)$$

$$\text{(ii) } b = \max(-\Delta R^2(shape_i, shape_i - shape_i), 0)$$

From these two values, then the ΔR^2 per position (ΔR^2_p) is defined as

$$\text{(iii) } \Delta R^2_p = \min(a, b)$$

We considered *tiled k*-mers for the calculation of this value in CAP-SELEX data, so as the improvements are summarized by the median across all aligned positions in *tiled k*-mers. Similar to the trim-and-summarize R^2 comparisons in HT-SELEX data, we tested whether this scheme produces reliable agreements between ΔR^2_p profiles obtained between reference *k*-mers and their shorter *tiled k*-mers. For a number of trimmed positions equal to three, we have obtained a positive correlation distribution between k-mers and trim-and-summarize using shorter, tiled k-mers, which validates this approach for small trimming values (three or less) (**Fig S2.1a-b**).

Clustering of shape profiles across SELEX datasets

Comparing the similarity of ΔR^2_p values between all SELEX datasets requires alignment and assessment of similarity between binding models generated by *k*-mer tables of different length. To align such cases we introduced an unbiased clustering scheme. Briefly, we applied a cubic spline interpolation to all shape profiles of a TF binding model to normalize them to 1000 points (function interp1d, from scikit-learn). Sometimes shape profiles can be mirrored and maximum ΔR^2_p values can be recovered in opposite positions across binding models (e.g. a TF binding model of length 10 with maximum ΔR^2_p value at position 3 contains its complementary model with maximum ΔR^2_p at position 7). To account for these cases we inverted the shape profile if the improvement in maximum performance was located at positions after the respective profile mean (position 500).

Using these shape profiles with common length, we clustered them using a partitioning around medoids (PAM) routine implemented by the pam function (package cluster in R) with a defined number of clusters between 2 and 10. For each cluster, we calculated a cluster-specific TF family and TF enrichment as the odds ratio between the number of datasets for a TF family or a TF associated to this cluster versus the number of datasets for that same TF family or TF in all others clusters (Fisher's exact test, using function `fisher_exact`, from `scipy`). Significance *p*-values were corrected using the Benjamini Hochberg procedure. To select the reported number of clusters (five), we iteratively assessed the total number of TF and TFs families reported as enriched, stopping at the minimum clustering value that maximizes the number of raw p-values lower than 0.05 (**Fig S1.5a**).

Analysis of Forkhead-Ets members using shape profiles

To compare shape profiles of double and matched single TF datasets that have a common Forkhead TF member, we studied the ΔR^2_p values for FOXO1 and FOXI1, as the corresponding CAP-SELEX datasets are enriched in cluster 1 and most of their TF-pair combinations have an equivalent topology. To align TF binding models generated from CAP-SELEX and HT-SELEX, we used the consensus sequence motif of the Forkhead TF (listed in the reference *k*-mer) as an anchor point. Then, we maximized the number of matches between the Forkhead motif region and the reference consensus sequence across all composite motifs (FOXO1: RWMAAAC;FOXI1: RTMAAC). For ETS1, we used the GGAA pattern for alignment. HT-SELEX data for comparison was retrieved for FOXO1, FOXI1 and ETS1 using the available IDs in each case. Since these datasets capture short motifs, shape profiles can be generated using a single *k*-mer representing the consensus binding motif. Reference *k*-mers were used as in the CIS-BP database [**Weirauch *et al.* 2014**]. For aligning and

comparing the profiles with the respective profiles for FOXO1, FOXI1 and ETS members we matched HT-SELEX *k*-mers to the respective composite *k*-mers reported for FOXO1 and ELK3 (ETS1 paralog), using the individual core motif for alignment, respectively.

FOXO1:ETS1 crystal structure is visualized in PyMOL [**DeLano *et al.* 2009**] from PDB ID:4LG0, and enhanced with the PDIviz software [**Ribeiro *et al.* 2016**]. Conservation of positive charge in the ETS1 residue 409 is calculated from the Pfam ID PF00178 (Ets-domain).

TF-TF motif enrichments in co-occupied ChIP-seq peaks

ChIP-seq peaks used to assess TF-TF motifs *in vivo* were retrieved from the ReMap2 database [**Cheneby *et al* 2018**]. Matched TF pairs from CAP-SELEX data were associated to ChIP-seq data when peaks for both TFs were available. For obtaining common summit regions, we intersected the respective peak ranges centered around the peak summit with fixed length of 200 bp using bedtools (function `intersect`). These co-occupied regions are defined as the foreground set of peaks for each TF-TF pair. We discarded TF pair datasets that had less than 50 co-occupied peaks, recovering a total of 105 datasets. The background set, was defined as follows: For each foreground set, an equal number of 200 bp-long sequences with similar %GC content distribution was obtained from mappable hg19 regions (wgEncodeCrgMapabilityAlign36mer, downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability), using the *BiasAway* software package [**Worsley-Hunt *et al.* 2014**].

To map motifs in these sequences, we prepared Position Weight Matrices (*PWMs*) from the Position Frequency Matrix provided in the CAP-SELEX dataset [**Jolma *et al.* 2015**]. In both foreground and background sequences we scored the best *PWM* motif hit per sequence, as the sequence that generates the highest score. These scores are used to

define single feature models, labeled as *PWM*. Additionally, to define *PWM+shape* models [**Mathelier *et al* 2015**], we extracted the DNA-shape features obtained for genomic regions aligned to all positions where a motif hit was obtained, using bwtool.

The ability of these features to separate foreground from background regions in each dataset was assessed as a classification task using Gradient boosting tree classifiers (XGBClassifier library [**Chen *et al* 2016**]). Predictive features were independently centered and scaled. In a 10-fold cross validation scheme, the overall classification performance for each model and dataset was summarized as the median Area Under The Receiver Operating Characteristic Curve (ROC-AUC).

To assess the improvement of TF families in *in vivo* datasets when using *1mer+shape* vs *1mer* models, we used their JASPAR family assignments, equivalently to the *in vitro* data analyses. For each TF family we specifically compared whether the ROC-AUC value differences between *PWM+shape* and *PWM* models (ΔROC-AUC) were significantly higher relative to all other datasets. Significance of the comparisons was assessed by a Wilcoxon rank sum test, and p-values were corrected using the Benjamini Hochberg procedure.

in vitro and *in vivo* positional improvement correlations

Similar to the ΔR^2_p calculation in SELEX data, we generated ΔROC-AUC values per position for *in vivo* data, calculating changes in classification performance after addition and removal of *shape* features in each position *i* on *in vivo* models.

$$\text{(i) } a = \max(\Delta\text{ROC-AUC}(PWM, PWM + shape_i), 0)$$

$$\text{(ii) } b = \max(-\Delta\text{ROC-AUC}(shape_i, shape_i - shape_i), 0)$$

$$\text{(iii) } \Delta\text{ROC-AUC per position} = \min(a, b)$$

In each matched TF-TF dataset with CAP-SELEX and ChIP-seq data, we aligned and compared ΔR^2_p values obtained from *in vitro 1mer+shape* models and ΔROC-AUC per position values obtained from *in vivo*

PWM+shape models using Spearman correlation. To estimate a False Discovery Rate (FDR) threshold for these correlation values, we scrambled correlation values in each model once and recalculated correlations.

Enrichment of cooperative and non-cooperative *k*-mers in Forkhead-Ets ChIP-seq data

Selected sequences from our structural validation were mapped into co-occupied peaks to assess their enrichment versus single TF occupied peaks across TF pairs from the Forkhead and Ets families. Briefly, we mapped consensus sequences representing ω-none and ω motifs (GTAAACAGGAA and AACAACAGGAA, respectively), against Forkhead and Ets ChIP-seq in pairs, allowing up to one mismatch in each reported match. This threshold is chosen as it increases the recovery of sequences similar to each pattern, with a minimum overlap between hits in both categories. To compare the number of hits between co-occupied and single TF occupied peaks in each TF pair combination, we calculated the odds ratio of the number of sequences that do or do not containing either of these patterns in co-occupied versus single TF occupied regions:

$$OR = [a / b] / [c / d]$$

a is the number of co-occupied peaks with the motif; *b* is the number of co-occupied peaks without the motif; *c* is the number of single TF occupied peaks with the motif, and *d* is the number of single TF occupied peaks without the motif. We used a Fisher's exact Test to assess the significance of these effect sizes across all assessed TF pairs, correcting *p*-values for multiple testing with the Benjamini Hochberg procedure.

TF pairs and ontology association analyses

Similar to the previous section, we prepared co-occupied and single TF occupied ChIP-seq regions for all TF pairs in the ReMap2 database, with full or partial match to CAP-SELEX data. Full match indicates that both TFs in the ChIP-seq data are the same as in the CAP-SELEX data. Partial matches are two TFs that belong to the same TF family, and are annotated based on the idea that TF pair share composite motifs that are conserved within paralogs of the same family [**Narasimhan *et al* 2015**]. This knowledge can be used to extend the search to TFs of the same family for which no CAP-SELEX data is available. An example of this is provided with the TF pair FOXO1:ETV4: Both TFs are present in a CAP-SELEX dataset, and there are ReMap2 ChIP-seq peaks available for FOXO1, ETV1, ETV4 and ETV6. Thus, the TF-TF *k*-mers for FOXO1:ETV4 are used to scan co-occupied ChIP-seq regions of FOXO1:ETV4 (full match), FOXO1:ETV1 (partial match) and FOXO1:ETV6 (partial match). To assign co-occupied and single TF occupied ChIP-seq peaks to biological processes, we first used the software GREAT [**CY McLean *et al* 2010**] with default parameters to map peaks to genes: Peaks are selected if located upstream of a Transcription Start Site (TSS) up to 5000 bp, downstream of a TSS up to 1000 bp, or nearby genes up to 1000 Kbp away from a TSS and in absence of other nearby genes. We then assigned genes to ontologies if they are listed in any of the three following ontologies: Gene Ontology Consortium (GO), Human Phenotypes Ontologies (HPO), and DISEASES database. We only considered terms with at least 10 and no more than 1000 genes, to focus our analysis on terms with an amount of associated genes that allows interpretation.

Using this information, we sought to predict the membership of one or two TFs in a given ontology term, and use this as a proxy for their joint binding being associated to a biological function. This prediction is calculated using co-occupied TF peaks and

counting the number of peak-gene pairs that are part of the ontology term in question. Co-occupied peaks are further stratified as cooperative (using the cooperativity *k*-mers) and non-cooperative.. We assumed that a TF pair (*A*,*B*) is more likely involved in an ontology term (*ont*) based on the number of genes (N_{gene}) and peaks (N_{peak}) reported as part of that ontology term when using co-occupied ($A \cap B$) peaks and their peak-gene pairs. For any *ont* and (*A*,*B*) combination, $N_{gene}$is lower or equal than N_{peak}, as multiple peaks can be mapped to the same gene ($N_{gene} \leq N_{peak}$).

The probability of a TF pair (*A*,*B*) to be associated with any *ont* $P(ont = 1 \mid A, B)$, is directly proportional to N_{gene} and N_{peak}.

$$P(ont = 1 \mid A, B) \propto N_{gene}(ont \mid A \cap B)$$

$$P(ont = 1 \mid A, B) \propto N_{peak}(ont \mid A \cap B)$$

To normalize N_{peak} and N_{gene} across all ontology terms and (*A*,*B*) combinations tested, we randomly sampled 200 times a number of unique regions equal to the observed number of $A \cap B$ from the original union of regions belonging to A and B ($A \cup B$), and recalculated decoy N_{peak} and N_{gene} values. From those we obtained mean (μ) and standard deviation (σ) estimates for the expected N_{gene} and N_{peak} associated to that ontology in case of a false association. This is used to convert N_{peak} and N_{gene} into z-scores

$$Z_{gene} = (N_{gene}(ont|A \cap B) - \mu_{gene}) / \sigma_{gene}$$

$$Z_{gene} = (N_{peak}(ont|A \cap B) - \mu_{peak}) / \sigma_{peak}$$

Equivalently, when using TF-TF *k*-mers the association between *ont* and (*A*,*B*) is proportional to the number of peaks and genes obtained when using $A \cap B$ peaks, with a selection for the presence of TF-TF *k*-mers in those peaks. This is indicated as $N_{gene,k}$and $N_{peak,k}$,where *k* refers to the specific *k*-mer used

$$P(ont = 1 \mid A, B, k) \propto N_{gene,k}(ont \mid A \cap B, k)$$

$$P(ont = 1 \mid A, B, k) \propto N_{peak,k}(ont \mid A \cap B, k)$$

Similar to the previous z-scores, we also normalize $N_{peak,k}$ and $N_{gene,k}$ into z-scores using 200 random samplings

$$Z_{gene,k} = (N_{gene,k}(ont|A \cap B, k) - \mu_{gene,k}) / \sigma_{gene,k}$$

$$Z_{peak,k} = (N_{peak,k}(ont|A \cap B, k) - \mu_{peak,k}) / \sigma_{peak,k}$$

K-mer mismatch thresholds for each $A \cap B$, ont k combination were defined so that they maximize $Z_{gene,k}$ and $Z_{peak,k}$ values. To do this, we allowed up to three mismatches in each k-mer to be mapped into a co-occupied peak, and recalculated the observed $Z_{gene,k}$ or $Z_{peak,k}$ values. When multiple k-mers for a pair (A, B) are available, we selected the one that gives the highest $Z_{gene,k}$ and $Z_{peak,k}$values.

Integrating the resulting four z-scores together, we defined an Ontology Association Probability (OAP) as the probability of an ontology term associated to a TF-pair (A,B)

$$OAP = P(ont = 1 \mid Z_{peak}, Z_{gene}, Z_{gene,k}, Z_{peak.k})$$

This probability is modeled based on the four Z-scores obtained above using a Logistic Regression

$$OAP = 1 / (1 + e^{-(\beta_0 + \beta_1 Z_{peak} + \beta_2 Z_{gene} + \beta_3 Z_{gene,k} + \beta_4 Z_{peak.k})})$$

where β_0 defines the intercept and β_1 to β_4 the Logistic Regression coefficients for each Z-score. This model is limited by the availability of ChIP-seq data, and can be potentially extended as new information is included. If no ChIP-seq data is available, then TSS k-mer information can be considered, and Z-score calculations can be obtained using down sampling from all genes (**Fig S2.7c**).

Classification benchmark and selection of strong associations between ontologies and TF pairs.

To benchmark this modeling scheme, we tested its ability to distinguish ontology terms deemed as positive if one or both TFs in (*A, B*) are listed as genes of the term: "TF1 or TF2" are TF-ontology relationships where one of the TFs is a gene member of that ontology term, whereas "TF2 and TF2" contain both TFs as members of that ontology term. Note these two examples: (i) the HPO term HP:0002488 (Acute leukemia) includes the TF ETV6, but not the TF FOXO1, thereby the TF pair FOXO1:ETV6 has a "TF1 or TF2" relationship to that particular ontology term (ii) The term HP:0002088 (Abnormal lung morphology) lists both MITF1 and FLI1 as gene members, and thereby the pair MITF:FLI1 has a "TF1 and TF2" relationship to that term. "background" terms are all ontology terms of which neither *A* nor *B* are members.

We assessed the predictive performance of the Logistic Regression using a full model with all *z*-scores together (*"peaks+kmers"* = $Z_{peak}, Z_{gene}, Z_{gene,k}, Z_{peak.k}$) and variants with only the peaks or *k*-mer *z*-scores (*peaks* = only Z_{peak} and Z_{gene}; *k-mers* = only $Z_{peak,k}$ and $Z_{gene,k}$). Performance metrics were defined by classification of "TF1 and TF2" terms versus "background" terms, or "TF1 or TF2" versus background terms, independently. Positive to negative ratios for between "TF1 and TF2" and "TF1 or TF2" versus background and total entries benchmarked in each ontology database are: HPO = 0.001 (*N*=99828) and 0.05 (*N*=1715112); DISEASES = 0.001 (*N*=102420) and 0.04 (*N*=2808801), and GO = 0.004 (*N*=166104) and 0.06 (*N*=906524). Using a 10-fold Cross Validation approach, we trained models on 9 portions of data and assessed the testing performance in the held-out portion, reporting the median ROC-AUC and Area Under the Precision Recall Curve (PR-AUC) values using the trapezoidal rule. We compared significant improvement using an

independent *t*-test between the 10 testing performance metrics obtained in each model, correcting *p*-values using the Benjamini Hochberg procedure.

To define strong and weak TF pairs and ontologies, we shuffled the ontology labels ten times to assess the *OAP* mean score that falsely labeled positive (decoy) terms when fitting a model with those. We observed that the majority of mean decoys OAP values are no bigger than 0.1, with slight variations across ontologies and models. Assuming that *OAP* values 0.1 units higher than this empirical mean threshold are unlikely false associations, we used this threshold to separate signal from noise: Any TF pair (*A, B*) and ontology association labeled as a "TF1 or TF2" or "TF1 and TF2" with a *OAP* value 0.1 units greater than the mean of its decoys cases is considered a strong association. If multiple ontology terms for a TF pair satisfy this criteria, we visualize and discuss only the association with the highest *OAP* score (**Fig S2.7b**). For generating the Forkhead-Ets network, we additionally restrict all four *Z*-scores to be greater than zero.

Overall survival calculations

We compared overall survival metadata and RNA expression levels from Chronic Lymphocytic Leukemia patients (*N*=184) from a Blood Cancer cohort [**Dietrich *et al* 2018**]. Groups were separated using high and low expression levels for any TF pair of interest using the normalized counts median of given TFs. We compared between basic models where both TFs have low expression (*low*/*low*), versus models in which both genes have high levels (*high*/*high*), or models of the configuration *high*/*low* or *low*/*high*. Hazard ratios and confidence intervals are calculated using the survival package in R [**TM Therneau et al 2018**]. To correct for the immunoglobulin heavy chain variable gene (IGHV) and p53 mutation statuses we assessed an additional model that indicates if the patient has either of those factors reported as positive, as a single category (*N*=88) (**Fig S2.7d**). Models with

patients with just one of the four combinations of these statuses are not reliable due to low sample numbers.

Related to Chapter 3

RNA-seq computational analysis

We mapped RNA-seq reads in each sample to the *M. musculus* mm10 genome with TopHat2 [**Kim et al 2013**], defining the Gencode v10 transcriptome as a reference. We used mapped reads to call differentially expressed genes using DESeq2 [**Love et al 2014**]. We compared each treatment and time point against its matched control samples using all samples together to estimate dispersions, and calling for differentially expressed genes (DE-genes, with a false discovery rate (FDR) of 10%).

We used unsupervised clustering to study the behavior of DE-peaks across conditions. Briefly, we selected genes with differential expression in at least one treatment and sorted them by significance and variance across conditions, selecting the top 5000 genes. We clustered the mean-corrected expression changes in each gene using partitioning around medoids (PAM) clustering, setting the number of clusters to ten. We compared the enrichment of gene ontology terms in each cluster versus other clusters using topGO [**Alexa et al 2006**], defining the whole genome as background.

Chromatin accessibility data analysis

We mapped ATAC-seq reads in each sample to the *M. musculus* genome build mm10 using bowtie [**Langmead et al 2010**] and with the following parameters. We used mapped reads to call peaks in each treatment and time point with MACS2 [**Zhang et al 2008**], considering pooled control samples as a background for all queries. The following parameters were defined to call peaks: "`—nomodel –shift -75 –extsize 150`". Then,

we jointly analyzed resulting peaks and ATAC-seq reads to call differentially accessible peaks using the R package DiffBind [**Ross-Innes et al 2010**]. To correct counts per peak in all conditions we applied LOESS normalization. We conducted comparisons between experimental conditions and matched controls, obtaining a set of differentially accessible peaks at FDR=10% (DA-peaks), which are labeled as gained or closed based on their positive or negative log2 fold changes versus controls in each time point, respectively.

We performed general genomic annotations for gained and closed DA-peaks in each treatment-time point pair using HOMER [**Heinz et al 2010**]. To assess the enrichment of neuron specific chromatin features, we applied a Hidden Markov Model generated from multiple chromatin marks and ChIP-seq data for mouse neurons, using ChromHMM [**Ernst et al 2012**]. This model considers 15 states in mm9. To interrogate our DA-peaks we converted ranges between mm10 and mm9 genome versions using liftOver [**Hinrichs et al 2006**]. We report the log2 fold enrichment between the number of nucleotides in one of the 15 states, versus the number of nucleotides overlapping with other states, using the function `OverlapEnrichment`.

We used DA-peak enrichments for gene ontologies using binomial and hypergeometric tests as implemented in the GREAT server [**CY McLean et al 2010**]**,** with default parameters to map peaks to genes. Peaks were selected if located upstream of a Transcription Start Site (TSS) up to 5000 bp, downstream of a TSS up to 1000 bp, or nearby genes up to 1000 Kbp away from a TSS and in the absence of other nearby genes. We used unchanged peaks as a background, to control for unspecific neuronal terms.

Motif enrichment analysis

We defined summit-centered 200-bp regions from all DA-peaks as foreground regions, and retrieved background regions for each one using GENRE **[Mariani et al. 2017]**, and a

custom mm10 background. Briefly, a representative background sequence is retrieved from a mouse-specific database of reference regions, with equivalent GC-content and CpG frequency, promoter overlap (extent of the sequence located within 2 kb upstream of a TSS), and repeat overlap. Further details are provided in **[Mariani *et al* 2017]**.

We used foreground and background sequences to map motifs using *(i) 8*-mers listed in a set of 108 transcription factor specificity groups generated from Protein Binding Microarray (PBM) data, and *(ii)* a library of Position Weight Matrices (*PWM*s) models (CIS-BP database [**Weirauch et al 2011; Lambert et al 2018]).** For *8*-mers, we defined the best *8*-mer score per sequence as the best E-score greater than 0.35, or -1 otherwise. For PWM motif hits we used the best motif score in each sequence as reported by FIMO [**CE Grant et al 2011**]. We used these scores to assess sensitivity and specificity using a Receiver Operating Characteristic (ROC) analysis, with foreground and background sequences in each treatment and time point as positive and negative groups, respectively. We used the Area Under the Curve (ROC-AUC), to define significantly enriched TF specificity modules and motifs. We used a Wilcoxon one-sided test, adjusted with a Benjamini Hochberg procedure, to assess enriched modules (FDR = 10% and ROC-AUC greater than 0.55).

TF modules co-enrichment analyses

We prioritized enriched TF-specificity modules (ROC AUC > 0.55) to assess their co-enrichment in DA-peaks for specific combinations. Briefly, we compared the abundance of peaks with *8*-mers for two of more modules. The enrichment of a specific combination of TFs together is DA-peaks is calculated using fold enrichment. The calculation of the exact probability for that fold enrichmentis calculated using the R package

`SuperExactTest` [**Wang et al 2015**], and *P* values are corrected using a Benjamini Hochberg procedure.

In addition to this, we compared DA-peaks log2 fold changes distributions between peaks with two TF modules versus only one of both modules, using a Wilcoxon two-sided test.

Genomic data co-variation and loop data analysis

Peak pair correlations between called peaks from the processed ATAC-data were generated using all peak-pairs that had a maximum distance of 50 Kbp. Peaks were further analyzed with Hi-C to assess changes in correlation distributions, filtering for peaks with one Hi-C loop reported in cortical neurons, embryonic stem cells and neuron progenitor cells. We defined true loops as peak-peak pairs where both peaks are less than 10 Kbp away from the coordinates of a Hi-C genomic peak.

We studied the effects of DA-peaks on DE-genes assessing the enrichment of up-regulated DE-genes proximal to gained DA-peaks. Briefly, we compared the amount of up DE-genes associated to a non-promoter gained DA-peak (peak-TSS distance greater than 2000 bp), and their log2 fold changes in each time point. When two DA-peaks are linked to one DE-gene, the one with the lowest *p*-adjusted value is selected. Then, we compared up-gained events (up-regulated DE-gene; gained DA-peak) with up-closed, down-gained and down-closed pairs using a 2 x 2 contingency table and Fisher's exact test for assessment of enrichment.

CTCF specific analyses at differentially accessible peaks

To assess the enrichment of DA-peaks for CTCF promoter-exon loops we used a previously released dataset of promoter-exon contacts [**Ruiz-Velasco et al 2017**] to compare the odds ratio between DA-peaks in these loops versus unchanged peaks. We

used Fisher's exact test to assess the overrepresentation of DA-peaks in those regions, versus unchanged peaks in those regions. We corrected *p*-values using the Benjamini Hochberg procedure [**Benjamini et al 1995**].

We assessed the enrichment of a convergent or divergent CTCF topology for motif hits in DA-peaks connected to promoter-exon loops. Briefly, gained and closed DA-peaks were that contain at least one CTCF motif hit (CIS-BP ID M06483_1.94d) [**Weirauch et al 2011**] were assessed for enrichment of promoter-exon loop pairs previously reported in [**Ruiz-Velasco et al 2017**]. We quantified Odds Ratios (OR) using the formula

$$OR = [a / b] / [c / d]$$

Where *a* is the number of DA-peaks with a CTCF motif and part of promoter-exon; *b* is the number DA-peaks without a CTCF motif and in a promoter-exon look; *c* is the number of non-DA peaks with a CTCF motif and in a promoter-exon loop, and *d* is the number of non-DA peaks without a CTCF motif and in a promoter-exon loop. We used a Fisher's exact Test to assess the significance of these effect sizes across treatments (BDNF and KCl) and directions (gained and closed peaks), correcting *p*-values for multiple testing with the Benjamini Hochberg procedure.

Differentially used exons analyses

To call differentially used exons we used mapped RNA-seq reads to obtain count tables for all listed exons in Gencode (v10). We used genes with non-zero counts in at least one treatment and time point to perform comparisons between BDNF and KCl matched time points, using DEXSeq [**Anders et al 2012**] to call differentially used exons (FDR=10%). To filter for cases related to CTCF promoter-exon loops, we selected genes containing differentially used exons and CTCF promoter-exon loops from [**Ruiz-Velasco et al 2017**].

Related to Chapter 4

Expression datasets of tissue and cell type gene classification

Using RNA-seq expression data from GTEx and Tabula Muris, as well as their tissue and cell type annotations, we classified genes as part of one specific tissue or cell type using a Z-score normalization. The top 1000 genes belonging to one specific tissue were selected for further analysis. For GTEx data, we averaged the TPM values for each gene in each and tissue across multiple samples, and then calculated Z-scores across tissues using those mean values per tissue. For scRNA-seq data from Tabula Muris, we used normalized CPM units provided by Seurat, and used the mean value per cell type for comparison across cell types.

Motif bias calculations

Calculation of motif biases is based on enrichment of TF-specific motifs in target genes versus non-target genes. We compared the motif fold change in TSS regions surrounding target genes versus non-target genes. To calculate a consensus value for all possible pairwise comparisons between target versus non-target cell types, this comparison was done multiple times for each target-specific gene versus each non-target gene. We removed common genes in each pairwise comparison, focusing only on disjoint genes sets. To summarize all motif fold changes across all comparisons into a single value, we calculated the median of those. This final value is re-scaled into a Z-score metric for each cell type or tissue across all motifs, for comparison of motifs with higher or lower abundance in their target genes versus all other non-target gene groups.

Scoring and prioritization of activators and repressors

Using the expression and motif bias values defined as $Z_{e,g}$ and $Z_{m,g}$, respectively, the final score for prioritization of each gene S_{g} as an activator or repressor is defined as the sum of their absolute values.

$$S_{\mathrm{g}} = \sqrt{Z_{m,g}^2 + Z_{e,g}^2}$$

The higher the S_{g} value, the higher a TF is to be a cell-specific activator or a repressor. The specific classification into activator or repressor is based on the sign of $Z_{m,g}^2$ (greater than zero = activator, lower than zero = repressor)

Appendix B

Experimental Materials and Methods

Related to Chapter 2

Protein cloning, expression and Purification

The ETS1 (331-440) and FOXO1 (143-270) sequences were purchased using Geneart (ThermoFisher). These were amplified and cloned using restriction free cloning into a pETM-22 vector, which comprises a cleavable N-terminal His_6- and Trx-tag. The single mutations were inserted in the pETM-22-ETS1 (331-440) vector using site directed mutagenesis.

Both proteins were expressed and purified from *E. coli* BL21 (DE3), grown in LB medium. The cultures were induced with 0.2 mM IPTG at an OD_{600} of 0.8 and grown overnight at 18°C.

After resuspension of the cells in a buffer containing 50 mM Tris (pH 7.5), 300 mM NaCl, 0.5 mg/ml lysozyme, EDTA free protease inhibitor (Roche) and benzonase, the cells were lysed using a french press. The cleared lysate was applied to a first Ni-NTA column and after washing eluted with an imidazole gradient from 0 to 300 mM. The eluted protein fractions were then cleaved with 3C-protease overnight at 4°C to remove the His_6-tag and simultaneously dialyzed against 0 mM imidazole, 50 mM Bis-Tris (pH 6.5) and 150 mM NaCl. After a second Ni-NTA purification FOXO1 was applied to a S75 gel-filtration

column (GE) equilibrated at 50 mM Bis-Tris and 150 mM NaCl. ETS1 purification involved an additional purification step using a Heparin column to remove DNA. The protein was eluted using a salt gradient from 50 mM to 2 mM NaCl. For NMR titration and backbone assignment experiments of FOXO1 the same purification steps were performed but Minimal medium M9 has been used to isotopically enrich the protein. For ^{15}N- and ^{15}N, ^{13}C-labeled protein expression, $^{15}NH_4Cl$ or $^{15}NH_4Cl$ and ^{13}C-Glucose were used as sole nitrogen and carbon sources, respectively. The final NMR and ITC buffer was 50 mM Bis-Tris, 150 mM NaCl, pH 6.5.

Isothermal Titration Calorimetry

Titrations were carried out using either a MicroCal PEAQ-ITC or a MicroCal iTC200 calorimeter at 25°C. All protein and DNA samples were dialyzed overnight at 4°C against the same buffer containing 50 mM Bis-Tris, pH 6.5 and 150 mM NaCl. For each titration 20 injections of 2 µl of titrant were made at 120 s intervals, while stirring at 750 rpm. Data were reduced with heat spikes from control and baseline corrected. The raw data integration, normalization and titration curve fitting was done using the MicroCal PEAQ-ITC analysis software provided by Malvern.

NMR

All NMR measurements were performed at 298 K on an Avance III Bruker NMR spectrometer with a magnetic field strength, corresponding to a proton Larmor frequency of 600 MHz, equipped with a cryogenic triple resonance gradient probe head. Backbone resonance assignment of FOXO1 was achieved to a completion of 85 % (excluding prolines) using ^{1}H,^{15}N-HSQC, HNCA, CBCA(CO)NH and HNCACB triple resonance experiments [**Sattler et al 1999**] analyzed with CARA (http://cara.nmr.ch).

For all NMR titration experiments a series of ^{1}H,^{15}N-HSQC spectra were recorded of ^{15}N-labeled FOXO1 in absence or presence of equimolar unlabeled ETS1. Different DNAs

(labeled as ω-none, ω and ω-high) were titrated always with the same series of molar equivalents (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0, 1.2) to protein concentration (100 µM). As the DNA stock solution was highly concentrated (10 mM), the dilution effect was negligible but still taken into account. All spectra were processed using NMRPipe [**Delaglio *et al* 1995**] and data analysis was performed using the program Sparky [**Lee *et al* 2015**] for chemical shift perturbation analysis and CCPN for determining dissociation constants by fitting the fast exchange chemical shift perturbations vs. DNA concentration using $A(B + x - \sqrt{(((B + x)^2 - 4x))}$ as a fitting function [**Vranken et al 2005**].

Related to Chapter 3

Primary cortical neuron culture

Prenatal embryos of CD-1 mouse at embryonic day 15 (E15) were used for the isolation of cortical neurons. Embryonic cortex was isolated and dissociated by chopping with scalpel followed by digestion in Accutase (ThermoFisher, A1110501) for 12 mins at 37°C. During digestion we treated the tissue with 250 unit/µL of Benzonase (Millipore, 71206-3) to prevent neuronal clumping due to genomic DNA released from dead cells. Following digestion, neurons were triturated gently and passed through the 40µm cell strainer (BD Falcon, 352340) before plating them onto 6 well plate at a density of $1x10^6$ cells per well. Tissue culture plates were coated with 0.1 mg/mL of Poly-D-Lysine (Sigma, P0899) and 2.5 µg/mL of laminin (Sigma, 11243217001). Primary neuronal cultures were maintained in Neurobasal medium (ThermoFisher, 21103) containing 1% penicillin/streptomycin (ThermoFisher, 15140122), 1% GlutaMAX (ThermoFisher, 35050), and 2% B27 supplement (ThermoFisher, 12587) at 37°C with 5% carbon dioxide in the incubator.

Post-seeding after 1 day in vitro (DIV1), half of the media was replaced with fresh pre-warmed Neurobasal media with all the supplements.

Stimulation with BDNF, KCl and forskolin.

Prior to every stimulation on DIV7, neurons were made quiescent for 2 hours with 100 µM DL-2-amino-5-phosphonopentanoic acid (DL-AP5; Fisher) and 1 µM tetrodotoxin (TTX; Tocris). KCl (55mM) depolarization was performed by adding warmed KCl depolarization buffer (170 mM KCl, 2 mM $CaCl_2$, 1 mM MgCl2and 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)) to a final concentration of 31% directly into the neuronal culture medium and incubated for 1, 6, and 10 hour For the BDNF and forskolin stimulations, neurons were incubated with BDNF (10 ng/ml) and forskolin (10 µM) on DIV7 for 1, 6, and 10 hour.

RNA-seq sample preparation and analysis

Mouse cortical neurons were collected at 1, 6, and 10 hours after each stimulation for RNA isolation. The RNAeasy kit (Qiagen) was used to extract RNA and genomic DNA was digested using the Turbo DNAse kit (Ambion). To assess the quality of RNA all samples were analyzed using Bioanalyser (Agilent Genomics). Only samples with a RIN (RNA integrity number) score above 9 were used for library preparation. To prepare libraries we used the oligo-dT capture kit (NEB) in combination with the NEBNext Ultra II kit. We pooled 24 samples with each sample carrying a distinct barcode and sequenced on NextSeq 500 at EMBL, Heidelberg Gene Core facility.

Differentially used exons analyses

Exon differential exon usage quantification with qPCR

For experimental validation, exons were selected based on the presence of a differential DUE and a promoter-exon CTCF loop. Primary cortical neuron cultures on DIV 7 were stimulated and RNA was isolated for expression analysis of the DUEs. Primers for exons of three gene examples were selected, in addition to neighboring exons for internal control. The exon 29 with additional 3'UTR sequence is a DUE for Trio. Two primers were designed to differentiate exon 29 (DUE) from constitutive exon 28. A forward primer lies within the coding region of exon 29 with a reverse primer in the immediately downstream 3'UTR region. Both forward and reverse primer lies in the coding region of exon 28. For Stxbp5 again primers were designed to allow selective amplification of DUE which is exon 1 against a constitutive exon 5. Similarly for Cpe-201 two set of primers were designed, one for exon 6 which is a DUE and another for constitutive exon 9. The primer sequences used to perform qPCR analysis for DUE are following. Trio: Exon 29+3'UTR Forward (5' CTCAGAGCAACGGGGTAAGAG 3') and Reverse (5' GTGCTGGAGAGCTGGAGTTAG 3'); Exon 28 Forward (5' TGAGTTGCCTCTGCTTGGAG 3') Reverse (5' GGACGCTTGGACTGGATGAA 3'). Stxbp5: Exon 1 Forward (5' CAACATCAGGAAGGTGCTGG 3') Reverse (5' GAAGTGTTCGGACTGGAGCG 3'); Exon 5 Forward (5' TGCCATCTGCCTTTCCAGAG 3') Reverse (5' TGACATAGCCTGAGAGTGTGA 3'). Cpe-201: Exon 6 Forward (5' TGCTTCGAGATCACTGTGGAG 3') Reverse (5' CTGCTCCAGGTAGCTGATGA 3'); Exon 9 Forward (5' TGTCTGGATCTACTTCATTCTTACA 3') Reverse (5' CGCAGTACAGGGTTCACAGA 3'). Following stimulation with BDNF and KCl, the fold change of each exon was quantified using qPCR and comparison against Rpl13 as a reference gene at 1 and 6 hours. Fold change values were used to calculate the exon ratio between each tested exon and their internal exon as a reference.

Related to Chapter 4

Reprogramming experiments with Myt1l/Prox1/Tbx5

Lentiviruses carrying the coding sequences for rtTA, GFP, MyoD, Ascl1, the four hepatocyte reprogramming factors (Foxa3, Gata4, Hnf1a and Hnf4a), Myt1l, Prox1 and Tbx5 were generated in HEK293T cells. Briefly, 3rd generation lentiviral packaging vectors were combined with plasmids containing the sequence of each gene of interest under control of a tetracycline-dependent promoter, complexed to polyethylenimine and transfected into 293T cells. After 48 hours the cell supernatant was harvested, concentrated 100-fold by ultracentrifugation and stored at -80°C.

Mouse embryonic fibroblasts (MEFs) were isolated from limbs of E13.5-E15.5 C57BL/6 mice and grown in DMEM media supplemented with 10% serum, penicillin-streptomycin, MEM non-essential amino acids, sodium pyruvate, glutaMAX-I and ß-mercaptoethanol. Prior to reprogramming cells were plated in 12-well plates (collagen-coated for hepatocyte reprogramming), and infected the next day by the addition of the different lentiviral combinations in the presence of 4 µg/mL polybrene.

After 24 hours the medium was replaced by MEF growth medium with 2 µg/mL of doxycycline, and 72 hours after infection by hepatocyte complete medium (Lonza) + 5% fetal bovine serum + doxycycline for hepatocyte reprogramming, or DMEM/F12 with N2 and B27 supplements, insulin, penicillin-streptomycin and doxycycline for neuronal/muscle reprogramming. From this point onwards, medium was replaced (hepatocytes) or half-exchanged (neuron/muscle) every other day for the duration of the experiment.

Western blotting

Cells were harvested in RIPA buffer (25 mM Tris, 150 mM NaCl, 1% nonidet P-40, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate (SDS)) and sonicated with a Bioruptor Plus before protein quantification using bicinchoninic acid assay and a Multiskan FC microplate photometer. Samples were then diluted to equal concentrations with 5xSDS sample buffer (10% SDS, 50% glycerol, 0.5M dithiothreitol, 0.25% bromophenol blue) and boiled at 98°C.

Protein electrophoresis was achieved by loading samples into 4-12% SDS-PAGE gels and subjecting them to a 100-160V electric current. Proteins were then transferred to nitrocellulose membranes, and total protein was detected using a REVERT total protein stain kit and a Licor Odyssey CLx infrared imager. Membranes were then blocked in 2% bovine serum albumin (BSA), incubated with primary and fluorescent-conjugated secondary antibodies and fluorescence detected with a Licor Odyssey CLx.

Marker gene quantification

Band intensities were obtained by using ImageStudioLite and normalized to total protein loaded in the well. These values were further normalized to the control (GFP-overexpression) sample in each reprogramming set, and the resulting fold-change values along with standard errors of the mean were plotted in R.

Immunofluorescence

Cells were fixed in 4% paraformaldehyde (PFA), permeabilized with 0.5% Triton X-100 and blocked in 2% BSA. Samples were then incubated with the indicated primary antibodies followed by their corresponding fluorescent-conjugated secondary antibodies, fixed again in 4% PFA and counterstained with DAPI before imaging in a Leica DM IL LED fluorescent microscope. Images were processed using the software Fiji [**Schindelin et al 2012**]

Appendix C

Supplementary Material

Related to Chapter 2

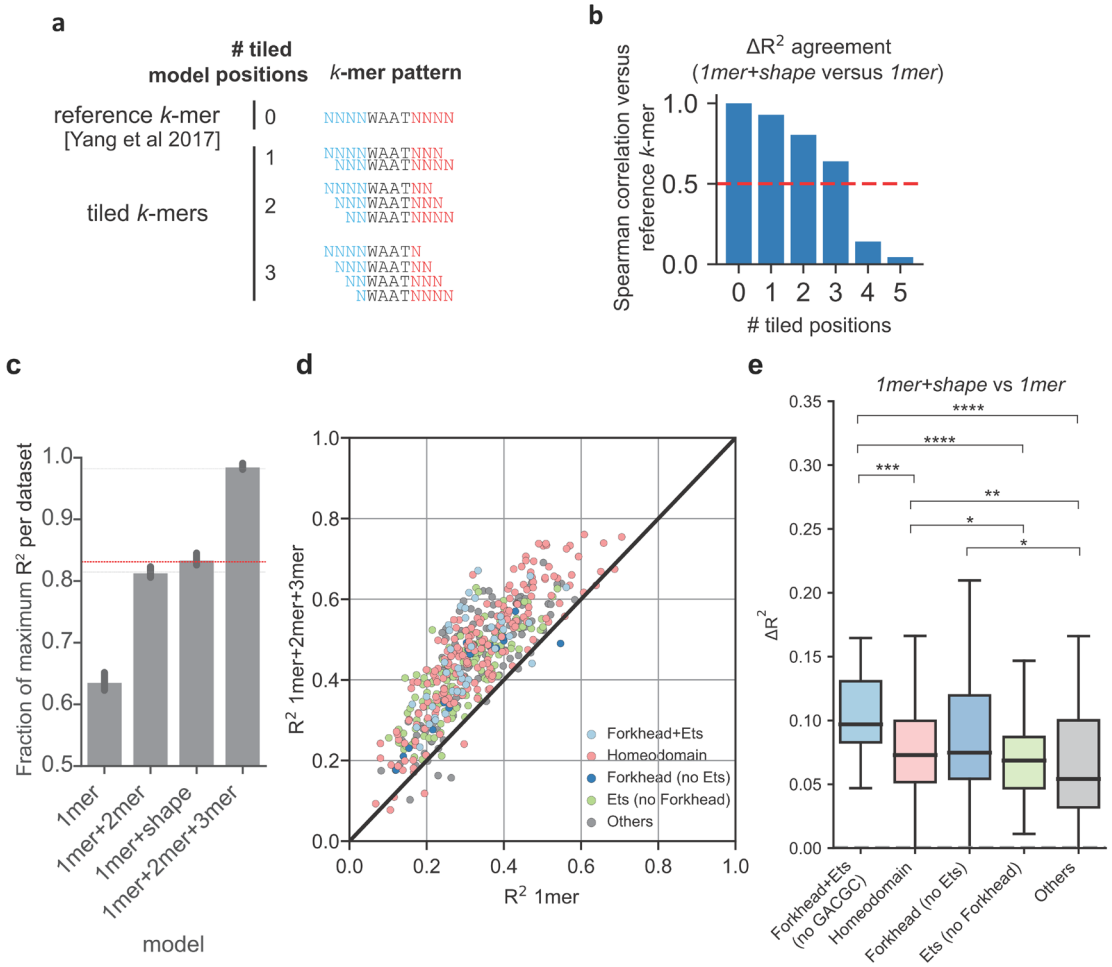

Figure S2.1. Selection of tiled *k*-mers cutoff using HT-SELEX data.

(**a**) Depiction of tiled k-mer approach applied to HT-SELEX data analyzed by Yang et al [**Yang et al 2017**]. Briefly, reference k-mers defined in this work are tiled increasingly, and the correlation between ΔR^2 estimates from *1mer+shape* versus *1mer* models is calculated. (**b**) Spearman correlation between tiled *k*-mers and reference *k*-mers. Red line indicates rho = 0.5, and threshold for selection of tiled *k*-mers in CAP-SELEX data. (**c**) Fraction of maximum variance per R^2 dataset using different model configurations. 1mer+shape model have higher performances than *1mer+2mer* and lower than *1mer+2mer+3mer*, which require more features (**d**) Comparison between *1mer+2mer+3mer* and *1mer* models. Forkhead+Ets datasets show similar improvements and trends as in **Fig 1b**. (**e**) ΔR^2 values observed for Forkhead+Ets datasets after removing *k*-mers containing bi-specificity related pentamers (GACGC) from Forkhead datasets (**Appendix A**). Asterisks as defined in **Fig 1c**.

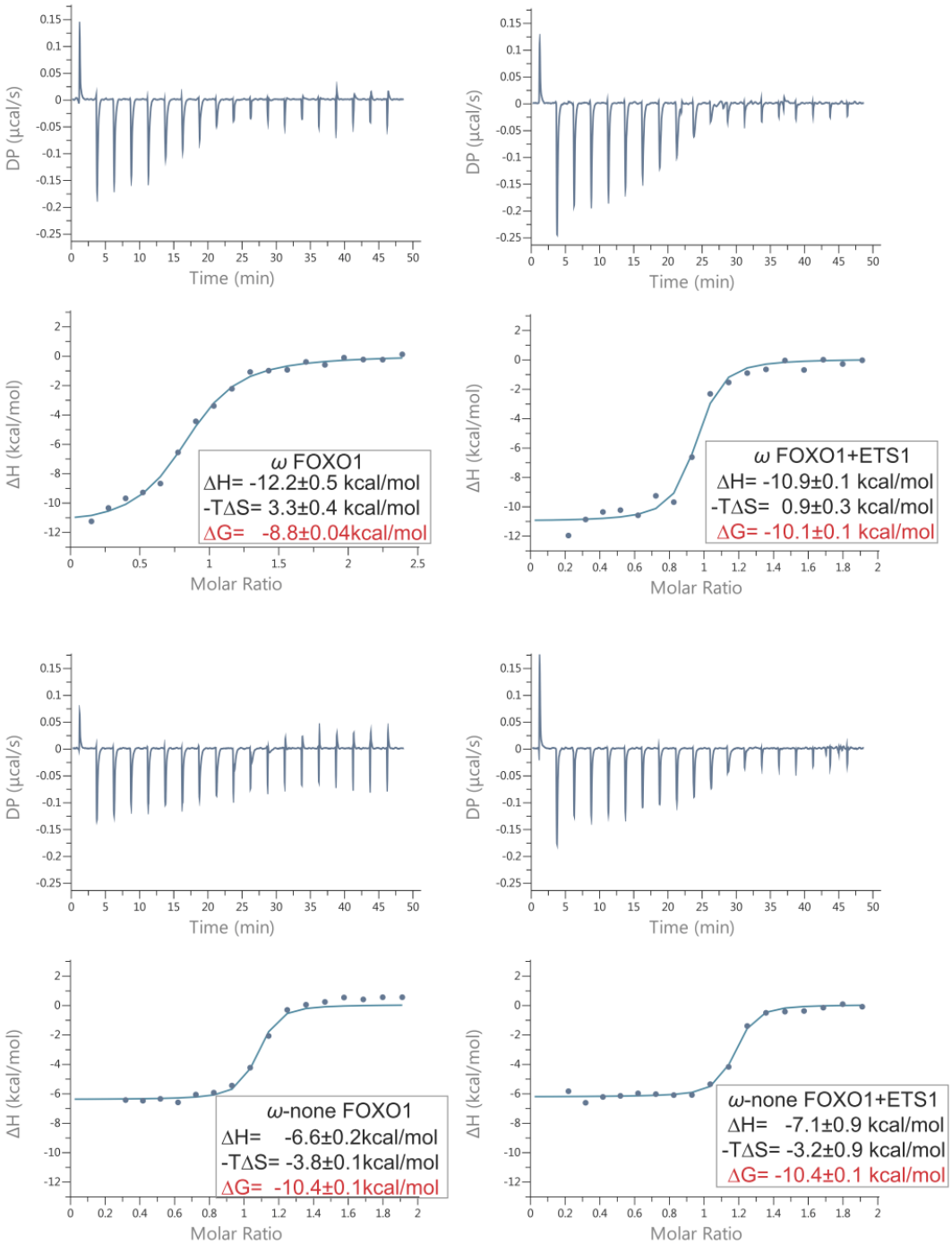

Figure S2.2. FOXO1:ETS1 binding against cooperative and non-cooperative DNA sequences measured with ITC.

ITC calorimetric titration data. Upper plots represent raw data, and lower plots indicate integrated heat of binding reaction. Line in scatter plots represent best fit to the data. Values in lower-right box indicate thermodynamic parameter estimates and their standard deviation (N=2 in each case). Lower panels indicate ΔG parameters (red) for FOXO1 and FOXO1:ETS1 when tested against DNA sequences predicted to have high binding affinity (ω-none) and lower affinity and cooperativity upon addition of ETS1 (ω)

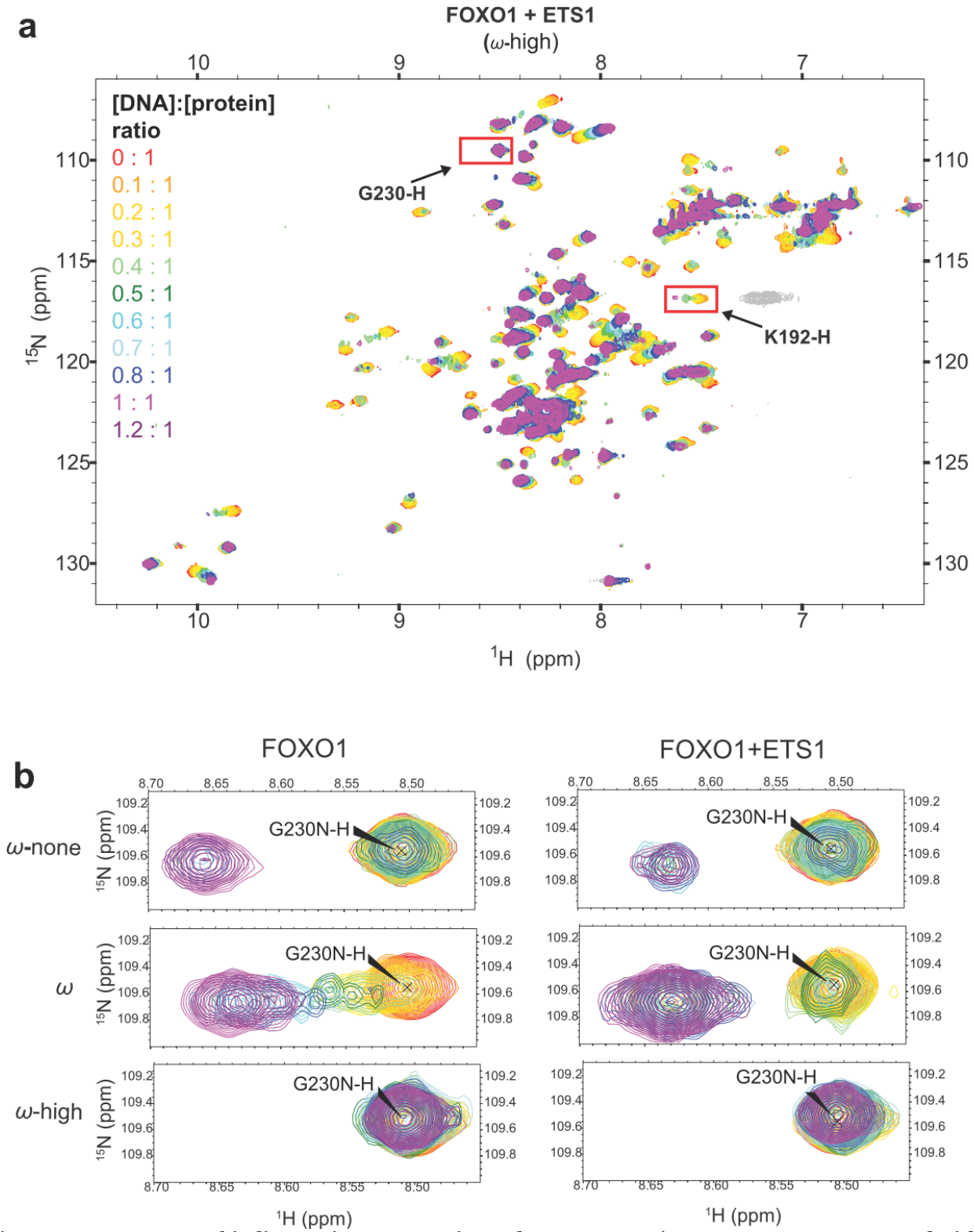

Figure S2.3. FOXO1:ETS1 binding against cooperative and non-cooperative DNA sequences measured with NMR.

(**a**) Full 1H–15N HSQC spectra for FOXO1:ETS1 protein interacting with ω-high. Colors indicate DNA to protein concentration ratios. Highlighted peak residue indicate titration peak for FOXO1 residue K192H-H (**b**) Titrated peak G230N-H from FOXO1 and behavior for three DNA sequences for FOXO1 and FOXO1+ETS1. Titration color scale as in **Fig 2.2e**

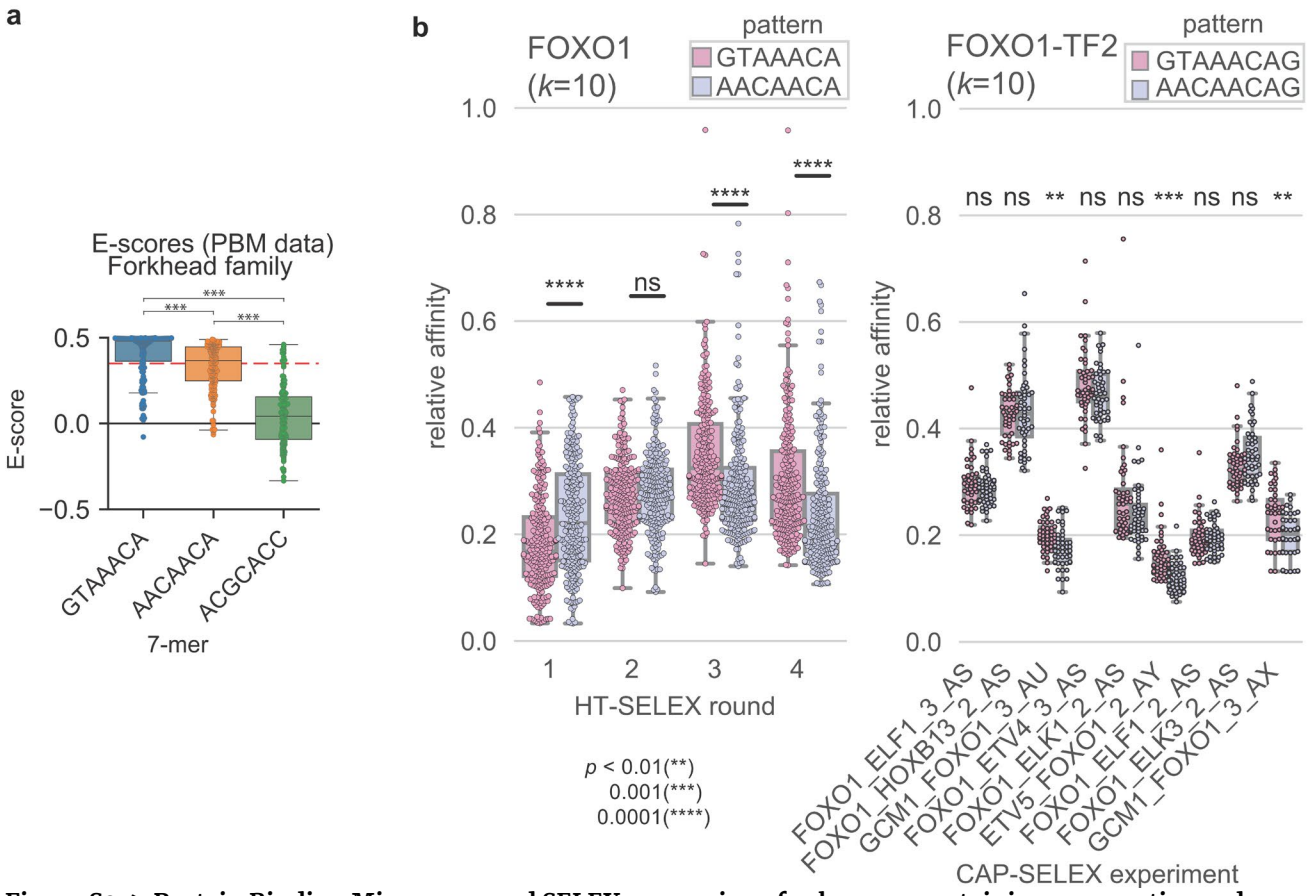

Figure S2.4. Protein Binding Microarray and SELEX comparison for k-mers containing cooperative and non-cooperative Forkhead-Ets sequences.

(a) Protein Binding Microarray E-scores distributions for Forkhead family members. Each boxplot shows a group of 8-mers containing the 7-mer in Forkhead region (ω-none = GTAAACA; ω = AACAACA; ω-high = ACGCACC). Red line indicates 0.35, threshold that defines high affinity sites (b) (left) Relative affinity values for 10-mers containing patterns related to ω-none and ω. X-axis indicates increasing HT-SELEX selection rounds. The distribution of relative affinities for k-mers containing the high affinity FOXO1 site is inverted at round 3 and 4 with respect to initial rounds. (right) Relative affinities for 10-mers obtained from CAP-SELEX data from FOXO1 and TF partners. X-axis indicates experiment ID, including TF1, TF2, selection round and barcode ID. 6 out of 9 comparisons show no differences in relative affinities distributions. Statistical comparisons were done using independent t-test, with Benjamini Hochberg correction procedure.

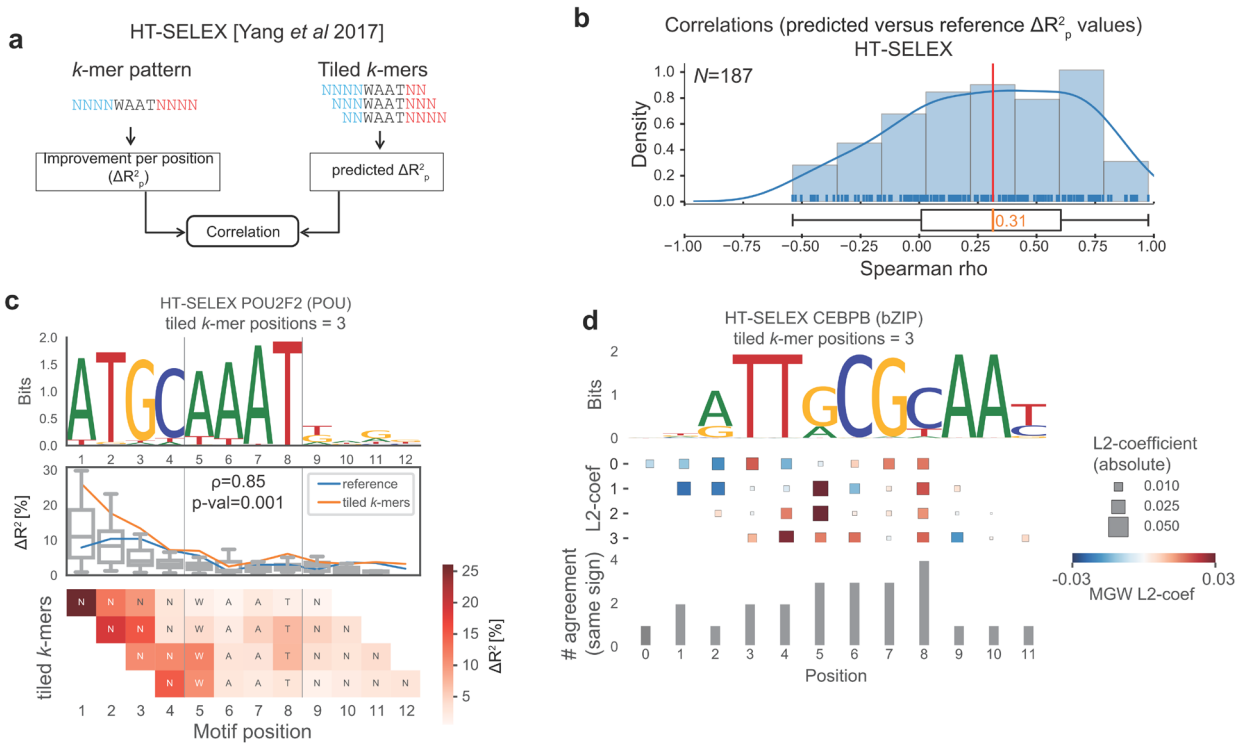

Figure S2.5. Benchmark of improvements per position consistency in HT-SELEX data.

Related to Figure 3 (a) tiled k-mers defined as shorter versions of the core motifs in HT-SELEX data analyses are used to study agreements observed in positional improvements by addition and removal of DNA-shape features. The estimated positional effect of all k-mers is compared with the effect observed with the model that uses the longest k-mer, using correlations (b) Distribution of correlations for all observed datasets (median Spearman correlation=0.31). (c) (upper) POU2F2 motif generated from top 100 k-mers by relative affinity (reference k-mer = NNNNWAATNNNN) (middle) Comparison between improvements per position generated using reference k-mer (blue line) and tiled k-mers (orange line) (Spearman correlation = 0.85) (bottom) Heatmap depicting ΔR^2_p values obtained for each tiled k-mer. Stronger biases are consistently obtained between positions one and five (d) Comparison of L2-coefficients for Minor Groove Width Features in CEBPB binding model obtained from HT-SELEX data, tiling 3 positions from reference k-mer. (top) CEBPB motif generated as in S4c. (middle) L2-coefficients for tiled k-mer aligned to that region of the reference k-mer. (bottom) Number of times coefficients are in agreement for the same sign in each position.

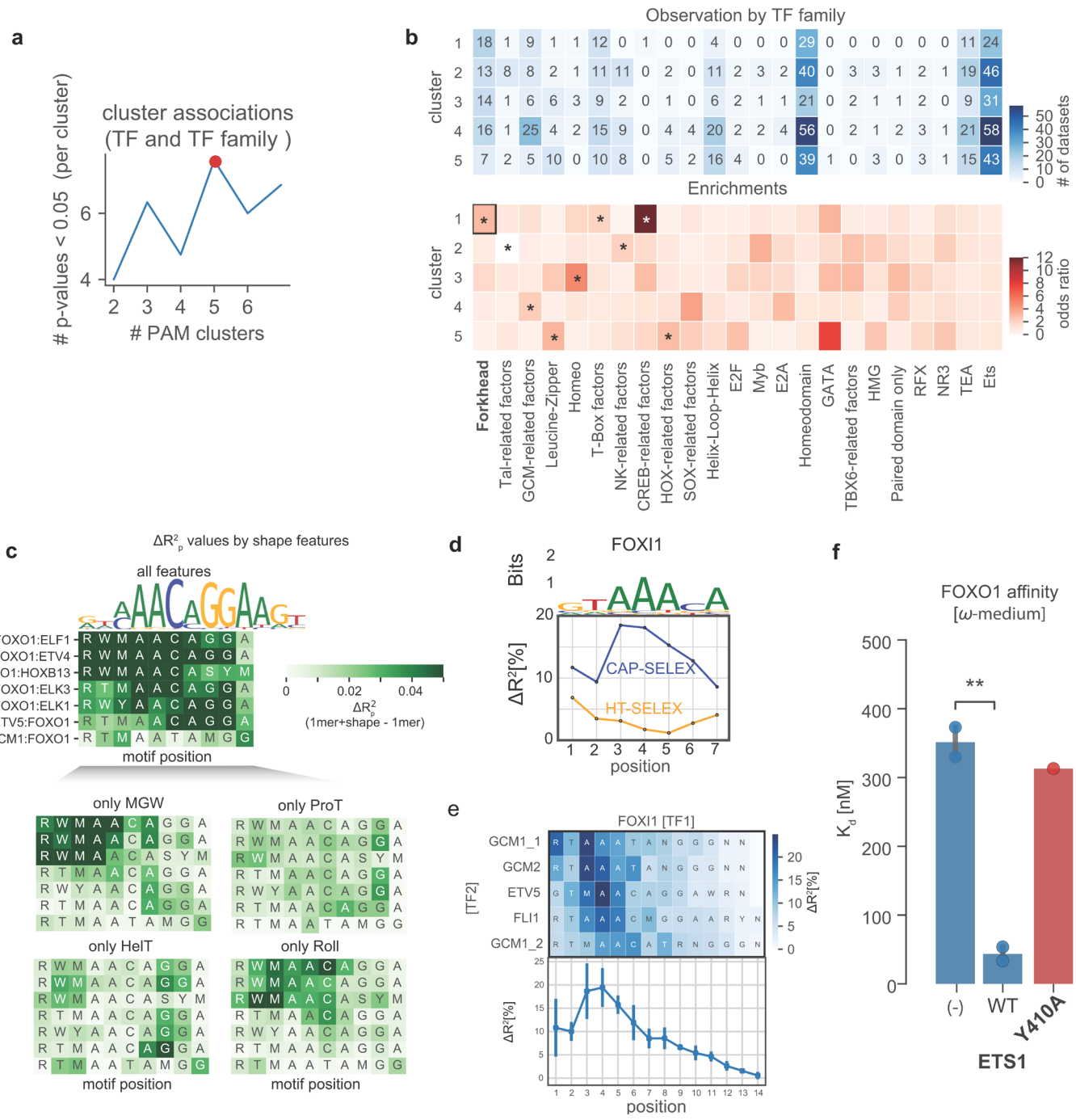

Figure S2.6. Positional improvements in CAP-SELEX data and Forkhead+Ets interactions

(**a**) Analysis of TF families enriched during PAM clustering of shape profiles for increasing number of clusters. Plateau is reached at five (three TFs families and two TFs. (**b**) Related to Fig 3b. (upper) Distribution of CAP-SELEX datasets by TF family (lower) enrichment values (* = p-value < 0.05. Black box = Adjusted $P < 0.1$) (**c**) Absolute changes in R^2 for CAP-SELEX models containing FOXO1, using individual DNA-shape properties. (**d**) *Related to Fig 2.3c.* (right) Comparison between relative R^2 FOXI1 in CAP-SELEX (blue) and values obtained from HT-SELEX data (orange line) (**e**) Related to Fig 2.3d. Reference *k*-mers of FOXI1 datasets aligned, and comparison of positional improvements after alignment (**f**) *Related to Fig 2.3f.* Dissociation constant measurements for FOXO1 to ω-medium alone (N=2) upon addition of ETS1 (N=2) or mutant Y410A (N=1).

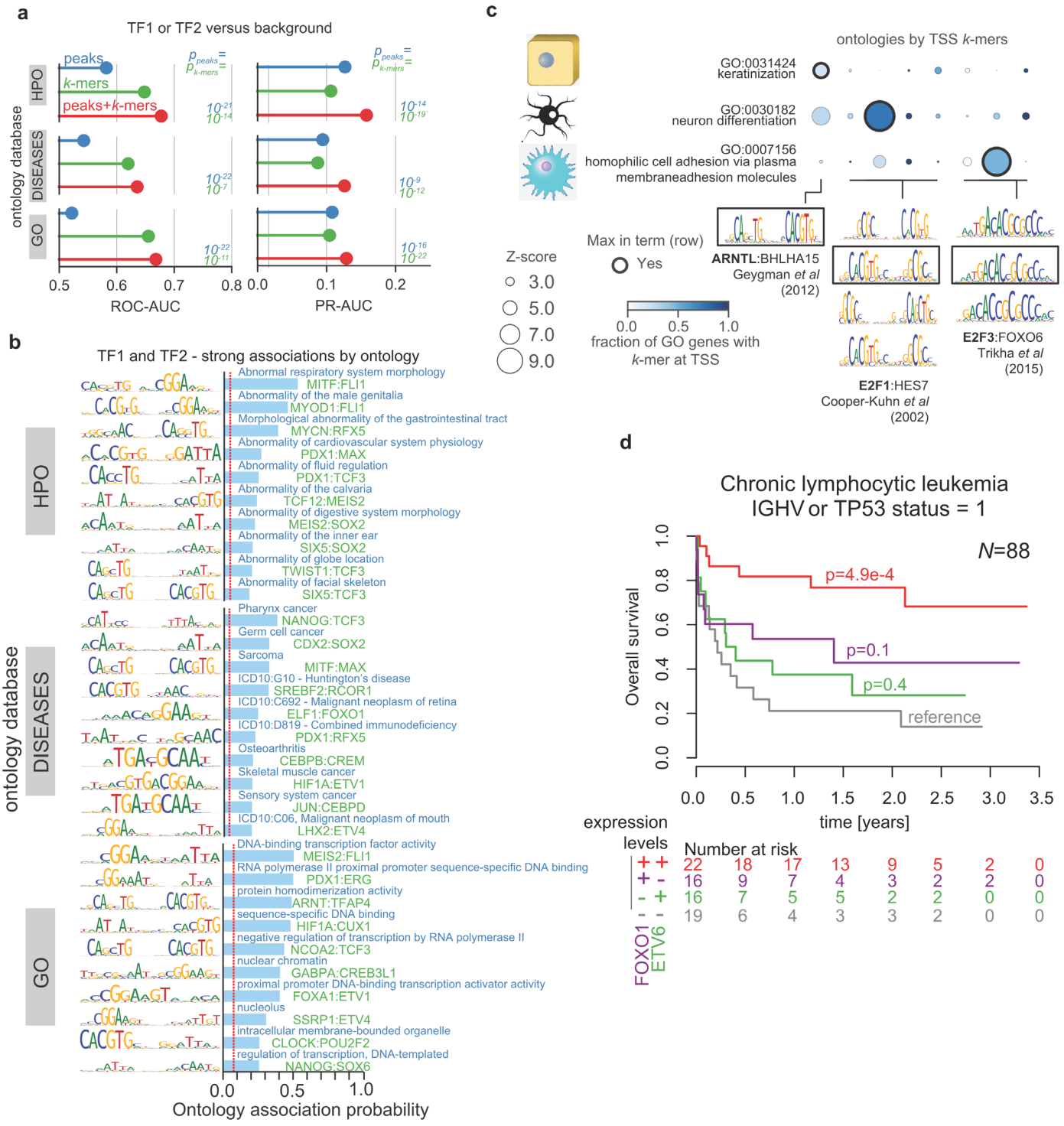

Figure S2.7. TF-TF associations to phenotypes using Ontology Association Probability.

Related to Figure 5 (**a**) "TF1 or TF2" benchmark results, when using OAPs to recover those associations with respect to background terms. Labels and p-values equivalent to Figure 5c. (**b**) Five strongest associations by ontology database using the category "TF1 and TF2" to assign TF pairs to ontologies. Red line indicates mean OAP for decoy terms (**c**) Related to Figure 6 (*top*) Ontologies related to differentiation and disease show a strong association to specific TF-TF pairs where at least one TF is known be related to that particular process. (bottom) Motifs related to each column are highlighted and grouped by their TF-TF names. When more than one topology exists, the motif with the highest score is highlighted in a black rectangle. (**d**) Kaplan-Meier plot of overall survival in CLL patients subsetted by IGHV and p53 mutation statuses (*N*=88). Data, *P* values and expression groups for FOXO1 and ETV6 are defined equivalently to Fig 6b.

Related to Chapter 3

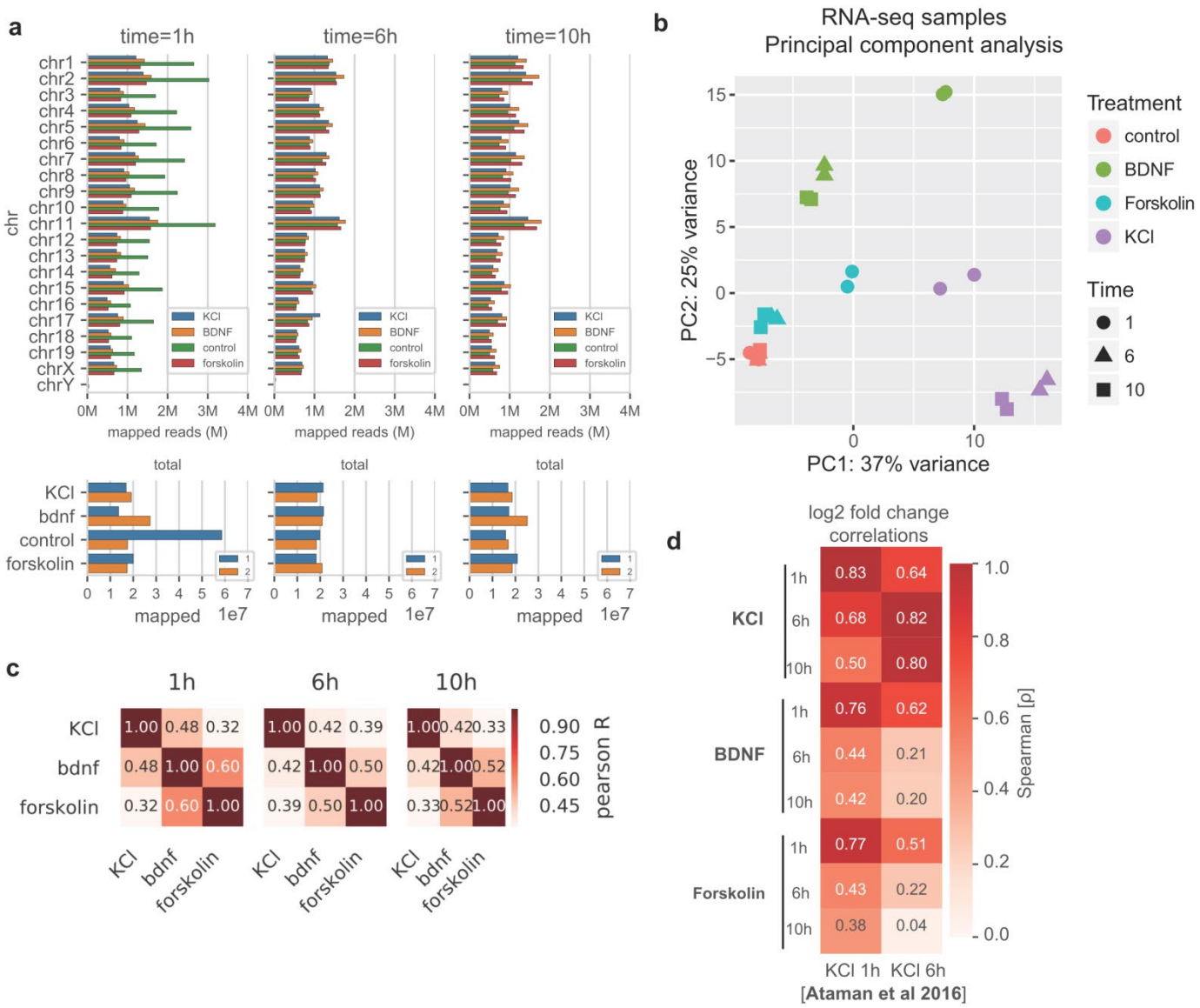

Figure S3.1 RNA-seq computational data processing

(**a**) (*top*) Mean of two replicates for mapped reads per chromosome in each time point and treatment combination (*bottom*) Total number of reads mapped in each biological replicate, in each condition and time point (**b**) Principal component analysis visualization for treatment and time points according to DESeq2 normalized counts. (**c**) Pearson correlation between mapped reads across treatments. (**d**) Correlation between log2 fold changes obtained in our analysis with matches DE-genes reported mouse neurons stimulated with KCl [**Ataman et al 2016**]

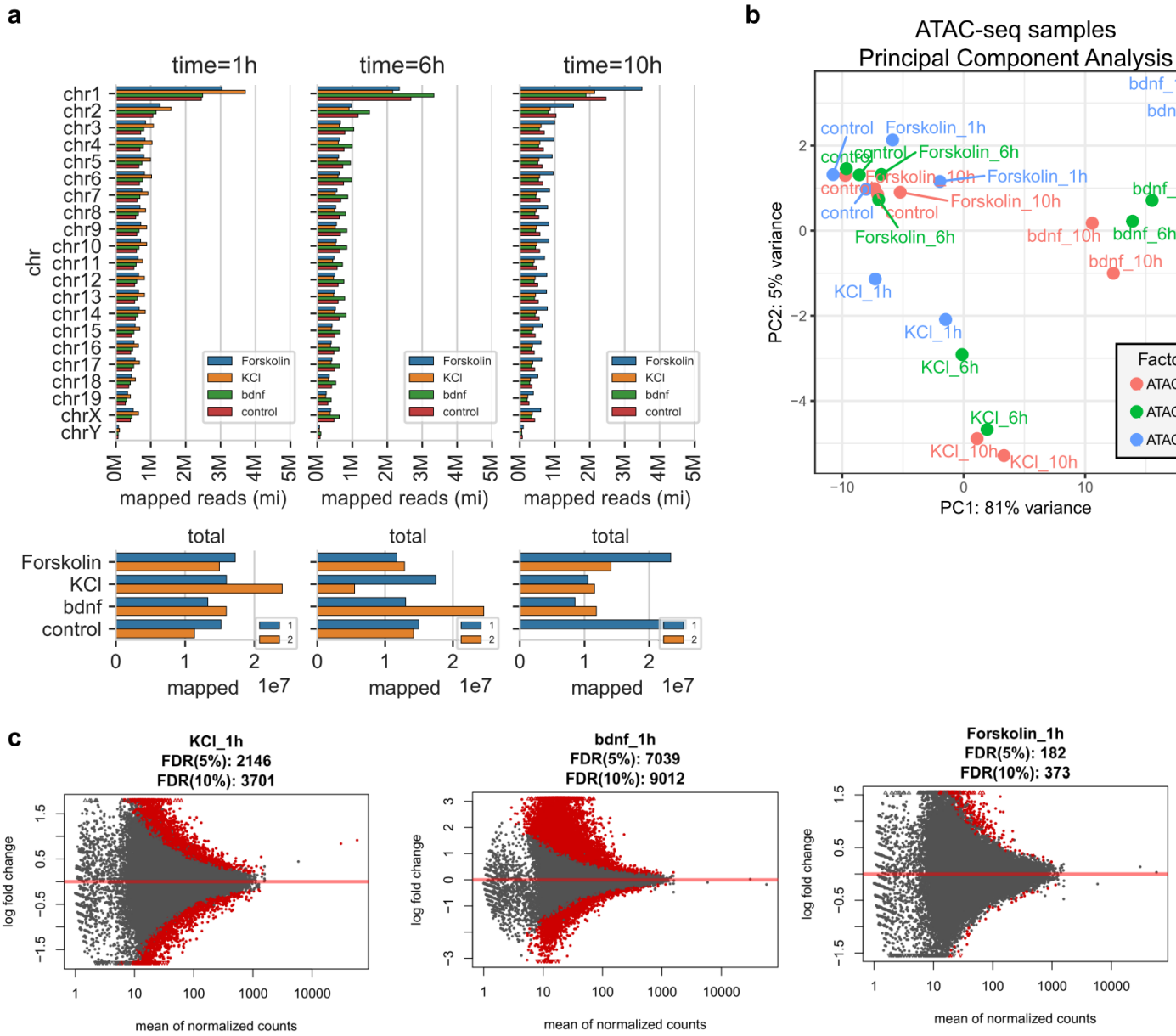

Figure S3.2 ATAC-seq computational data processing

(a) (*top*) Mean of two replicates for mapped reads per chromosome in each time point and treatment combination (*bottom*) Total number of reads mapped in each biological replicate, in each condition and time point **(b)** Principal component analysis visualization for treatment and time points using normalized counts generated with DiffBind **(c)** Differentially accessible peaks (DA-peaks) obtained when comparing 1h ATAC-seq samples versus controls for KCl (*left)*, BDNF (*middle*) and Forskolin (*right*). Red points indicate a peak with accessibility levels higher (log fold change > 0) or lower (log fold change < 0) than matched controls (FDR=10%).

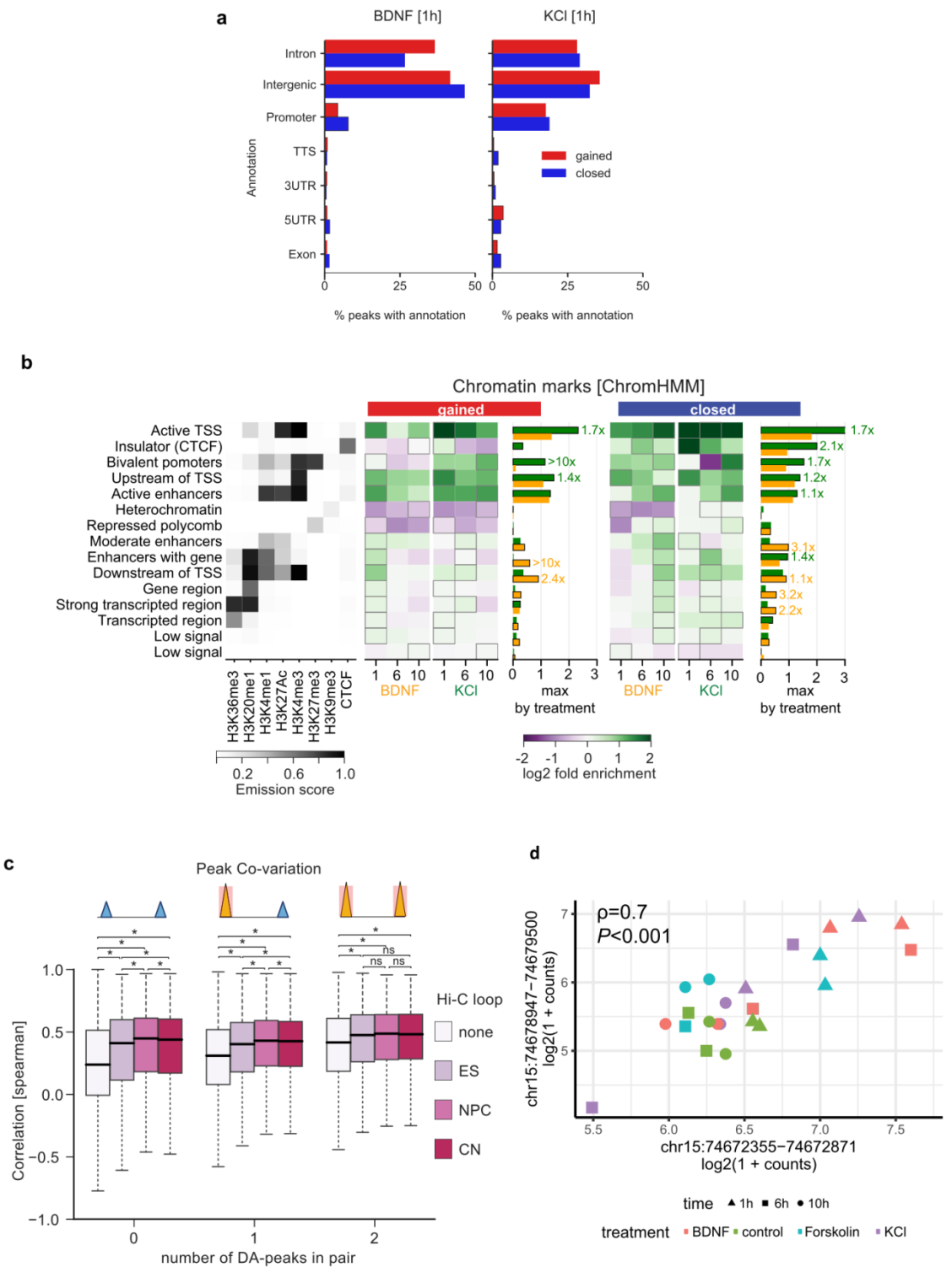

Figure S3.3 Genomic features of ATAC-seq peaks

Related to Figure 3.2c **(a)** Percentage distribution of DA-peaks for BDNF and KCl 1h in peak annotations (HOMER) **(b)** Chromatin states enrichments in DA-peaks for neuronal marks, based on ChromHMM 15-states model. Labels and descriptions as in Figure 3.2. **(c)** Correlation distributions for normalized counts of proximal peak pairs (CN=cortical neurons; NPC=neural progenitor cells; ES=Embryonic stem cells) (**Appendix A**). (**d**) Example of correlation between ATAC-seq peak in *Arc* gene and a DRE upstream of it

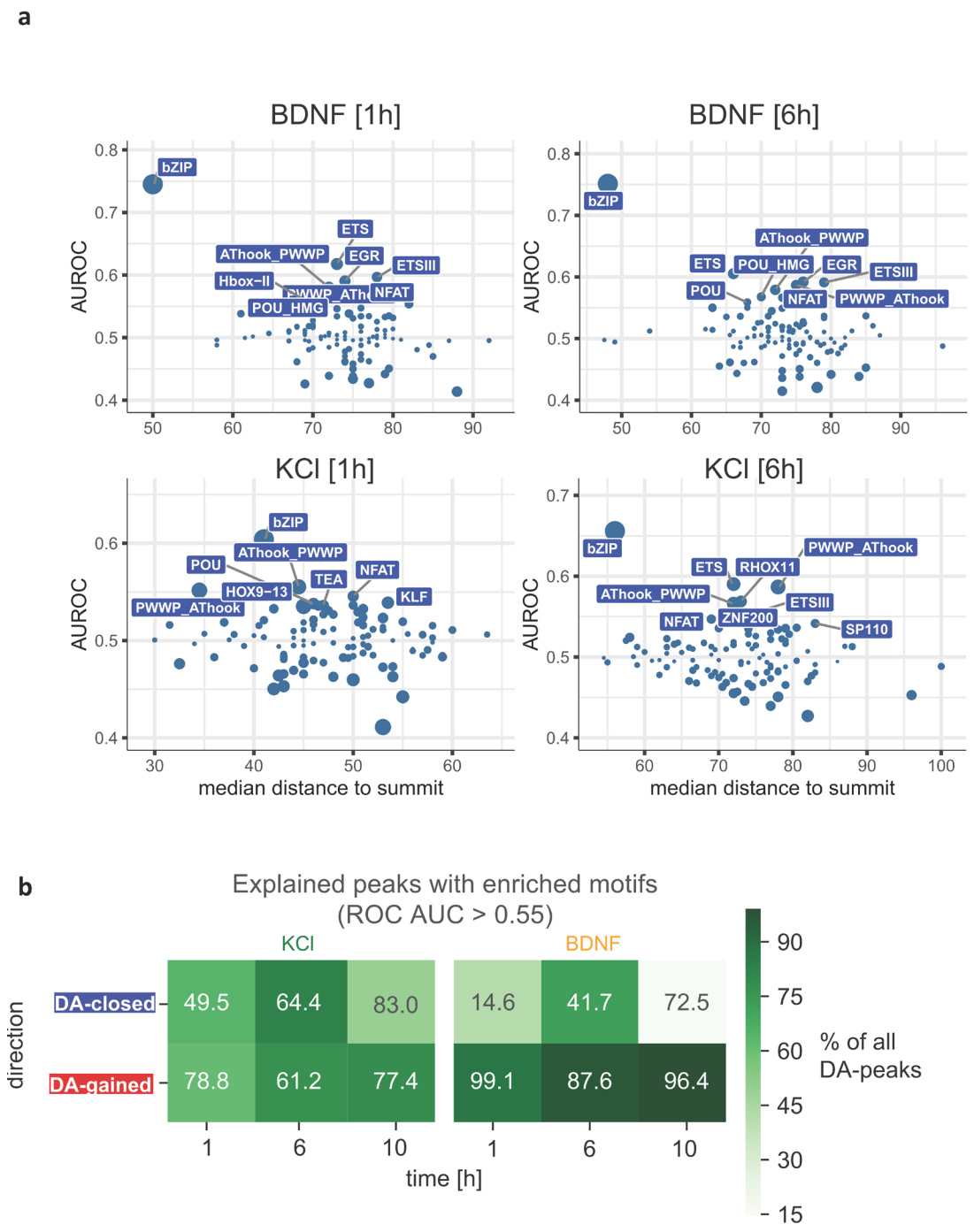

Figure S3.4 Features of *8*-mers for known TFs in ATAC-seq DA-peaks

(**a**) For DA-gained peaks, *8*-mers modules are visualized by its overall enrichment (Area Under the Receiver Operating Characteristic Curve, AUROC) versus the median relative distance to the peak summit for all observations. bZIP has the highest enrichment and additionally the lowest median distance to summit for BDNF 1 and 6h. (**b**) Percentage of DA-peaks explained by enriched TF motifs (ROC AUC > 0.55).

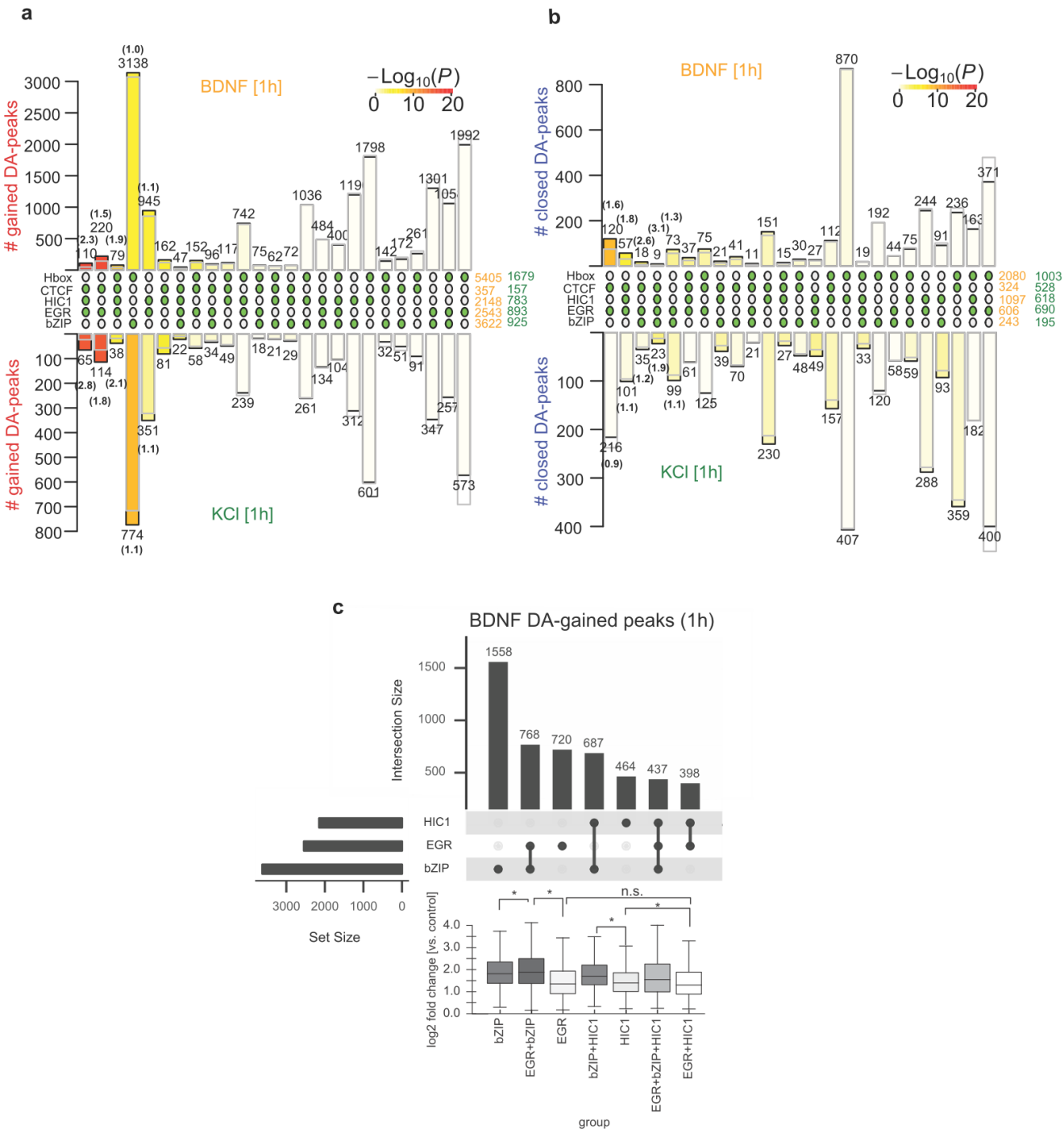

Figure S3.5. k-mers co-enrichments in BDNF vs KCl data

(a) Co-enrichments for selected TF specificity models in BDNF and KCl for gained DA-peaks **(b)** Similar to *(a)* for closed DA-closed peaks. KCl terms are sorted similar to BDNF for visual comparison. Fold Enrichments for five first terms are highlighted in parenthesis (*P* indicates p-values calculated hypergeometric [**Wang et al 2015**] (**Appendix A**). **(c)** Relationship between accessibility changes in ATAC-seq and presence of HIC1, EGR, and bZIP motifs in BDNF at 1h. Asterisks indicate $P < 0.1$ using one sided Wilcoxon's test, after Benjamini Hochberg procedure.

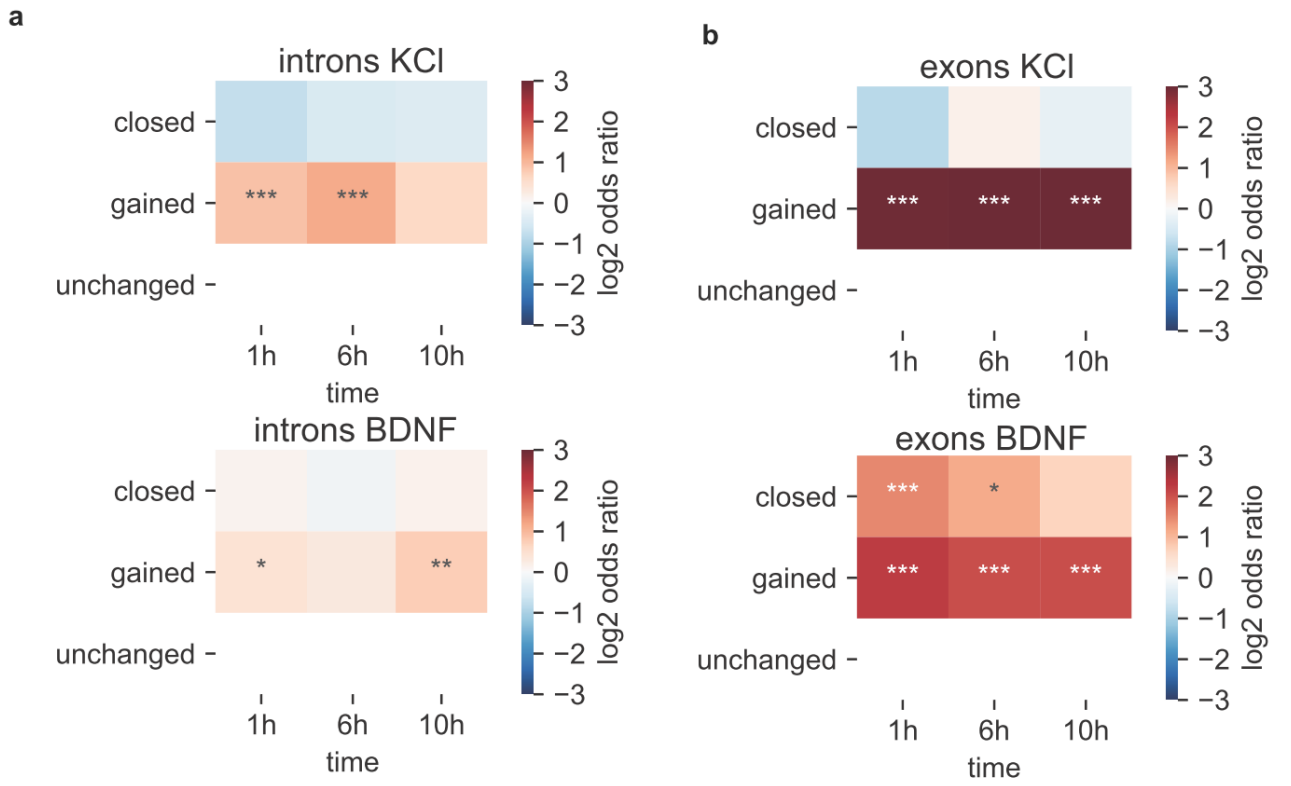

Figure S3.6. Enrichment of ATAC-seq peaks with CTCF motifs in introns and exons

(a) Enrichment scores for CTCF motifs in gained, closed and unchanged ATAC-seq associated to introns, and **(b)** exons. * indicates Fisher's one sided exact test, with *P* values adjusted by Benjamini Hochberg procedure *, **, *** = $P < 0.05$, 0.01 and 0.001, respectively (Fisher's exact test).

Related to Chapter 4

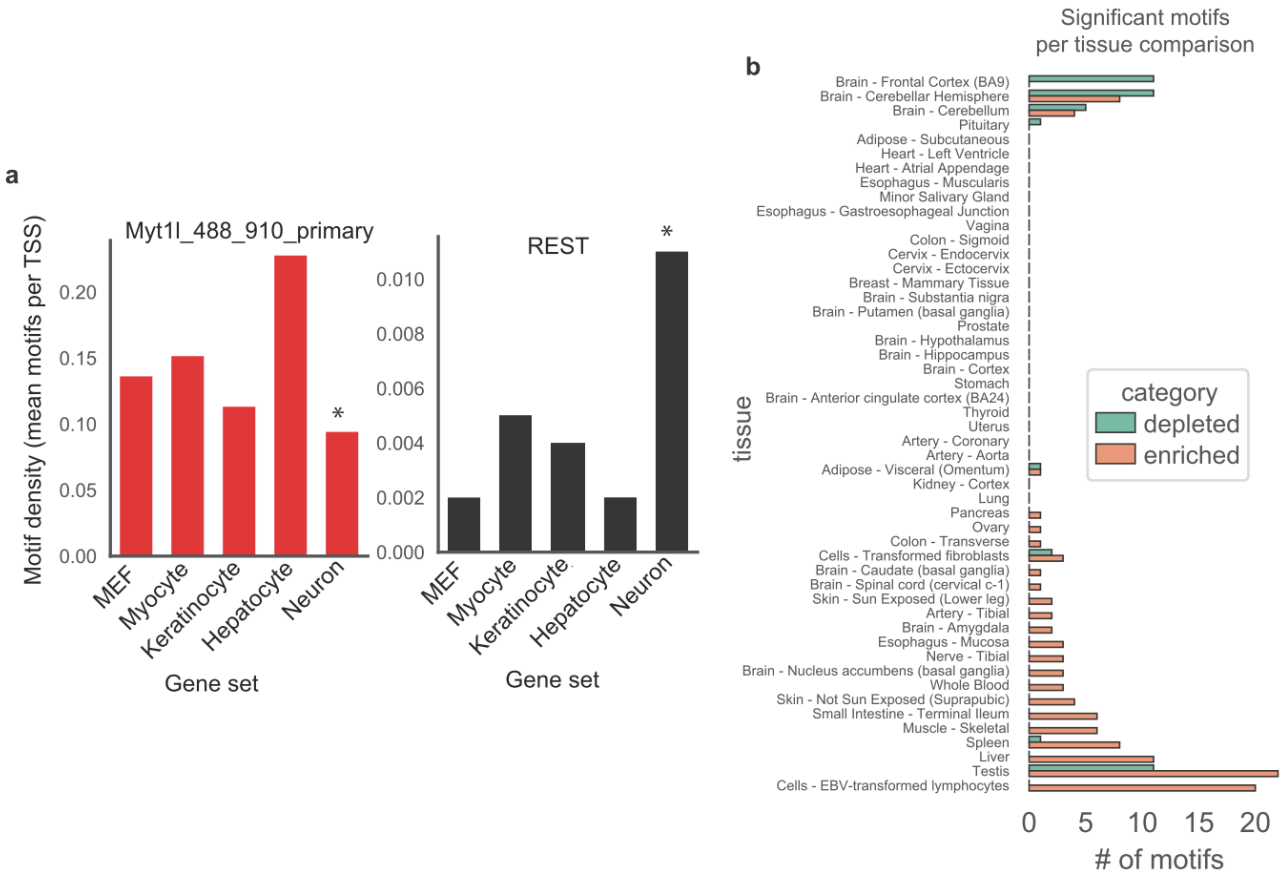

Figure S4.1 Enrichment and depletion of PWMs motifs in gene sets

(a) (*left*) Enrichment of Myt1l SELEX motif using top-1000 genes based on expression levels in Mouse Embryonic Fibroblasts, Myocytes, Keratinocytes, Hepatocytes and Neurons. (*right*) equivalent bar plot for REST motifs. * indicated $P < 0.05$ based on two-sided t-test between Neuron versus all other cell types. Adapted from **[Mall et al 2017] (b)** Number of significant PWMs reported as enriched or depleted based on comparison between signature genes in GTEx tissues versus all other tissues.

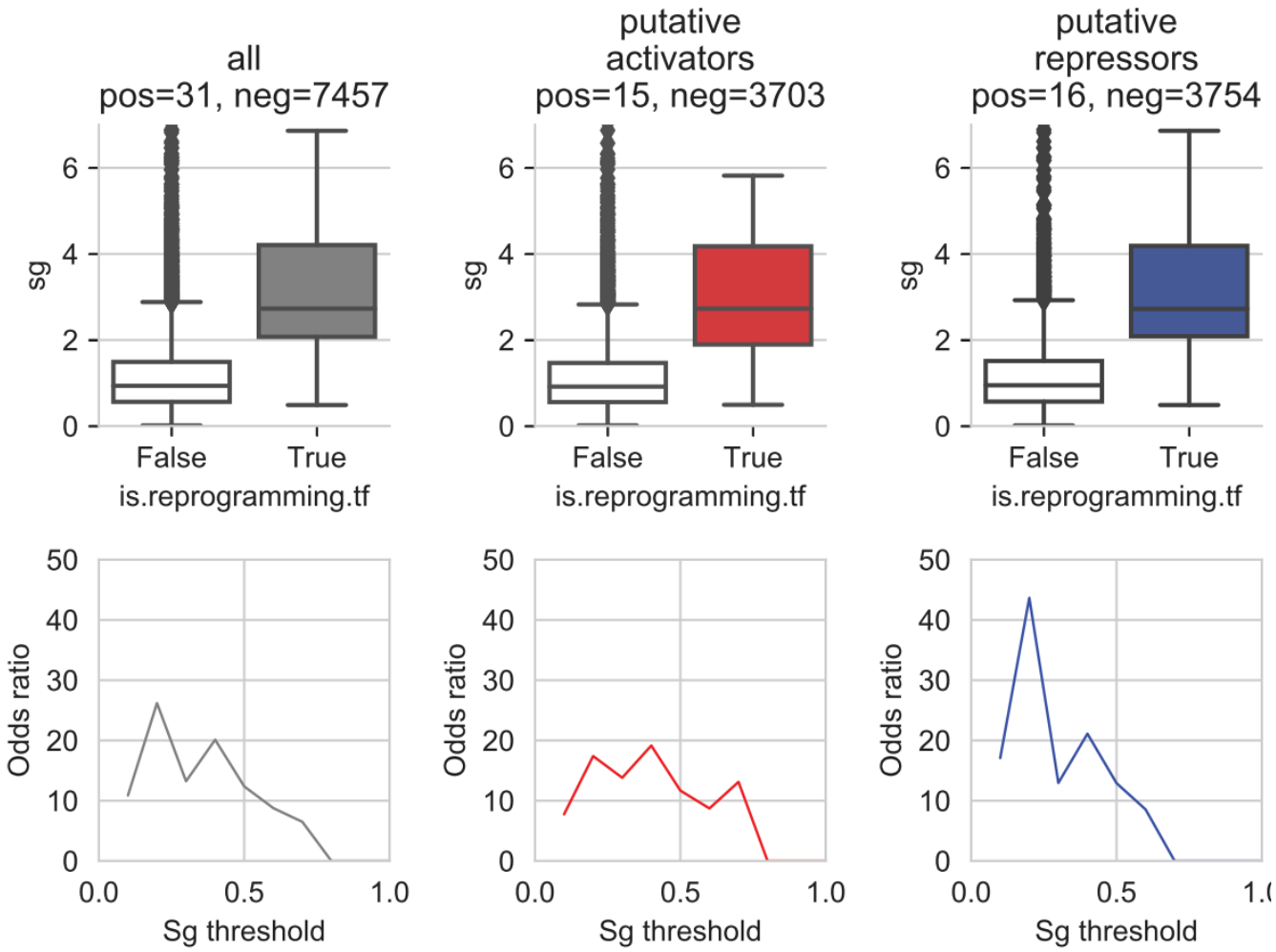

Figure S4.2 Enrichment and depletion of reprogramming TFs using Sg score

(*top*) Comparison between Sg scores obtained for known reprogramming TFs versus unknown cases, for (*left*) all, (*middle*) only putative activator TFs, and (*right*) only repressor TFs. (*bottom*) enrichment of reprogramming TFs using a moving threshold for Sg score (step=0.1). Maximum odds ratio is observed at 0.2, 0.4 and 0.2 for all TFs, only activator TFs and only repressor TFs, respectively.

Supplementary Table S4-I. Known reprogramming TFs with high S_g scores
(Score threshold defined as 0.4. Selected TFs for validation in bold)

Cell type	Gene name	Sg	Zm	Ze	predicted role
hepatocyte	**Prox1**	**2.53**	**-1.85**	**1.73**	**repressor**
oligodendrocyte	Olig2	4.16	-1.78	3.76	repressor
neuron	**Myt1l**	**5.16**	**-1.68**	**4.88**	**repressor**
large intestine goblet cell	Nr1i2	2.32	-1.31	1.91	repressor
neuron	Pou3f2	2.36	-1.19	2.04	repressor
cardiac muscle cell	Hand2	2.91	-1.00	2.74	repressor
cardiac muscle cell	**Tbx5**	**6.86**	**-0.79**	**6.81**	**repressor**
skeletal muscle satellite cell	Myod1	4.84	-0.61	4.80	repressor
Slamf1-positive multipotent progenitor cell	Gata2	2.00	-0.60	1.91	repressor
hepatocyte	Atf5	2.07	-0.48	2.02	repressor
oligodendrocyte	Sox10	4.27	-0.46	4.24	repressor
cardiac muscle cell	Gata4	3.06	-0.38	3.04	repressor
hepatocyte	Foxa3	2.94	-0.22	2.93	repressor
epithelial cell of large intestine	Foxa1	2.09	-0.16	2.09	repressor
oligodendrocyte	Runx2	0.49	-0.10	-0.48	repressor
epithelial cell of large intestine	Foxa3	1.89	-0.06	1.89	repressor
large intestine goblet cell	Cebpa	0.49	0.36	0.34	activator
large intestine goblet cell	Hnf4a	2.53	0.68	2.44	activator
type B pancreatic cell	Pdx1	5.82	0.73	5.77	activator
type B pancreatic cell	Mafa	7.61	0.77	7.57	activator
type B pancreatic cell	Neurog3	0.90	0.81	0.40	activator
epithelial cell of large intestine	Gata6	2.08	0.85	1.90	activator
epithelial cell of large intestine	Hnf4a	3.18	1.02	3.01	activator
epithelial cell of large intestine	Cdx2	4.94	1.04	4.83	activator
hepatocyte	Foxa2	1.71	1.41	0.98	activator
hepatocyte	Gata4	1.53	1.46	0.46	activator
hepatocyte	Hnf1a	3.06	1.79	2.48	activator
neuron	Ascl1	2.73	1.92	1.94	activator
cardiac muscle cell	Mef2c	2.37	2.34	0.37	activator
hepatocyte	Hnf4a	4.25	2.47	3.46	activator
Slamf1-positive multipotent progenitor cell	Erg	4.10	2.99	2.81	activator

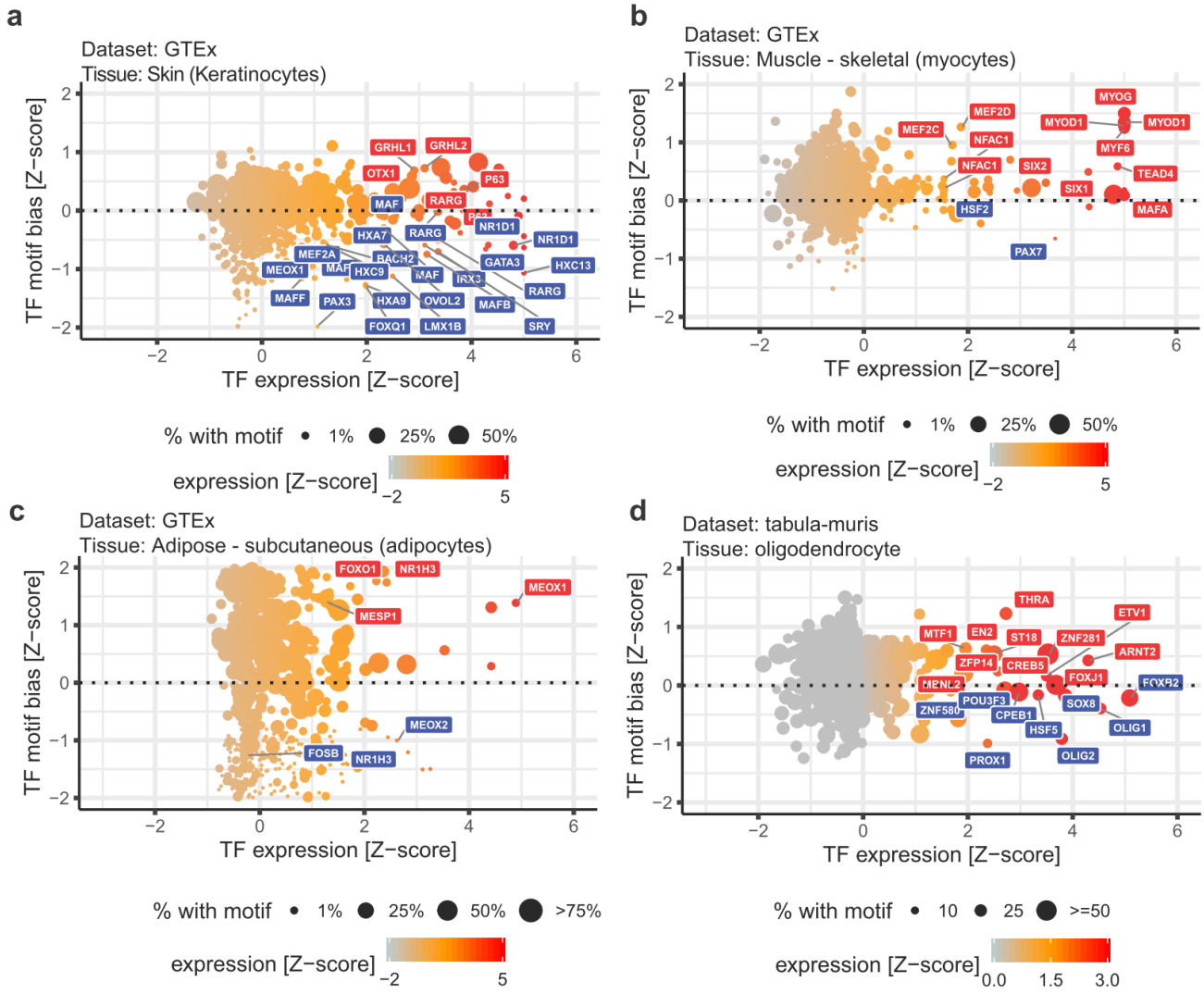

Figure S4.3 Expression and motif biases classify TF activators and putative terminal repressors

(a) Skin (keratinocytes) expression and motif biases highlight several TF activators and repressors. Among those, FOXQ1 has been validated to improve differentiation from MEFs to induced Keratinocytes [**Rackham et al 2016**]. **(b)** Skeletal muscle (myocytes) analysis highlights Myod1 as a known transcriptional activator of muscle genes [**Tapscott et al 2005**], is highlighted in the upper-right region. **(c)** Adipose tissue (adipocytes). Highlights Meox1 is predicted as a TF activator in adipocytes. **(d)** Oligodendrocytes analysis predicts Arnt2 as a strong activator, and Olig1/2 as terminal repressors.

Bibliography

1. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).

2. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* (2015). doi:10.1038/nature15393

3. AbdulMajeed, A. A., Dalley, A. J. & Farah, C. S. Loss of ELF3 immunoexpression is useful for detecting oral squamous cell carcinoma but not for distinguishing between grades of epithelial dysplasia. *Ann. Diagn. Pathol.* **17,** 331–340 (2013).

4. Abe, N. *et al.* Deconvolving the Recognition of DNA Shape from Sequence. *Cell* **161,** 307–318 (2015).

5. Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22,** 1600–1607 (2006).

6. Ando, M. *et al.* Mutational Landscape and Antiproliferative Functions of ELF Transcription Factors in Human Cancer. *Cancer Res.* **76,** 1814–1824 (2016).

7. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

8. Ataman, B. *et al.* Evolution of Osteocrin as an activity-regulated factor in the primate brain. *Nature* **539,** 242–247 (2016).

9. Bading, H., Ginty, D. D. & Greenberg, M. E. Regulation of gene expression in hippocampal neurons by distinct calcium signaling pathways. *Science* **260,** 181–6 (1993).

10. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science (80-.).* **324,** 1720–1723 (2009).

11. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).

12. Barrera, L. A. *et al.* Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science (80-.).* **351,** 1450–1454 (2016).

13. Beckmann, A. M. & Wilce, P. A. Egr transcription factors in the nervous system. *Neurochem. Int.* **31,** 477-510; discussion 517–6 (1997).

14. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300 (1995).

15. Berger, M. F. *et al.* Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell* **133,** 1266–1276 (2008).

16. Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* **4,** 393–411 (2009).

17. Biddie, S. C. *et al.* Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. *Mol. Cell* **43,** 145–155 (2011).

18. Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.* **39,** W68–W73 (2011).

19. Blanchet, P. *et al.* MYT1L mutations cause intellectual disability and variable obesity by dysregulating gene expression and development of the neuroendocrine hypothalamus. *PLoS Genet.* **13,** e1006957 (2017).

20. Bloch, G. & Grozinger, C. M. Social molecular pathways and the evolution of bee societies. *Philos. Trans. R. Soc. B Biol. Sci.* **366,** 2155 (2011).

21. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171,** 557–572.e24 (2017).

22. Boulay, G. *et al.* Hypermethylated in Cancer 1 (HIC1) Recruits Polycomb Repressive Complex 2 (PRC2) to a Subset of Its Target Genes through Interaction with Human Polycomb-like (hPCL) Proteins. *J. Biol. Chem.* **287,** 10509–10524 (2012).

23. Chao, M. V & Hempstead, B. L. p75 and Trk: a two-receptor system. *Trends Neurosci.* **18,** 321–6 (1995).

24. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. (2016). doi:10.1145/2939672.2939785

25. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx1092

26. Cheng, Y. *et al.* Neurotrophic factor-α1 prevents stress-induced depression through enhancement of neurogenesis and is activated by rosiglitazone. *Mol. Psychiatry* **20,** 744–54 (2015).

27. Cheng, Y., Cawley, N. X. & Loh, Y. P. Carboxypeptidase E (NF-α1): a new trophic factor in neuroprotection. *Neurosci. Bull.* **30,** 692–696 (2014).

28. Choy, W. W., Datta, D., Geiger, C. A., Birrane, G. & Grant, M. A. Crystallization and preliminary X-ray analysis of a complex of the FOXO1 and Ets1 DNA-binding domains and DNA. *Acta Crystallogr. Sect. FStructural Biol. Commun.* **70,** 44–48 (2014).

29. Cohen, N. M. *et al.* SHAMAN: bin-free randomization, normalization and screening of Hi-C matrices. *bioRxiv* (2017). doi:10.1101/187203

30. Crocker, J., Noon, E. P. Ben & Stern, D. L. *The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution*. *Current Topics in Developmental Biology* (Elsevier Inc., 2015). doi:10.1016/bs.ctdb.2015.11.018

31. de la Torre-Ubieta, L. *et al.* The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* **172,** 289–304.e18 (2018).

32. de Mendoza, A. *et al.* Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E4858-66 (2013).

33. Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).

34. Dietrich, S. *et al.* Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* **128,** 427–445 (2017).

35. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* **25,** 1268–1280 (2015).

36. Dudas, J. *et al.* Altered regulation of Prox1-gene-expression in liver tumors. *BMC Cancer* **8,** 92 (2008).

37. Feng, R. *et al.* PU.1 and C/EBPα/β convert fibroblasts into macrophage-like cells. *Proc. Natl. Acad. Sci.* **105,** 6057–6062 (2008).

38. Flavell, S. W. & Greenberg, M. E. Signaling Mechanisms Linking Neuronal Activity to Gene Expression and Plasticity of the Nervous System. *Annu. Rev. Neurosci.* **31,** 563–590 (2008).

39. Fu, X., He, F., Li, Y., Shahveranov, A. & Hutchins, A. P. Genomic and molecular control of cell type and cell type conversions. *Cell Regen. (London, England)* **6,** 1–7 (2017).

40. Fujita, Y. *et al.* Tomosyn: a syntaxin-1-binding protein that forms a novel complex in the neurotransmitter release process. *Neuron* **20,** 905–15 (1998).

41. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* **13,** 840–52 (2012).

42. Gallegos, D. A., Chan, U., Chen, L.-F. & West, A. E. Chromatin Regulation of Neuronal Maturation and Plasticity. *Trends Neurosci.* **41,** 311–324 (2018).

43. Geerts, C. J. *et al.* Tomosyn associates with secretory vesicles in neurons through its N- and C-terminal domains. *PLoS One* **12,** e0180912 (2017).

44. Gheorghe, M. *et al.* A map of direct TF–DNA interactions in the human genome. *Nucleic Acids Res.* **47,** e21–e21 (2019).

45. Goebbels, S. *et al.* Genetic targeting of principal neurons in neocortex and hippocampus of NEX-Cre mice. *genesis* **44,** 611–621 (2006).

46. Greer, P. L. & Greenberg, M. E. From Synapse to Nucleus: Calcium-Dependent Gene Transcription in the Control of Synapse Development and Function. *Neuron* **59,** 846–860 (2008).

47. GTEx Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).

48. Guerrero-Ramirez, G., Valdez-Cordoba, C., Islas-Cisneros, J. & Trevino, V. Computational approaches for predicting key transcription factors in targeted cell reprogramming (Review). *Mol. Med. Rep.* **18,** 1225–1237 (2018).

49. Guo, Y., Tian, K., Zeng, H., Guo, X. & Gifford, D. K. A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Res.* **28,** 891 (2018).

50. Guturu, H., Doxey, A. C., Wenger, A. M. & Bejerano, G. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20130029 (2013).

51. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46,** D380–D386 (2018).

52. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).

53. Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9,** 20–32 (2018).

54. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39,** 311–318 (2007).

55. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).

56. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* (2006). doi:10.1093/nar/gkj144

57. Holmberg, J. & Perlmann, T. Maintaining differentiated cellular identity. *Nat. Rev. Genet.* **13,** 429–439 (2012).

58. Hörmanseder, E. *et al.* H3K4 Methylation-Dependent Memory of Somatic Cell Identity Inhibits Reprogramming and Development of Nuclear Transfer Embryos. *Cell Stem Cell* **21,** 135–143.e6 (2017).

59. Huang, Y.-H., Jankowski, A., Cheah, K. S. E., Prabhakar, S. & Jauch, R. SOXE transcription factors form selective dimers on non-compact DNA motifs through

multifaceted interactions between dimerization and high-mobility group domains. *Sci. Rep.* **5,** 10398 (2015).

60. Ieda, M. *et al.* Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. *Cell* **142,** 375–386 (2010).

61. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43,** 110–119 (2017).

62. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43,** 110–119 (2017).

63. Isakova, A., Berset, Y., Hatzimanikatis, V. & Deplancke, B. Quantification of cooperativity in heterodimer-DNA binding improves the accuracy of binding specificity models. *J. Biol. Chem.* **291,** 10293–10306 (2016).

64. Iwata, A. *et al.* Quality of TCR signaling determined by differential affinities of enhancers for the composite BATF–IRF4 transcription factor complex. *Nat. Immunol.* **18,** 563–572 (2017).

65. Jain, D., Baldi, S., Zabel, A., Straub, T. & Becker, P. B. Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.* **43,** 6959–6968 (2015).

66. Jankowski, A., Prabhakar, S. & Tiuryn, J. TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics* **15,** 208 (2014).

67. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20,** 861–73 (2010).

68. Jolma, A. & Taipale, J. in *Sub-cellular biochemistry* **52,** 155–173 (2011).

69. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152,** 327–339 (2013).

70. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527,** 384–388 (2015).

71. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527,** 384–8 (2015).

72. Jolma, A. *et al.* Binding specificities of human RNA binding proteins towards structured and linear RNA sequences. *bioRxiv* 317909 (2019). doi:10.1101/317909

73. Jugnia, L.-B., Manno, D., Dodard, S., Greer, C. W. & Hendry, M. Manipulating redox conditions to enhance in situ bioremediation of RDX in groundwater at a contaminated site. *Sci. Total Environ.* **676,** 368–377 (2019).

74. Junion, G. *et al.* A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell* **148,** 473–486 (2012).

75. Kandel, E. R. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. *Mol. Brain* **5,** 14 (2012).

76. Katrancha, S. M. *et al.* Trio Haploinsufficiency Causes Neurodevelopmental Disease-Associated Deficits. *Cell Rep.* **26,** 2805–2817.e9 (2019).

77. Katrancha, S. M. *et al.* Trio Haploinsufficiency Causes Neurodevelopmental Disease-Associated Deficits. *Cell Rep.* **26,** 2805–2817.e9 (2019).

78. Kinney, J. B. & McCandlish, D. M. Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annu. Rev. Genomics Hum. Genet.* **20,** annurev-genom-083118-014845 (2019).

79. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47,** D1018–D1027 (2019).

80. Kribelbauer, J. F. *et al.* Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Rep.* **19,** 2383–2395 (2017).

81. Kundaje, A. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* (2015). doi:10.1016/j.cell.2015.07.048

82. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172,** 650–665 (2018).

83. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinforma.* (2010). doi:10.1002/0471250953.bi1107s32

84. Lazebnik, Y. Can a biologist fix a radio?--Or, what I learned while studying apoptosis. *Cancer Cell* **2,** 179–82 (2002).

85. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31,** 1325–1327 (2015).

86. Lenz, G. *et al.* Stromal Gene Signatures in Large-B-Cell Lymphomas. *N. Engl. J. Med.* **359,** 2313–2323 (2008).

87. Li, C. X. & Poznansky, M. J. Characterization of the ZO-1 protein in endothelial and other cell lines. *J. Cell Sci.* **97,** (1990).

88. Li, J. *et al.* Structure of the Forkhead Domain of FOXA2 Bound to a Complete DNA Consensus Site. *Biochemistry* **56,** 3745–3753 (2017).

89. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science (80-.).* **352,** 600–604 (2016).

90. Lindskog, C. *et al.* The human cardiac and skeletal muscle proteomes defined by transcriptomics and antibody-based profiling. *BMC Genomics* **16,** 475 (2015).

91. Lou, H. *et al.* Sorting and Activity-Dependent Secretion of BDNF Require Interaction of a Specific Motif with the Sorting Receptor Carboxypeptidase E. *Neuron* **45,** 245–255 (2005).

92. Luna-Zurita, L. *et al.* Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell* **164,** 999–1014 (2016).

93. Macías, W., Carlson, R., Rajadhyaksha, A., Barczak, A. & Konradi, C. Potassium chloride depolarization mediates CREB phosphorylation in striatal neurons in an NMDA receptor-dependent manner. *Brain Res.* **890,** 222–32 (2001).

94. Malik, A. N. *et al.* Genome-wide identification and characterization of functional neuronal activity–dependent enhancers. *Nat. Neurosci.* **17,** 1330–1339 (2014).

95. Mall, M. *et al.* Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. *Nature* **544,** 245–249 (2017).

96. Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A. & Bulyk, M. L. Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. *Cell Syst.* **5,** 187–201.e7 (2017).

97. Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A. & Bulyk, M. L. Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. *Cell Syst.* **5,** 187–201.e7 (2017).

98. Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A. & Bulyk, M. L. Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. *Cell Syst.* **5,** 187–201.e7 (2017).

99. Mathelier, A. *et al.* DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst.* **3,** 278–286 (2016).

100. Mathelier, A. *et al.* JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** 142–147 (2014).

101. Mayran, A. & Drouin, J. Pioneer transcription factors shape the epigenetic landscape. *J. Biol. Chem.* **293,** 13795–13804 (2018).

102. Mayran, A. *et al.* Pioneer and nonpioneer cooperation drives lineage specific chromatin opening. *bioRxiv* 472647 (2019). doi:10.1101/472647

103. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).

104. Mohindru, B. *et al.* Health State Utility Data in Cystic Fibrosis: A Systematic Review. *PharmacoEconomics - open* (2019). doi:10.1007/s41669-019-0144-1

105. Monahan, K. *et al.* Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *Elife* **6,** (2017).

106. Monahan, K. *et al.* Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *Elife* **6,** (2017).

107. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47,** 1–8 (2017).

108. Narasimhan, K. *et al.* DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Res.* **43,** 1513–1528 (2015).

109. Ng, R. K. & Gurdon, J. B. Epigenetic memory of an active gene state depends on histone H3.3 incorporation into chromatin in the absence of transcription. *Nat. Cell Biol.* **10,** 102–109 (2008).

110. Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4,** (2015).

111. O'Rourke, N. A., Weiler, N. C., Micheva, K. D. & Smith, S. J. Deep molecular diversity of mammalian synapses: why it matters and how to measure it. *Nat. Rev. Neurosci.* **13,** 365–379 (2012).

112. Paredes, S. H., Melgar, M. F. & Sethupathy, P. Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index. *Bioinformatics* **29,** 1485–1487 (2013).

113. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

114. Peixoto, L. & Abel, T. The Role of Histone Acetylation in Memory Formation and Cognitive Impairments. *Neuropsychopharmacology* **38,** 62 (2013).

115. Peker, D., Quigley, B., Qin, D., Papenhausen, P. & Zhang, L. Burkitt Lymphoma Arising From Lymphoplasmacytic Lymphoma Following Acquisition of *MYC* Translocation and Loss of the *ETV6* Tumor Suppressor Gene. *Arch. Pathol. Lab. Med.* **137,** 130–133 (2013).

116. Pengelly, R. J. *et al.* Mutations specific to the Rac-GEF domain of *TRIO* cause intellectual disability and microcephaly. *J. Med. Genet.* **53,** 735–742 (2016).

117. Pengelly, R. J. *et al.* Mutations specific to the Rac-GEF domain of *TRIO* cause intellectual disability and microcephaly. *J. Med. Genet.* **53,** 735–742 (2016).

118. Pinte, S. *et al.* The Tumor Suppressor Gene *HIC1* (*Hypermethylated in Cancer 1*) Is a Sequence-specific Transcriptional Repressor. *J. Biol. Chem.* **279,** 38313–38324 (2004).

119. Plath, N. *et al.* Arc/Arg3.1 Is Essential for the Consolidation of Synaptic Plasticity and Memories. *Neuron* **52,** 437–444 (2006).

120. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: Text mining and data integration of disease–gene associations. *Methods* **74,** 83–89 (2015).

121. Poo, M. Neurotrophins as synaptic modulators. *Nat. Rev. Neurosci.* **2,** 24–32 (2001).

122. Rastogi, C. *et al.* Accurate and sensitive quantification of protein-DNA binding affinity. *Proc. Natl. Acad. Sci.* **115,** E3692–E3701 (2018).

123. Ray, D. *et al.* RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. *Methods* **118–119,** 3–15 (2017).

124. Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347,** 1155–9 (2015).

125. Reményi, A., Schöler, H. R. & Wilmanns, M. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.* **11,** 812–815 (2004).

126. Ren, G. *et al.* CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol. Cell* **67,** 1049–1058.e6 (2017).

127. Ribeiro, J., Melo, F. & Schüller, A. PDIviz: analysis and visualization of protein–DNA binding interfaces: Fig. 1. *Bioinformatics* **31,** 2751–2753 (2015).

128. Riley, T. R. *et al.* in *Methods in molecular biology (Clifton, N.J.)* **1196,** 255–278 (2014).

129. Rogers, J. M. *et al.* Bispecific Forkhead Transcription Factor FoxN3 Recognizes Two Distinct Motifs with Different DNA Shapes. *Mol. Cell* (2019). doi:10.1016/j.molcel.2019.01.019

130. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79,** 233–69 (2010).

131. Rosa, F. F. *et al.* Direct reprogramming of fibroblasts into antigen-presenting dendritic cells. *Sci. Immunol.* **3,** eaau4292 (2018).

132. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* (2012). doi:10.1038/nature10730

133. Ruan, S. & Stormo, G. D. Inherent limitations of probabilistic models for protein-DNA binding specificity. *PLoS Comput. Biol.* **13,** e1005638 (2017).

134. Ruiz-Velasco, M. *et al.* CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Syst.* **5,** 628–637.e6 (2017).

135. Sadybekov, A., Tian, C., Arnesano, C., Katritch, V. & Herring, B. E. An autism spectrum disorder-related de novo mutation hotspot discovered in the GEF1 domain of Trio. *Nat. Commun.* **8,** 601 (2017).

136. Sams, D. S. *et al.* Neuronal CTCF Is Necessary for Basal and Experience-Dependent Gene Regulation, Memory Formation, and Genomic Structure of BDNF and Arc. *Cell Rep.* **17,** 2418–2430 (2016).

137. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res.* **47,** D212–D220 (2019).

138. Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34,** 93–158 (1999).

139. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* (2017). doi:10.1038/nmeth.4401

140. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (2012).

141. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18,** 6097–100 (1990).

142. Seamon, K. B. & Daly, J. W. Forskolin: a unique diterpene activator of cyclic AMP-generating systems. *J. Cyclic Nucleotide Res.* **7,** 201–24 (1981).

143. Shi, W., Fornes, O., Mathelier, A. & Wasserman, W. W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* **44,** 10106–10116 (2016).

144. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479,** 74–79 (2011).

145. Singh, S. *et al.* Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition. *Elife* **5,** (2016).

146. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell* **147,** 1270–1282 (2011).

147. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13,** 613–626 (2012).

148. Stadhouders, R. *et al.* Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J.* **31,** 986–99 (2012).

149. Stengel, A. *et al.* Detection of recurrent and of novel fusion transcripts in myeloid malignancies by targeted RNA sequencing. *Leukemia* **32,** 1229–1238 (2018).

150. Stormo, G. D., Schneider, T. D. & Gold, L. M. Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10,** 2971–2996 (1982).

151. Su, Y. *et al.* Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nat. Neurosci.* **20,** 476–483 (2017).

152. Su, Y. *et al.* Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nat. Neurosci.* **20,** 476–483 (2017).

153. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126,** 663–676 (2006).

154. Tapscott, S. J. The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. *Development* **132,** 2685–2695 (2005).

155. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci.* **110,** 18602–18607 (2013).

156. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47,** D330–D338 (2019).

157. Tyssowski, K. M. *et al.* Different Neuronal Activity Patterns Induce Different Gene Expression Programs. *Neuron* **98,** 530–546.e11 (2018).

158. Tzingounis, A. V. & Nicoll, R. A. Arc/Arg3.1: Linking Gene Expression to Synaptic Plasticity and Memory. *Neuron* **52,** 403–407 (2006).

159. Ubaid Ullah *et al.* Transcriptional Repressor HIC1 Contributes to Suppressive Function of Human Induced Regulatory T Cells. *Cell Rep.* **22,** 2094–2106 (2018).

160. van Bömmel, A., Love, M. I., Chung, H.-R. & Vingron, M. coTRaCTE predicts co-occurring transcription factors within cell-type specific enhancers. *PLOS Comput. Biol.* **14,** e1006372 (2018).

161. van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation. *Microbiol. Mol. Biol. Rev.* **73,** 481–509 (2009).

162. Vandel, J., Cassan, O., Lèbre, S., Lecellier, C.-H. & Bréhélin, L. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC Genomics* **20,** 103 (2019).

163. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–263 (2009).

164. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463,** 1035–41 (2010).

165. von Bohlen und Halbach, O. Immunohistological markers for staging neurogenesis in adult hippocampus. *Cell Tissue Res.* **329,** 409–420 (2007).

166. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins Struct. Funct. Bioinforma.* **59,** 687–696 (2005).

167. Waldron, L. *et al.* The Cardiac TBX5 Interactome Reveals a Chromatin Remodeling Network Essential for Cardiac Septation. *Dev. Cell* **36,** 262–275 (2016).

168. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5,** 16923 (2015).

169. Wei, B. *et al.* A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. *Nat. Biotechnol.* **36,** 521–529 (2018).

170. Wei, G.-H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* **29,** 2147–60 (2010).

171. Weirauch, M. T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31,** 126–34 (2013).

172. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158,** 1431–1443 (2014).

173. Wittmann, M. *et al.* Synaptic Activity Induces Dramatic Changes in the Geometry of the Cell Nucleus: Interplay between Nuclear Structure, Histone H3 Phosphorylation, and Nuclear Calcium Signaling. *J. Neurosci.* **29,** 14687–14700 (2009).

174. Woronowicz, A. *et al.* Carboxypeptidase E knockout mice exhibit abnormal dendritic arborization and spine morphology in central nervous system neurons. *J. Neurosci. Res.* **88,** 64–72 (2010).

175. Worsley Hunt, R., Mathelier, A., Del Peso, L. & Wasserman, W. W. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics* **15,** 472 (2014).

176. Xie, L. *et al.* FOXO1 is a tumor suppressor in classical Hodgkin lymphoma. *Blood* **119,** 3503–3511 (2012).

177. Yang, L. *et al.* Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.* **13,** 1–14 (2017).

178. Zhang, B. *et al.* Prognostic Significance of Phosphorylated FOXO1 Expression in Soft Tissue Sarcoma. *Ann. Surg. Oncol.* **16,** 1925–1937 (2009).

179. Zhang, L. *et al.* SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. *Genome Res.* **28,** 111–121 (2018).

180. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* (2008). doi:10.1186/gb-2008-9-9-r137

181. Zhou, T. *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U. S. A.* 1–6 (2015). doi:10.1073/pnas.1422023112

182. Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature* **562,** 76–81 (2018).