A computational study of nucleosomal binding and alternative isoforms of human transcription factors

Sviatoslav Sidorov

Imperial College London

(The Institute of Clinical Sciences)

and

The Francis Crick Institute

Supervisors

Professor Nicholas Luscombe, Professor Boris Lenhard

A thesis submitted for the degree of

Doctor of Philosophy

Imperial College London

September 2022

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work.

This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Declaration

I hereby declare that the present thesis is my own work, except where a due reference is made in the text.

Abstract

Eukaryotic transcription factors (TFs) are proteins that bind short DNA motifs and regulate gene transcription. Because genomic DNA is organised into nucleosomes via binding histone octamers, TFs compete with histones for binding DNA. Also, the functions of a TF are mainly defined by its domains; therefore, a TF gene can vary the characteristics of its protein product through the expression of alternative isoforms with different domains. However, the mechanisms of TF-nucleosome interactions and the functional importance of alternative TF isoforms are not fully understood. Here, I address these two problems computationally via the integrative analysis of publicly available *in vivo* human sequencing data. First, I evaluated a novel, gyre-spanning, mode of TF-nucleosome binding proposed recently by another lab based on *in vitro* evidence. Analysing the nucleosome occupancy and TF binding in the human genome, I found no evidence of such binding and concluded that it must be extremely rare, if at all present. Secondly, I studied the alternative isoforms of human TFs genome-wide. I found that independently of the gene length and the number of exons, TF genes more efficiently sample the set of possible alternative isoforms than non-TF genes, suggesting the particular importance of alternative isoforms for TFs. Also, I found that TF isoforms without a DNA-binding domain (DBD) are produced by almost a third of all human TFs, tend to be tissue-specific and likely reverse the transcription regulation effect of DBD-containing isoforms. Moreover, I demonstrated that the switches of the highest-expressed TF isoforms across human adult tissues may represent a widespread functional mechanism. Finally, I collected a compendium of human TFs with experimentally characterised alternative isoforms which will hopefully serve as a resource for future studies. In summary, my analysis further developed the fundamental knowledge about the TF-nucleosome interactions and the alternative isoforms of TFs in humans.

Acknowledgements

First of all, I would like to thank my supervisors, Professor Nicholas Luscombe and Professor Boris Lenhard, for giving me a chance to do a PhD in their brilliant labs, for helping me to learn the art of doing science, for the opportunity to define my own projects but also for the careful scientific guidance from the very start till the finish line of my PhD studies.

Secondly, I would like to express my gratitude to my thesis committee – Dr. James Briscoe, Professor David Jones and Dr. Tobias Warnecke – for their invariable attention to my research and for priceless feedback and support.

I also would like to thank The Francis Crick Institute for the generous funding of my PhD studies for four years, for the perfect administration of the PhD Programme and for a plethora of learning opportunities.

My PhD studies would not be possible without my outstanding colleagues in both labs. I would like to thank you for our scientific discussions, for all the feedback and support you provided to me during these four years and, of course, for our picnics and trips together that I will always remember!

I would like to thank Mahmoud-Reza Rafiee for giving me the opportunity to take part in his projects and through that – to further develop my knowledge of chromatin biology and my skills in bioinformatics. I also would like to thank Koustav Pal for his help in my project about alternative isoforms of human transcription factors that I present in this thesis.

My special thanks go to the many of my colleagues who carefully read and corrected different parts of my PhD thesis: Charlotte Capitanchik, Anob Chakrabarty, Christopher Cheshire, Jake Cornwall-Scoones, Beibei Du-Harpur, Radina Georgieva, Nejc Haberman, Ira Iosub, Marc Jones, Sarvesh Nikumbh, Koustav Pal, Mahmoud-Reza Rafiee and Oliver Ziff. Thank you so much for your time and effort! I am fully responsible for any remaining errors or incongruences.

Pursuing a PhD is inevitably demanding and challenging, and I would like to express my warmest gratitude to all my friends who kindly supported me throughout my journey! My special thanks go to Anton Afanasev, Rahul Arora, Anton Atamas, Pavel Avdeev, Amit Samani, Olga Sigalova, Sergey Simanovsky, Paul Villain and Pauline Misson, Johannes Welbl, and Karina Zile. You have always been kind and understanding, just somewhere near me in London or at a distance of a Zoom call, and you helped me immensely through my ups and downs. Thank you!

Finally, I would like to thank my parents – who have always unconditionally supported me in pursuing my interests and goals, who instilled a love of Nature and an interest in Science into me – and my passionate biology teacher, a life-long mentor and a dear friend Dr. Stephan Famelis. Without your kindness, support and belief in me, I would not be able to become a scientist.

Contents

Declaration 3

Abstract 4

Acknowledgements 5

Contents 7

List of Figures 9

List of Tables 12

List of Abbreviations 13

Chapter 1. Introduction 16

Chapter 2. Gyre-spanning nucleosomal binding of transcription factors in the human genome 20

2.1. Introduction 20

2.2. Methods 24

2.3. Results 29

2.4. Discussion 36

Chapter 3. Production of alternative isoforms by human transcription factors 39

3.1. Introduction 39

3.1.1. Non-DNA-binding domains of TFs 39

3.1.2. Alternative isoforms of TFs 40

3.2. Methods 45

3.2.1. Preprocessing of domain matches in TF isoforms 45

3.2.2. Preliminary classification of matched domains and definition of TF families 53

3.2.3. Selection of TF and non-TF isoforms 57

3.3. Results 59

3.3.1. Transcription factors produce more alternative isoforms than other coding genes 59

3.3.2. Compendium of human transcription factors with studied isoforms provides a basis for genome-wide predictions 62

3.4. Discussion 85

Chapter 4. Alternative isoforms without a DNA-binding domain 86

4.1. Introduction 86

4.2. Methods 88

4.2.1. Annotation of non-DNA-binding domains 88

4.2.2. Analysis of TF isoform expression 91

4.3. Results 93
4.3.1. Loss of DNA-binding domains is widespread among human transcription factors 93
4.3.2. Expression and tissue specificity of DBD- isoforms suggest their functional importance 99
4.3.3. Non-DBDs enable functional predictions for a part of DBD- isoforms 103
4.4. Discussion 109
4.4.1. Summary and interpretation of the results 109
4.4.2. Limitations of the analysis 111
4.4.3. Perspective 112

Chapter 5. Switches of major alternative isoforms of human transcription factors 114
5.1. Introduction 114
5.2. Methods 115
5.2.1. Analysis of switches of major TF isoforms across adult human tissues 115
5.2.2. Analysis of switch patterns across adult human tissues 118
5.3. Results 122
5.3.1. TFs switch major isoforms across adult human tissues with predictable functional outcomes 122
5.3.2. Switches of major isoforms demonstrate biologically relevant patterns 129
5.4. Discussion 148
5.4.1. Summary and interpretation of the results 148
5.4.2. Limitations of the analysis 150
5.4.3. Perspective 153

Chapter 6. Discussion 156
6.1. Summary 156
6.2. Perspective 161

References 163

List of Figures

Chapter 1

1.1 Structures and DNA-binding motifs of some classical DBDs…….. 16

1.2 Transcription regulation of a eukaryotic gene……... 18

Chapter 2

2.1 Structure of the nucleosome core…….. 21

2.2 TBXT-DNA binding……. 22

2.3 Log10-scaled numbers of matches of TBXT motifs……. 30

2.4 Positions of the GTM matches in strong nucleosomes……. 31

2.5 Average TBXT ChIP-seq coverage……... 33

2.6 Proportions of strong nucleosomes in genomic regions……... 34

2.7 Log10-scaled numbers of the GTM matches of selected TFs in strong nucleosomes…….. 35

Chapter 3

3.1 Schematic of the alternative splicing of a nascent RNA molecule to three different mRNA isoforms translated into three different proteins…… 41

3.2 Integrated coordinates of a domain match…… 45

3.3 General and specific InterPro entry matches…… 47

3.4 Examples of the rank-based filtering of InterPro entry matches in TF isoforms……. 49-50

3.5 The numbers of InterPro entry matches…… 51

3.6 Examples of merged signatures from Gene3D and PRINTS in non-C2H2 zinc finger matches…… 52

3.7 The numbers of signature matches…… 53

3.8 The numbers of unique InterPro entries, by type…… 54

3.9 TFs produce more isoforms than non-TF coding genes…… 60

3.10 Experimentally studied functional differences in alternative isoforms of human TFs. Part 1……. 63

3.11 Experimentally studied functional differences in alternative isoforms of human TFs. Part 2……. 66

Chapter 4

4.1 Examples of the deduplication of overlapping non-DBD matches……... 90

4.2 TF isoforms without a DBD…….. 95

4.3 Proportions of TFs with different numbers of DBDs…… 96

4.4 Proportions of single-DBD TFs with different numbers of DBD sequences… 97

4.5 Proportions of multi-DBD TFs with constant or changing DBD combinations……. 98

4.6 Expression of DBD+ and DBD- TF isoforms in adult human tissues…….. 101

4.7 Tissue specificity of DBD+ and DBD- TF isoforms……. 102

4.8 Expression and tissue specificity of DBD+ and DBD- TF isoforms per TF family…… 103

4.9 Proportions of DBD- isoforms by the types of non-DBDs that they contain... 104

4.10 Proportions of DBD- isoforms produced by the C2H2 zinc finger TFs and other TFs……... 105

Chapter 5

5.1 The definition of a major isoform switch…….. 116

5.2 Hypergeometric test of the overrepresentation of a gene set in a signature, in comparison to the background…….. 117

5.3 Switches between isoforms of the same TF…….. 119

5.4 Colour-coding of TF scores……... 121

5.5 Switching TFs……... 123

5.6 The types and functional categories of major isoform switches……... 124

5.7 Artificial “changes” in the DBD sequences of switching major isoforms……..……... 126

5.8 Changes in the DBD sequences of switching major isoforms. Part 1……... 127

5.9 Changes in the DBD sequences of switching major isoforms. Part 2……... 127

5.10 Overrepresentation of TF families in the functional categories of switches…. 129

5.11 Expression patterns of switching major isoforms across adult human tissues.. 131

5.12 PCA of human adult tissue samples…….. 132

5.13 Groups of TFs with pronounced brain-non-brain switch patterns…… 135

5.14A Major isoform switches in particular TFs. FOXJ3…… 138

5.14B Major isoform switches in particular TFs. ZNF12…… 139

5.14C Major isoform switches in particular TFs. IRF3……... 140

5.14D Major isoform switches in particular TFs. ZNF571…….. 141

5.14E Major isoform switches in particular TFs. TFEB……. 142

5.14F Major isoform switches in particular TFs. PRDM2…….. 143

5.14G Major isoform switches in particular TFs. WT1……... 144

5.14H Major isoform switches in particular TFs. MEF2C…….. 145

5.14I Major isoform switches in particular TFs. TBX3……. 146

5.14J Major isoform switches in particular TFs. ZNF655…….. 147

List of Tables

Table 1	Experimentally studied human TF isoforms with functional differences explained by domain differences…….	68
Table 2	Experimental details about TFs from Table 1 and their literature-based isoform expression patterns…….	71
Table 3	Other human TFs or small groups of TFs with studied alternative isoforms……	82
Table 4	Predicted functions of DBD- isoforms……..	106
Table 5	Groups of TFs with pronounced switches of major isoforms between the brain and non-brain tissues……	134
Supplementary Table 1	Domain matches in human TF isoforms……	*
Supplementary Table 2	Functional annotation of non-DBDs matched in human TF isoforms……..	*
Supplementary Table 3	Switches of major TF isoforms across adult human tissues..	*
Supplementary Table 4	Clusters of switching human TFs……..	*

* See sheets Suppl Table 1-4 in https://github.com/ComputationalRegulatoryGenomicsICL/human-tf-isoforms-thesis/blob/main/Supplementary_Tables.xlsx

List of Abbreviations

ATI	Alternative Transcription Initiation
ATT	Alternative Transcription Termination
bHLH	Basic Helix-Loop-Helix
bZIP	Basic Leucine Zipper
CAGE	Cap Analysis of Gene Expression
CAP-SELEX	Consecutive Affinity Purification SELEX
CATH	A database hierarchically classifying domains into four levels: Class, Architecture, Topology, Homologous superfamily
CDD	Conserved Domains Database
ChIP	Chromatin ImmunoPrecipitation
ChIP-exo	ChIP combined with lambda exonuclease digestion followed by sequencing
ChIP-seq	ChIP followed by sequencing
CNN	Convolutional Neural Network
CUT&RUN	Cleavage Under Targets & Release Using Nuclease
CUT&Tag	Cleavage Under Targets & Tagmentation
DBD	DNA-Binding Domain
EBI	European Bioinformatics Institute
ESC	Embryonic Stem Cell
FANTOM	Functional ANnoTation Of the Mammalian genome
FDR	False Discovery Rate
FIMO	Find Individual Motif Occurrences
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype-Tissue Expression database

GTF	General TF
GTM	Gapped TBXT Motif
HLH	Helix-Loop-Helix
HOCOMOCO	HOmo sapiens COmprehensive MOdel COllection
HT-SELEX	High-Throughput SELEX
IDR	Intrinsically Disordered Region
LBD	Ligand-Binding Domain
LOESS	LOcally Estimated Scatterplot Smoothing
MANE	Matched Annotation between NCBI and EBI
MEME	Multiple Em for Motif Elicitation
NCAP-SELEX	Nucleosome CAP-SELEX
NCBI	National Center for Biotechnology Information
NPC	Neural Progenitor Cell
NPIDB	Nucleic acid-Protein Interaction DataBase
NSC	Neural Stem Cell
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PCA	Principal Component Analysis
PDB	Protein Data Bank
PEAR	Paired-End reAd mergeR
Pfam	Database of Protein families
PFM	Position Frequency Matrix
PHD	Plant HomeoDomain
PIRSF	Protein Information Resource SuperFamily
PTM	Post-Translational Modification
PWM	Position Weight Matrix
RefSeq	Reference Sequence database

RNA-seq	RNA sequencing
SELEX	Systematic Evolution of Ligands by EXponential enrichment
SMART	Simple Modular Architecture Research Tool
TBP	TATA-box Binding Protein
TBXT	T-BoX transcription factor T
TF	Transcription Factor
TPM	Transcripts per kilobase Per Million
TRANSFAC	TRANScription FACtor database
TSL	Transcript Support Level
TSS	Transcription Start Site
UniPROBE	Universal Protein binding microarray Resource for Oligonucleotide Binding Evaluation
UTR	UnTranslated Region
ZF	Zinc Finger

Chapter 1. Introduction

Transcription factors (TFs) are proteins that bind DNA in a sequence-specific manner and influence transcription (Lambert et al., 2018). TFs interact with DNA via **DNA-binding domains** (DBDs). A DBD is a special case of a **protein domain** which is an independently folded functional part of the protein sequence (Janin and Wodak, 1983; Orengo et al., 1997; Wetlaufer, 1973). Like many other protein domains, DBDs are evolutionarily stable, and hence their sequence similarity is used to classify TFs into **families** (Weirauch and Hughes, 2011). Classical examples of the DBD types (and the corresponding TF families) include the basic leucine zipper (bZIP) (Jakoby et al., 2002; Noman et al., 2017; Tsukada et al., 2011), the basic helix-loop-helix (bHLH) (Jones, 2004; Skinner et al., 2010) and the homeodomain (Bürglin and Affolter, 2016; Gehring et al., 1994) (Fig. 1.1A-C). Of note, different residues within the DBD sequence have different roles in DNA binding. While residues that contact the DNA bases confer the binding specificity, those contacting the sugar-phosphate backbone stabilise the binding (Luscombe et al., 2001).

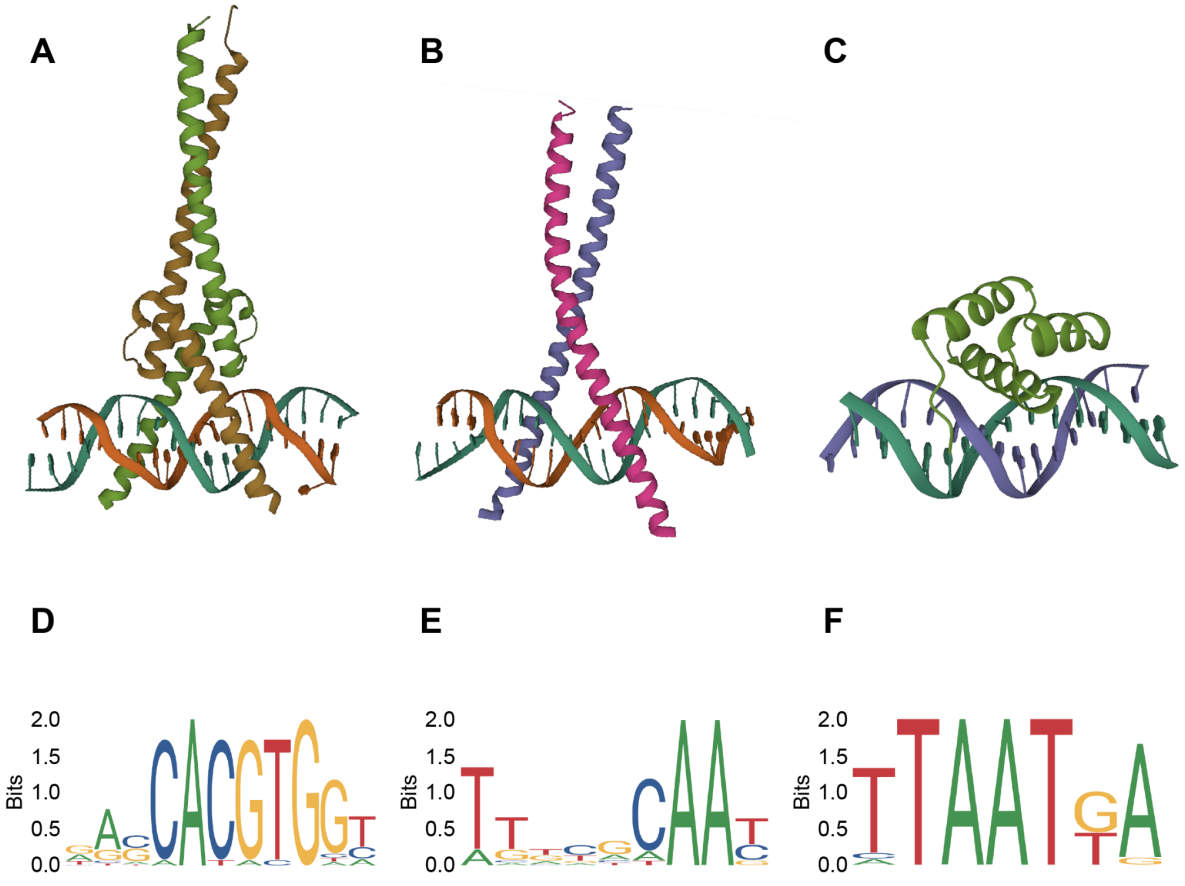


Fig. 1.1. Structures and DNA-binding motifs of some classical DBDs. (A) Example of bHLH TFs binding DNA: the structure of the heterodimer of the human MYC (green) and MAX (brown) TFs in complex with DNA (PDB ID 1NKP version 1.3: 2021-02-03) (Nair and Burley, 2003). **(B)** Example of bZIP TFs binding DNA: the structure of the *Rattus norvegicus* Cebpa homodimer in complex with DNA (PDB ID 1NWQ version 1.2: 2011-07-13) (Miller et al., 2003). **(C)** Example of a homeodomain binding DNA: the structure of the *Drosophila melanogaster* antennapedia homeodomain in complex with DNA (PDB ID 9ANT version 1.3: 2021-11-03) (Fraenkel and Pabo, 1998). The images in panels A-C were created using Mol* Viewer (Sehnal et al., 2021) from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) website (rcsb.org) (Berman et al., 2000; Burley et al., 2021). **(D)** DNA binding motif of the human MYC-MAX heterodimer (JASPAR ID MA0059.1). **(E)** DNA binding motif of Cebpa from *Mus musculus* and *Rattus norvegicus* (JASPAR ID MA0102.2). **(F)** DNA binding motif of Antp (Homeotic protein antennapedia) from *Drosophila melanogaster* (JASPAR ID MA0166.1). Images in panels D-F were adapted from the JASPAR CORE 2022 database (Castro-Mondragon et al., 2022).

Eukaryotic DBDs typically recognise and bind short (6-12 bp) **binding sites** (Lambert et al., 2018). A set of sequences derived from the binding sites of a specific TF can be summarised as a **binding motif**. TF binding sites have been studied for a long time as they play a key role in ensuring specificity of transcription regulation. Hundreds of motifs representing TF binding specificities have been collected in various databases, including TRANSFAC (Matys et al., 2006), UniPROBE (Hume et al., 2015), JASPAR (Castro-Mondragon et al., 2022), and HOCOMOCO (Kulakovskiy et al., 2018). These motifs are usually represented as position weight matrices (PWMs) where each element is a log-likelihood of a given nucleotide at a given position in a set of TF binding sites, calculated using chosen background nucleotide frequencies (Stormo, 2000). Motifs are typically visualised using **sequence logos** (Schneider and Stephens, 1990), where at each position letters are stacked, and the height of each letter is proportional to its frequency. For instance, Fig. 1.1D-F show the DNA binding motifs of the TFs from Fig. 1.1A-C.

Many TFs form **dimers** (complexes of two molecules) which can be subdivided into **homodimers** (complexes of two identical TF molecules) and **heterodimers** (complexes of two different TF molecules). For instance, Fig. 1.1A shows a heterodimer formed between human MYC and MAX TFs, while Fig. 1.1B demonstrates a rat Cebpa homodimer. Dimerisation is characteristic of certain TF families, including bHLH (Jones, 2004), bZIP (Jakoby et al., 2002;

Tsukada et al., 2011), nuclear receptors (Pardee et al., 2011), MADS-box (Shore and Sharrocks, 1995), STAT (Takeda and Akira, 2000), T-box (Papaioannou, 2014) and others. Dimerisation plays a prominent role in the regulation of the TF functionality (Amoutzias et al., 2008). It enhances the affinity and specificity of TF-DNA binding (Todeschini et al., 2014) and allows to switch a function of a TF by regulating the expression of its dimerisation partner (Lüscher and Vervoorts, 2012) or via a post-translational modification of the partner TF (Lluís et al., 2005). Additionally, dimerisation with TFs of disparate DNA binding specificities differentially regulates the DNA binding specificity of a TF (Hai and Curran, 1991) or the regulatory outcome (Lüscher and Vervoorts, 2012).

Of note, homodimers often bind palindromic DNA motifs, as the binding specificities of the monomers are identical but they face each other with their dimerisation surfaces. Fig. 1.1E shows a palindromic binding motif for the Cebpa homodimer, while Fig. 2.2B in section 2.1 shows a palindromic binding motif for the homodimer of a human T-box factor TBXT (Brachyury).

Whether in a monomeric or dimeric form, TFs regulate transcription via binding the DNA in **regulatory regions** of the genome, promoters and enhancers (Fig. 1.2).

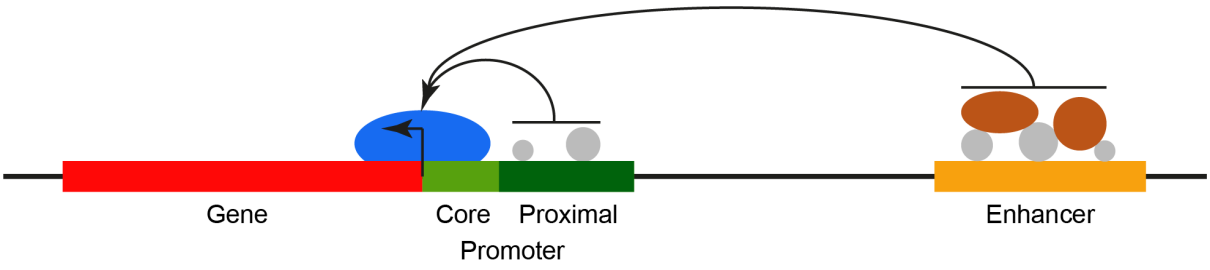


Fig. 1.2. Transcription regulation of a eukaryotic gene. The transcription is regulated by TFs bound to the proximal promoter and an enhancer and by recruited cofactors. The RNA polymerase complex is depicted by the blue ellipse, TFs are depicted by grey circles, and cofactors are depicted by the brown ellipse and circle.

Promoter is a region that contains the **transcription start site(s)** (TSSs) of a gene and where the RNA polymerase complex assembles (Lenhard et al., 2012). The particular locus where the assembly of the RNA polymerase complex takes place is called **core promoter**. The level of transcription from the core promoter is regulated by TFs that bind DNA upstream of the core promoter, in a region called **proximal promoter**, and in distant regions called

enhancers (Lenhard et al., 2012). **Enhancer** is a short genomic region that is distal from TSSs, contains TF-binding sites and regulates transcription. Specifically, an enhancer affects the transcription of its target gene when it is brought into a close proximity to the target core promoter and is bound by TFs which, in turn, recruit transcriptional **cofactors: coactivators** or **corepressors** (Shlyueva et al., 2014) (Fig. 1.2).

I discuss other facets of the TF biology in the introductory sections of the analysis chapters below.

Chapter 2. Gyre-spanning nucleosomal binding of transcription factors in the human genome

2.1. Introduction

TF binding motifs can be derived from various *in vitro* and *in vivo* experiments. Currently, one of the main *in vitro* methods used to study TF binding specificity is **HT-SELEX** (High-Throughput Systematic Evolution of Ligands by Exponential Enrichment) (Jolma et al., 2010). It enables the selection of DNA sequences with high affinity for a specific TF. Starting with all DNA sequences of a certain length (for example, 14 bp), it enriches sequences that are bound by the TF over the course of several cycles of binding and selection. This method was recently modified to study binding specificities of TF pairs on naked DNA (Consecutive Affinity Purification **(CAP) SELEX**, (Jolma et al., 2015)) and on nucleosomes (Nucleosome CAP **(NCAP) SELEX**, (Zhu et al., 2018)).

In vivo methods to assess TF binding specificities are based on immunoprecipitation of the TF of interest crosslinked with bound DNA fragments, followed by high-throughput sequencing. The first and most widely used method of this type is **ChIP-seq** (Chromatin Immunoprecipitation followed by Sequencing) (Johnson et al., 2007; Robertson et al., 2007). In the last decade, thousands of ChIP-seq experiments have been done for various TFs yielding several hundred different binding motifs for TFs in human and mouse (Kulakovskiy et al., 2018). A plethora of chromatin immunoprecipitation methods that utilise high-throughput sequencing were developed after ChIP-seq, including **ChIP-exo** (Rhee and Pugh, 2011) and **CUT&RUN** (Cleavage Under Targets and Release Using Nuclease) (Skene and Henikoff, 2017). Both methods identify TF-bound sites more precisely than ChIP-seq by cutting DNA fragments close to a bound TF.

When a DNA binding motif of a TF is known, it can be searched for (**scanned**) in a set of DNA sequences (for example, in ChIP-seq or CUT&RUN peaks) to find the TF binding sites. In a typical approach, each position in each sequence is scored using the position weight (or frequency) matrix that represents the motif (see Chapter 1), and a **match** of a motif is called at a given position of a given sequence if the p-value of the position, calculated based on the score and adjusted for multiple testing, is below a set cutoff (Grant et al., 2011).

However, TFs are not the only binders of genomic DNA; another prominent binder is the octamer of histone proteins H2A, H2B, H3 and H4. The histone octamer is wrapped ~1.7

times by the 145-147 bp **nucleosomal DNA,** and together the octamer and the DNA comprise the **nucleosome core.** The pair of nucleotides in the middle of the nucleosomal DNA is called the **dyad** (Luger et al., 1997). The two ~72-bp fragments of the nucleosomal DNA on both sides of the dyad are called the **gyres** (Zhou et al., 2019) (Fig. 2.1). Nucleosome cores are connected by fragments of exposed DNA called **linker DNA** and are stabilised by the **linker histone** H1. The nucleosome core, the linker DNA and the linker histone comprise the **nucleosome** (Luger et al., 1997). Nucleosomes with a stable genomic position across cells are called **strongly positioned nucleosomes** (or **strong nucleosomes**) (Gaffney et al., 2012).

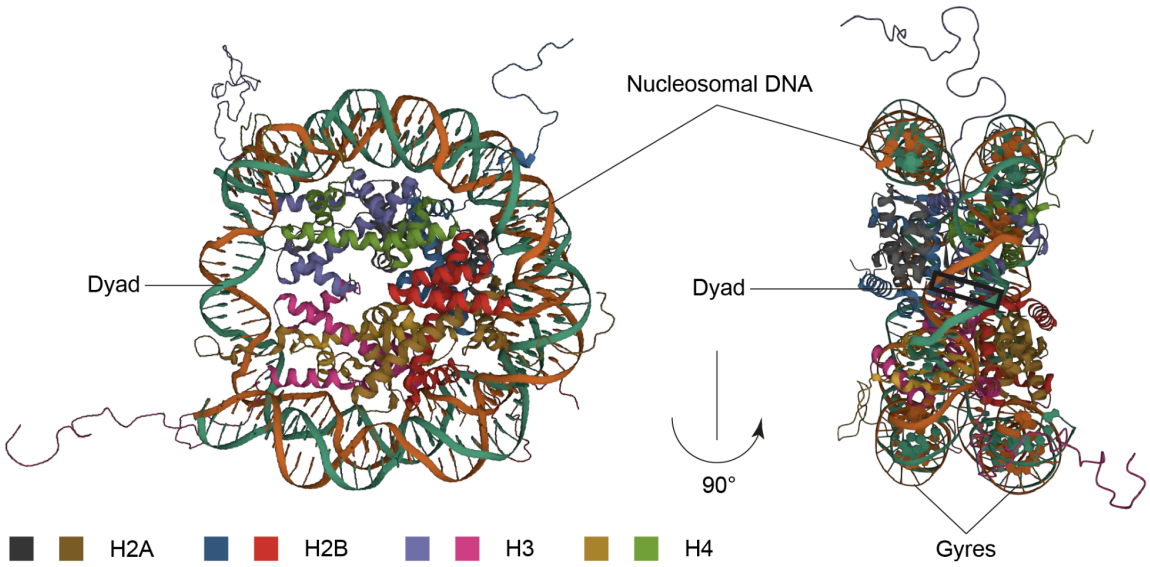


Fig. 2.1. Structure of the nucleosome core. Adapted from PDB ID 1KX5 version 1.2: 2011-07-13 (Davey et al., 2002). The image was created using Mol* Viewer (Sehnal et al., 2021) from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) website (rcsb.org) (Berman et al., 2000; Burley et al., 2021).

Nucleosomes frequently occupy inactive regulatory regions and pose a barrier for TF binding (Shlyueva et al., 2014). Hence, TFs have to compete with histone proteins for DNA binding, and this competition may occur in various forms (Klemm et al., 2019; Ramachandran and Henikoff, 2016). First, TFs can evict histones by achieving high concentrations at their target binding sites and thus disrupting the histone turnover on the DNA (Joseph et al., 2017; Miller and Widom, 2003; Mirny, 2010; Workman and Kingston, 1992). Secondly, TFs can bind the linker DNA and destabilise the nucleosome by evicting the linker histone (Ghosh et

al., 2010; Iwafuchi-Doi et al., 2016; Lone et al., 2013). They also can occupy enhancers and displace nucleosomes from target promoters (Taberlay et al., 2011). Finally, it has been argued that certain TFs that bind DNA first and establish its accessibility for other TFs (**pioneer factors**) are able to evict histone octamers via binding the nucleosomal DNA (Klemm et al., 2019). However, recent studies suggest that such pioneer TFs as OCT4 and SOX2 bind the entry and exit parts of the nucleosomal DNA rather than the nucleosome core *per se* and hence may actually compete with the linker histone or bind transiently accessible parts of the nucleosomal DNA (Echigoya et al., 2020; Michael et al., 2020). The competition with the linker histone was indeed shown for OCT4 and for another pioneer TF, FOXA1 (Echigoya et al., 2020; Iwafuchi-Doi et al., 2016).

To clarify the ability of human TFs to bind to the nucleosomal DNA, (Zhu et al., 2018) tested more than 200 human TFs *in vitro* using NCAP-SELEX. With this method, the authors selected DNA sequences bound by a specific TF in the presence of a nucleosome. Analysing the selected sequences for TF binding motifs, the authors obtained a particularly interesting result for a well-studied developmental TF TBXT (Brachyury). TBXT binds to a palindromic DNA motif (Fig. 2.2B) as a homodimer (Fig. 2.2A). However, apart from this known continuous binding motif, the authors discovered a disrupted motif in which the two halves of the continuous motif were separated by a 76-bp gap (Fig. 2.2D). Because the size of the gap was close to the length of one turn of the nucleosomal DNA around the histone octamer, the authors hypothesised that the TBXT homodimer could bind nucleosomes across DNA gyres (Fig. 2.2C).

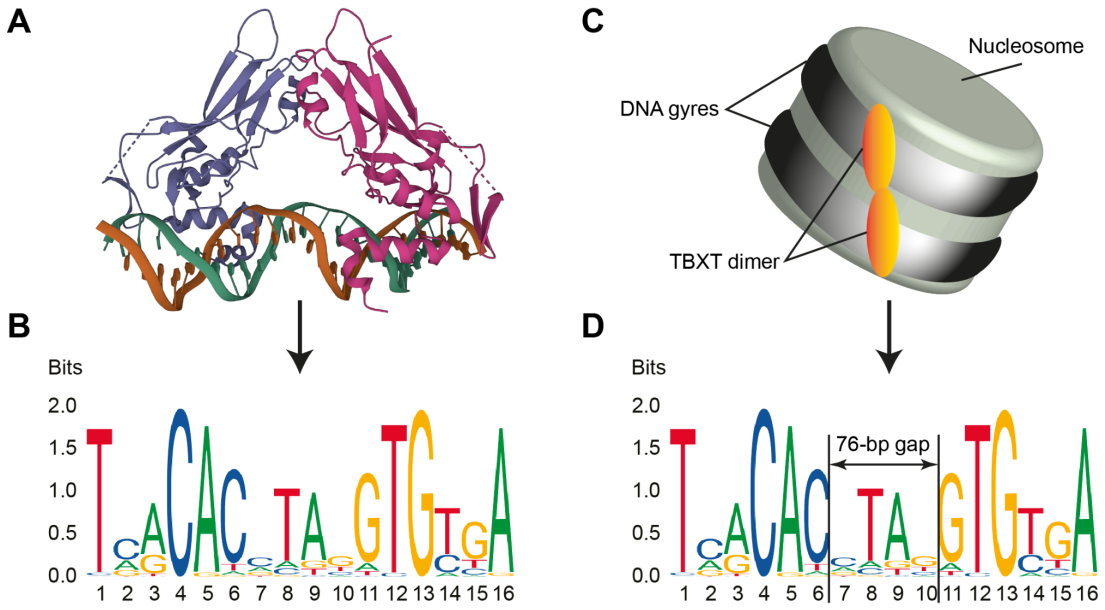


Fig. 2.2. TBXT-DNA binding. (A) Structure of the human TBXT homodimer in complex with DNA (PDB ID 6F58 version 1.0: 2017-12-13). The image was created using Mol* Viewer (Sehnal et al., 2021) from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) website (rcsb.org) (Berman et al., 2000; Burley et al., 2021). **(B)** DNA-binding motif of the human TBXT homodimer (JASPAR CORE ID MA0009.2). The image is adapted from the JASPAR CORE database (Castro-Mondragon et al., 2022). **(C)** Model of the gyre-spanning nucleosomal binding of the human TBXT dimer, inferred by (Zhu et al., 2018). See fig. 2c in the cited article for the original schematic. **(D)** Schematic of the disrupted nucleosomal DNA-binding motif, with the 76-bp gap, of the human TBXT homodimer. The original motif was obtained by (Zhu et al., 2018) in an NCAP-SELEX experiment and can be found in fig. 2b in the cited article. The schematic shown here is based on the regular, continuous, motif from panel B.

To test this hypothesis, (Zhu et al., 2018) searched for the **g**apped **T**BXT **m**otif (GTM) with a 76-bp gap (GTM76) in TBXT ChIP-seq peaks from human endoderm and mesoderm progenitors (Faial et al., 2015) but did not find any GTM76 matches. Nevertheless, I hypothesised that this negative result was not ultimate and decided to take another approach to search for the evidence of the novel, gyre-spanning binding of TBXT in the human genome. Specifically, in (Zhu et al., 2018), the authors noted that the binding of TBXT stabilised nucleosomes in the NCAP-SELEX experiment. Based on this observation, I supposed that if the gyre-spanning binding of TBXT is present in the human genome, then the strongly positioned nucleosomes are the most likely *loci* for this binding. However, because (Zhu et al., 2018) did not find GTM76 in the TBXT ChIP-seq peaks, I also supposed that the matches of this motif, if present, did not reach a set significance cutoff. Namely, (Zhu et al., 2018) scanned all TBXT peaks and used a raw p-value cutoff 0.0001; however, if they had scanned only the peaks intersecting with strong nucleosomes, they may have used a less stringent cutoff, and the GTM76 matches, if present in such peaks, may have become significant. Consequently, I decided to scan DNA sequences from strongly positioned nucleosomes in the human genome for GTMs of different gap sizes and then check the TBXT ChIP-seq signal specifically in strong nucleosomes with GTMs, if such nucleosomes exist.

2.2. Methods

Preprocessing and plotting scripts for the described analysis can be found at https://github.com/ComputationalRegulatoryGenomicsICL/gyre-spanning-binding-thesis.

To obtain the sequences of the strong nucleosomes in the human genome, I used *awk* and generated 151-bp sequences around the strong nucleosome dyads (N = 1,037,801) found by (Gaffney et al., 2012). In this way, the dyad occupied 1 bp, and the DNA gyres occupied 75 bp on both sides of the dyad. I chose the nucleosomal DNA length of 151 bp instead of the typical 147 bp to account for a possible slight variation in the strong nucleosome positions across the seven human lymphoblastoid cell lines assayed by (Gaffney et al., 2012). The dyad coordinates that I used were lifted over from the human genome assembly hg18 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.12/) to hg19 (Church et al., 2011) by Dr. Cai Li from the Luscombe Lab.

To obtain the sequences of the DNA fragments selected for the nucleosomal binding of TBXT in the NCAP-SELEX experiment performed by (Zhu et al., 2018) (a positive control), I downloaded the paired-end reads sequenced from these fragments (European Nucleotide Archive ID ERR2399503) and followed the preprocessing steps described in (Zhu et al., 2018). Specifically, I merged each pair of reads using PEAR version 0.9.11 (Zhang et al., 2014) and requiring a minimum overlap of 5 bp between reads. Next, I checked the quality of the paired reads and the merged fragments with FastQC (Andrews, 2010) and, using a custom Python script, I selected the fragments of length 154 bp, as (Zhu et al., 2018) used sequences of this length in their NCAP-SELEX experiment. Next, I used TrimGalore version 0.4.4 (Krueger et al., 2021; Martin, 2011) with the options `-q 20 --stringency 7 --length 154` to discard fragments with low-quality or adapter-contaminated ends. I additionally excluded fragments with a 14-bp overlap with either of the two fixed flanking sequences that (Zhu et al., 2018) ligated to the selected 154-bp sequences (5' flank: CCCTACACGACGCTCTTCCGATCT (24 bp), 3' flank: AGATCGGAAGAGCACACGTCTG (22 bp)). For this, I used Cutadapt version 1.9.1 (Martin, 2011) with the options `-a AGATCGGAAGAGCACACGTCTG -g CCCTACACGACGCTCTTCCGATCT -O 14 --discard-trimmed`. In this way, I obtained the final set of the TBXT NCAP-SELEX sequences (N = 163,050).

Finally, I used the utility *fasta-shuffle-letters* from the MEME Suite (Bailey et al., 2015) with the option `-kmer 2` to obtain the dinucleotide-shuffled sequences of the strong

nucleosomes and the TBXT NCAP-SELEX fragments to use them as negative controls. I also used the *shuffle* utility from BEDTools (Quinlan and Hall, 2010) with the options `-excl <strong_nucleosomes_BED> -seed 47474656 -f 0 -noOverlapping -i <strong_nucleosomes_BED> -g hg19.chrom.sizes` to select random non-overlapping 151-bp regions in the human genome outside the strong nucleosomes (N = 1,037,801). Here, `hg19.chrom.sizes` is a list of human chromosome sizes in the hg19 genome assembly. To obtain the corresponding sequences, I used the *getfasta* utility from BEDTools and the reference sequence hg19.

To obtain the gapped TBXT motifs (GTMs) with the gap sizes of 69-89 bp, I inserted from 65-85 rows between the two halves of the TBXT motif in the MA0009.2 position frequency matrix (PFM) of TBXT. I obtained this matrix from the JASPAR CORE database (Castro-Mondragon et al., 2022). I set the probability of each of the four nucleotides in each of the inserted rows to 0.25. I calculated the gap size in the same way the authors of (Zhu et al., 2018) did (Fig. 2.2D). Namely, I counted as the gap the positions between the CAC and GTG 3-mers; hence the difference by 4 between the number of added rows and the gap size.

To scan the GTMs, along with the original, continuous, TBXT motif MA0009.2, I obtained the mononucleotide background probabilities (0-order Markov models) separately from the strong nucleosome sequences, the positive control set of sequences, the negative control sets of sequences and the random sequences outside strong nucleosomes from the human genome. For this, I used the *fasta-get-markov* utility from the MEME Suite. Using the generated mononucleotide background probabilities, I scanned GTMs in the mentioned sets of sequences. For this, I used MEME FIMO (Grant et al., 2011) and filtered the resulting matches by the adjusted p-values under the FDR = 5%. As the scanned motifs are palindromic, I used a custom Python script to deduplicate matches found in the same sequence and with the same coordinates but on the opposite strands. From a pair of duplicate matches, I retained the one on the positive strand. Additionally, I filtered out matches of the paired motifs that colocated with the matches of the continuous motif. I visualised the log10-scaled numbers of matches with the *ComplexHeatmap* R package (Gu et al., 2016) version 2.7.9.1005 and used the GTM matches with the gap sizes of 77 and 79 bp (GTM77 and GTM79 matches) for further analyses.

To measure the distances between the strong nucleosome dyads and the GTM midpoints, I used custom Python scripts, Bash commands and the *closest* utility from BEDTools.

To test the binding of TBXT to the GTM matches, I downloaded the TBXT ChIP-seq data obtained by (Faial et al., 2015) from the mesoderm progenitors differentiated from the human embryonic stem cells. I chose a mesodermal dataset, as TBXT is an established factor of the mesoderm formation in vertebrates (Bruce and Winklbauer, 2020; Papaioannou, 2014). From the two available mesodermal datasets, I selected the one obtained with the R&D antibody, as (Faial et al., 2015) mentioned that the corresponding dataset had a higher percentage of uniquely mapped reads (GEO ID GSE60606; dataset IDs GSM1483191 (IP) and GSM1483195 (Input)). First of all, I assessed the quality of the datasets using FastQC and mapped the reads onto the hg19 genome assembly using Bowtie2 (Langmead and Salzberg, 2012) and allowing the mapper, by default, to search for multiple alignments but to output only the best one. Secondly, I sorted the obtained BAM files using the *sort* utility from samtools (Danecek et al., 2021) and deduplicated the mapped reads using the *MarkDuplicates* utility from the Picard tools (http://broadinstitute.github.io/picard/). Next, I converted the sorted and deduplicated BAM files to the bedGraph format with the *genomecov* utility from BEDTools and used the *intersect* utility from the same toolkit to filter the resulting bedGraph files against the *wgEncodeDacMapabilityConsensusExcludable* blacklist of the hg19 regions that tend to produce an artificially high ChIP signal (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/). After that, I used the filtered bedGraph files to generate the metaplots of the ChIP-seq coverage with the *genomation* R-package (Akalin et al., 2014) version 1.22.0.

To study the localisation of GTMs in the genome, I used the primary human genome annotation version 19 from the GENCODE database (Frankish et al., 2019) to separate genic and intergenic regions. As a representation of regulatory regions, I used a list of enhancers annotated by the FANTOM5 consortium (Andersson et al., 2014; Lizio et al., 2015, 2019) (https://fantom.gsc.riken.jp/5/datahub/hg19/enhancers/human_permissive_enhancers_phase_1_and_2.bb) and a comprehensive list of promoters created by the Lenhard Lab. First, from the FANTOM5 database they obtained the Cap Analysis of Gene Expression (CAGE)-based coordinates of transcription start sites (TSSs) for 808 samples of human tissues, primary cells and cell lines (Forrest et al., 2014; Lizio et al., 2015). Next, they pooled all TSSs, represented by tag counts and genomic coordinates in hg19, in a single file which was loaded into CAGEr (Haberle et al., 2015) version 1.2 for further analysis. Specifically, they calculated normalised tag counts in Transcripts per kilobase Per Million (TPMs) using a normalisation method described in (Balwierz et al., 2009), clustered TSSs with the normalised tag count ≥1 TPM within 20 bp from each other and assigned the obtained clusters to individual transcripts

annotated in Ensembl release 75 (Cunningham et al., 2022). For this purpose only, the clusters were overlapped with the Ensembl promoter regions constructed by adding 500 bp upstream and 2000 bp downstream of the first nucleotide of each annotated transcript. Finally, they considered only the clusters located within the Ensembl promoter regions; in each Ensembl promoter, they excluded clusters with the total signal < 5% of the total signal of the highest expressed cluster in that promoter; lastly, they excluded the Ensembl promoters from the analysis and defined a CAGE-based promoter as a region under a TSS cluster, spanning from the most upstream TSS to the most downstream TSS of the cluster. Promoters defined in this way formed the final set of per-transcript promoters that Radina Georgieva from the Lenhard Lab provided to me.

For the purposes of my analysis, I concatenated the list of promoters with the list of enhancers, sorted the resulting list by chromosome and the start coordinate and used the *merge* utility from BEDTools to merge any overlapping regions. In this way, I obtained 84,750 regulatory regions.

Using the primary human genome annotation version 19 from the GENCODE database and the *intersect* utility from BEDTools, I defined the *genic* strong nucleosomes as those overlapping annotated features; I called other strong nucleosomes *intergenic*. Among the genic strong nucleosomes, I defined the *exonic* ones as those overlapping with exons; I called other genic strong nucleosomes *intronic*. Among the intergenic strong nucleosomes, I defined the *regulatory* ones as those overlapping the regulatory regions from the list described above; I called other intergenic strong nucleosomes *non-regulatory*. In this analysis and in the analyses described below, I called an overlap between two regions if the regions shared at least 1 bp.

To analyse the localisation of strong nucleosomes relative to repeats, I used the RepeatMasker annotation of the human repeats in hg19 (Smith et al., 2013). I assigned a strong nucleosome to a repeat type if the strong nucleosome overlapped a repeat; otherwise, I assigned it to the category "Outside repeats." I assigned each repeat-overlapping strong nucleosome to a repeat type by choosing the repeat with the maximum overlap. If more than one type of repeat produced the maximum overlap with a given strong nucleosome, I randomly selected the repeat type for the nucleosome assignment from the uniform distribution across all such types. I used the *intersect* utility from BEDTools to find the overlaps of the strong nucleosomes with the genic and intergenic regions and also to find and measure the strong nucleosome-repeat overlaps. Additionally, I used the same tool to classify the GTM matches into those located within an Alu repeat, partially overlapping it and located outside. If a GTM match overlapped more than one Alu repeat, I selected the maximum overlap.

To search for the gyre-spanning nucleosomal binding of other TFs with palindromic motifs, I used the JASPAR CORE 2018 database to obtain 579 motifs (PFMs) of vertebrate TFs. Next, using a custom Python script, I selected palindromic motifs. Specifically, for each motif, I checked if its length L was even, because otherwise a motif would have a single position in the middle which would never be complementary to itself. If L was even, I next checked if the most frequent nucleotides in the corresponding positions i and $L - i + 1$ were complementary, where $1 \leq i \leq L$. If this condition was also met, I counted a motif as palindromic. In this way, I obtained 59 palindromic motifs.

After that, from each palindromic motif, using a custom Python script, I constructed 21 gapped motifs with gap sizes from 65-85 bp (which corresponded to the gaps of 69 to 89 bp for TBXT). I did this by inserting from 65-85 rows into the corresponding PFMs between the two halves of the motifs and defining the probability of each nucleotide in each inserted row as 0.25. Then, I scanned the obtained gapped motifs in strong nucleosomes in the same way I did for the TBXT motifs, using the same mononucleotide background probabilities derived from the strong nucleosome sequences and deduplicating the resulting matches.

Finally, for further analysis I selected only the TFs with at least 1,000 matches. As among them were not only human TFs but also murine ones, I checked that the murine TFs had human orthologues using the Mouse Genome Informatics (MGI) database (Blake et al., 2021). To visualise the log10-scaled numbers of matches of the scanned gapped motifs, I used the *ComplexHeatmap* R package version 2.7.9.1005.

2.3. Results

I started my analysis by scanning GTMs of gap sizes from 69-89 bp in 151-bp sequences extracted from the human reference genome around the dyads of strongly positioned nucleosomes (N = 1,037,801) found by (Gaffney et al., 2012). Importantly, (Gaffney et al., 2012) used lymphoblastoid cell lines to generate the set of strong nucleosomes, and such cell lines express TBXT, or the T gene, as demonstrated in the GTEx database (Lonsdale et al., 2013): https://gtexportal.org/home/gene/ENSG00000164458. Of note, I measured the gap size in GTMs in the same way as (Zhu et al., 2018) did (Fig. 2.2D); consequently, the regular (continuous) TBXT motif had a gap size of 4 bp. I also tested a positive control (154-bp long NCAP-SELEX sequences selected for the nucleosomal binding of TBXT, N = 163,050), two negative controls (dinucleotide-shuffled strong nucleosome and NCAP-SELEX sequences) and 1,037,801 random 151-bp sequences from the human genome outside strong nucleosomes (Fig. 2.3).

Strikingly, in strong nucleosomes, I found more than 11,000 matches of GTM77 and GTM79 (N = 4,124 and N = 7,339, respectively), while GTM76 matched only once. The gap sizes of 77 and 79 bp could still be compatible with the gyre-spanning binding. However, the difference in the matched gap sizes between my results and the results obtained by (Zhu et al., 2018) was intriguing and suggested that if GTMs indeed serve for the gyre-spanning binding of TBXT, then this binding could be facilitated by different gap sizes *in vitro* and *in vivo*.

Additionally, the positive and negative controls worked as expected. Specifically, in the TBXT NCAP-SELEX sequences, I found both the regular TBXT motif (N = 2,697) and GTM76 (N = 2,732). Interestingly, I found many more matches of GTM74 (N = 6,980), which is not readily explained given the fact that I measured the gap size in the same way (Zhu et al., 2018) did. One explanation would be an artefact in my assembly of the NCAP-SELEX sequences. On the other hand, the gap size of 74 bp is closer to the half-turn of the DNA molecule around the nucleosome core (73..74 bp) than the gap size of 76 bp.

In contrast, in the dinucleotide-shuffled nucleosomal and NCAP-SELEX sequences, I found, in total, only 6 matches of GTMs with different gap sizes. Consequently, GTMs are a property of only particular DNA sequences. Overall, GTMs are present in the human strong nucleosomes but their gap sizes differ by 1 or 3 bp from those found *in vitro* by (Zhu et al., 2018).

However, in the random genomic 151-bp sequences outside strong nucleosomes, I found considerable numbers of matches of GTM76 ($N = 210$), GTM77 ($N = 389$) and GTM79 ($N = 1{,}500$). This fact suggested that GTMs are not unique to strong nucleosomes and, consequently, may have a broader role than the gyre-spanning binding of TBXT (or no specific biological role). Additionally, GTM77 and GTM79 matched only 1% (11,402) of all strong nucleosomes, suggesting that if the gyre-spanning binding of TBXT indeed occurs *in vivo*, it does not serve as the primary mechanism of stabilisation of the nucleosome position. Nevertheless, inspired by the presence of thousands of GTM77 and GTM79 matches in the human strong nucleosomes, I decided to study these two gapped motifs in detail.

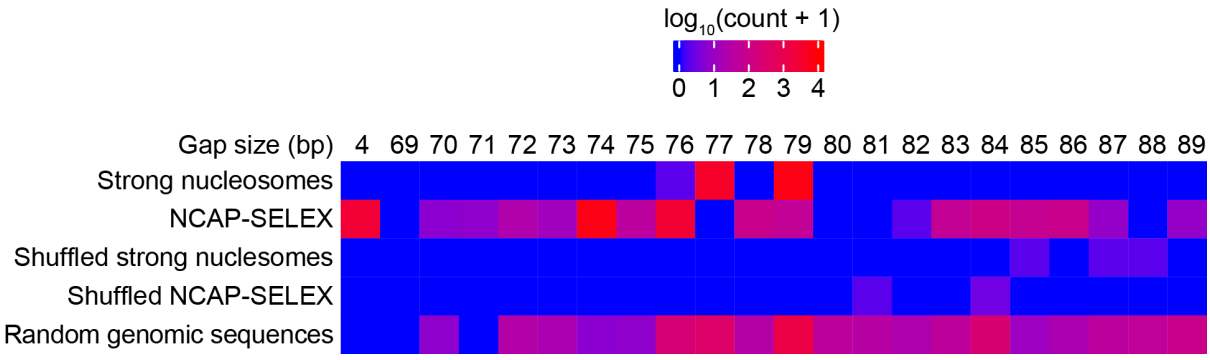


Fig. 2.3. Log10-scaled numbers of matches of TBXT motifs. The heatmap represents the matches of the regular (continuous) TBXT motif (gap size of 4 bp) and gapped TBXT motifs (GTMs) with gap sizes from 69-89 bp in the human strong nucleosomes, in the positive control (NCAP-SELEX sequences for TBXT), in negative controls (dinucleotide-shuffled strong nucleosomes and TBXT NCAP-SELEX sequences) and in random human genomic sequences outside strong nucleosomes.

First of all, I studied the positions of GTM77 and GTM79 matches in strong nucleosomes (Fig. 2.4). Namely, for each GTM match, I measured the distance d between the midpoint of the match and the dyad of the strong nucleosome that contained the match (Fig. 2.4A). Of note, $d = 0$ when the dyad and the midpoint reside in the same nucleotide and reaches a maximum value when a GTM match starts at the first or ends at the last nucleotide of the nucleosomal sequence. The maximum value is given by the matches of GTM77, as the length of the half-motif (TCACAC or GTGTGA) is 6 bp, the half-length of the gap, excluding the midpoint, is $(77 - 1)/2 = 38$ bp, and, hence, the remaining distance between the

midpoint and the dyad, excluding the dyad itself, is $(151-1)/2-(38+6)=31$ bp. The same calculations for the GTM79 matches give $max(d)=30$ bp.

I found that the distance d had a characteristic value of 9..10 bp; however, the longer distances ($\geq$ 23 bp) were more frequent, with $d=30..31$ bp being the most frequent (Fig. 2.4B). This trend held for both the GTM77 and GTM79 matches, although the GTM79 matches had another characteristic distance value, $d=16..19$ bp. The distance of 9..10 bp corresponds to a remote location of a GTM from the dyad (Fig. 2.4C, left), while the progressively longer distance corresponds to the progressively closer location of a GTM to the dyad (Fig. 2.4C, right).

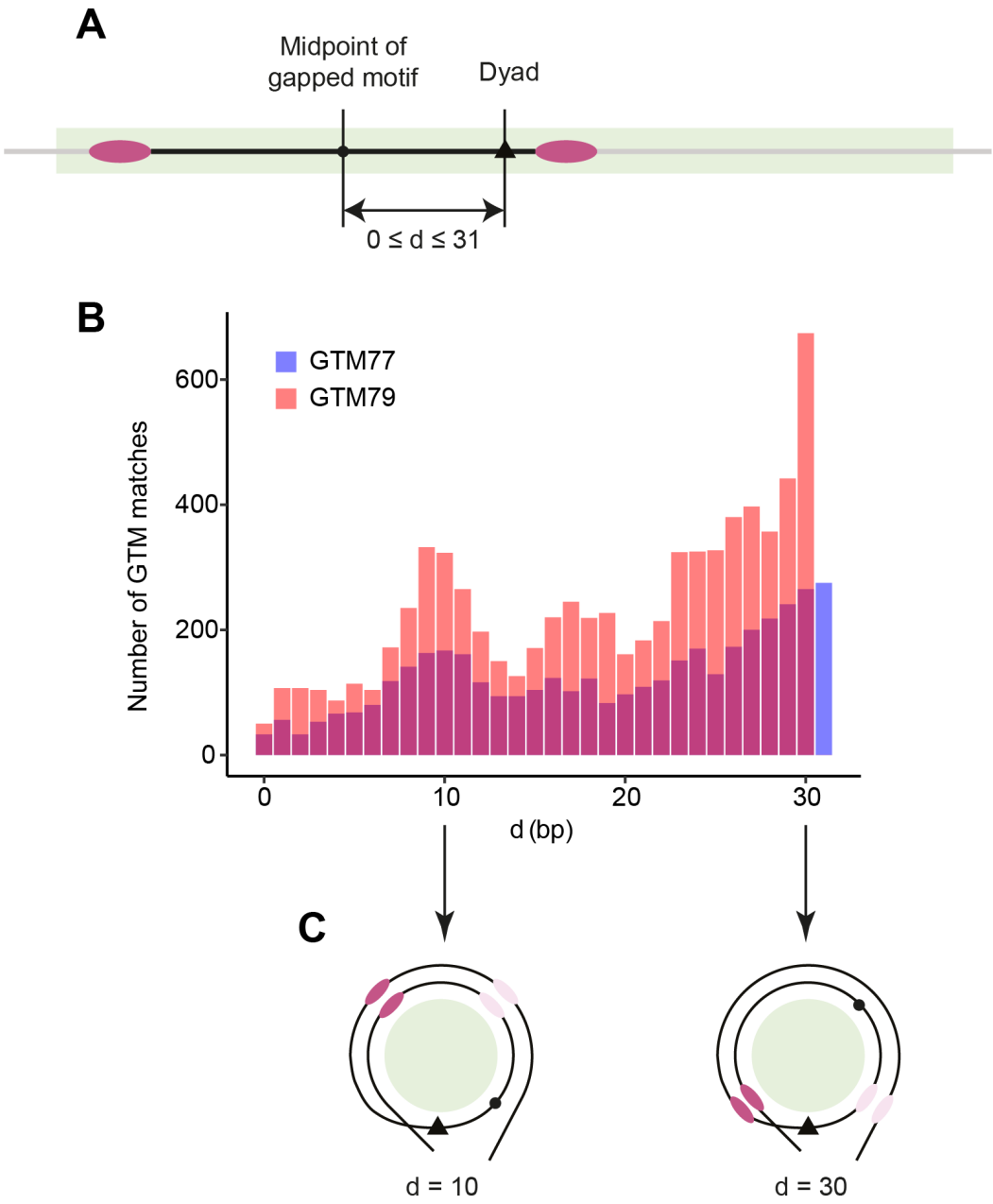


Fig. 2.4. Positions of the GTM matches in strong nucleosomes. (A) Measurement of the distance *d* between the midpoint of a GTM match and the dyad of a strong nucleosome that contains the match. The magenta ellipses depict the two matched halves of a GTM. **(B)** Histogram of the numbers of the GTM77 and GTM79 matches for all possible values of the distance *d*. **(C)** Schematics of the GTM positions along the nucleosome gyres when $d = 10$ (*left*) or $d = 30$ (*right*). As in panel A, the triangle denotes the dyad, and the dot denotes the GTM midpoint. Magenta ellipses depict the two matched halves of a GTM whose midpoint is marked by the dot. Light pink ellipses depict the symmetrical position of a GTM on the other side of the dyad for the same distance *d*.

Next, to identify if TBXT binds to the GTM77 and GTM79 matches in strong nucleosomes, I analysed TBXT ChIP-seq data from human mesoderm progenitors (Faial et al., 2015). From 23,555 significant TBXT ChIP-seq peaks found by (Faial et al., 2015), only 4 intersected the GTM77 and GTM79 matches. This finding disproved my hypothesis that the lack of GTMs in TBXT ChIP-seq peaks was a statistical artefact of the search for GTMs in all peaks. However, instead of peaks, I decided to test the TBXT ChIP-seq read coverage at strong nucleosomes (Fig. 2.5). I compared the mean coverage in strong nucleosomes with the GTM77 and GTM79 matches (GTM+ strong nucleosomes, N = 11,401) to the mean coverage in the same number of randomly chosen strong nucleosomes without GTMs (GTM- strong nucleosomes). My analysis showed almost no difference in the coverage between the two groups of strong nucleosomes and, importantly, the clear excess of the mean coverage of input reads over the mean coverage of IP reads in these *loci* (Fig. 2.5). Therefore, I obtained no evidence of the TBXT binding to GTMs in strong nucleosomes. Consequently, if this kind of TBXT-DNA binding occurs *in vivo*, it must be extremely rare.

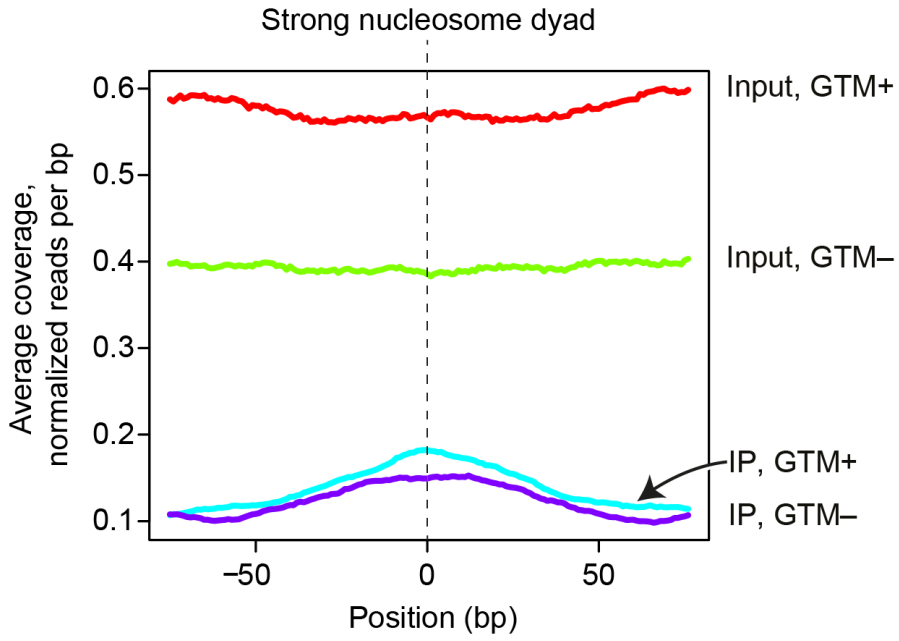


Fig. 2.5. Average TBXT ChIP-seq coverage. The coverage is calculated in strong nucleosomes with the GTM77 and GTM79 matches (GTM+ strong nucleosomes) and in strong nucleosomes without these matches (GTM- strong nucleosomes). The graph is centred around the strong nucleosome dyads.

Nevertheless, I was surprised by the lack of the TBXT ChIP-seq signal at the GTM77 and GTM79 matches, therefore I decided to investigate these matches further. Specifically, I analysed the localisation of the strong nucleosomes with the GTM77 and GTM79 matches (GTM+ strong nucleosomes, $N = 11{,}401$) in the human genome, using all strong nucleosomes ($N = 1{,}037{,}801$) as a control. To this end, I counted strong nucleosomes overlapping different types of regions in the genome.

I found that the localisation of the GTM+ strong nucleosomes was similar to the localisation of all strong nucleosomes and that 98% of GTM+ strong nucleosomes lay in intronic or non-regulatory intergenic regions (Fig. 2.6A). This result suggested that the majority of the GTM+ strong nucleosomes were in repetitive DNA. To test this hypothesis, I used the RepeatMasker annotation of the human repeats (Smith et al., 2013). I found that 99% of the GTM+ strong nucleosomes overlapped repeats, in comparison to only 86% of all strong nucleosomes. Furthermore, 98% of the GTM+ strong nucleosomes had the maximum overlap with Alu elements, in comparison to 43% of all strong nucleosomes (Fig. 2.6B). These findings suggested that GTM77 and GTM79 are a part of the Alu sequence. To show this explicitly, I counted the GTM77 and GTM79 matches that were fully within, overlapped or located outside

an Alu repeat. I found that 97% of the GTM matches located within an Alu sequence (Fig. 2.6C), which proved that the obtained GTM matches are Alu-specific.

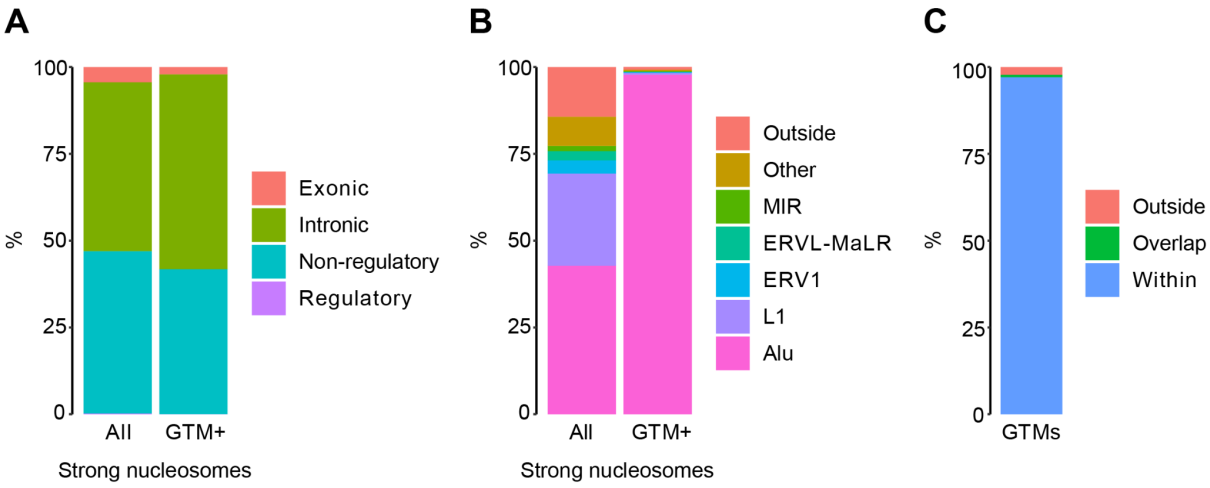


Fig. 2.6. Proportions of strong nucleosomes in genomic regions. (A) Proportions of all strong nucleosomes (*left*) and GTM+ strong nucleosomes (*right*) overlapping genic regions (exons or introns) or intergenic regions (regulatory or non-regulatory). **(B)** Proportions of all strong nucleosomes (*left*) and GTM+ strong nucleosomes (*right*) overlapping different repeat families. If a strong nucleosome overlapped repeats of more than one type, it was assigned to the type that produced the maximum overlap. If more than one type of repeats produced the maximum overlap with a strong nucleosome, then the type for the nucleosome assignment was selected randomly from the uniform distribution across all such types. **(C)** Proportions of the GTM77 and GTM79 matches within, partially overlapping or outside Alu elements.

Finally, because I did not find evidence for the gyre-spanning binding of TBXT in the human genome *in vivo*, I decided to explore the possibility of the gyre-spanning binding of other human TFs with palindromic motifs, supposing that the gyre-spanning binding of those TFs would also stabilise nucleosomes and hence be characteristic of strong nucleosomes. First, I obtained the set of vertebrate TF motifs (N = 579) available in the JASPAR CORE 2018 database (Castro-Mondragon et al., 2022) and selected palindromic motifs (N = 59). Next, for each palindromic motif, I generated a set of 21 gapped motifs with gap sizes from 65-85 bp and scanned the gapped motifs in all strong nucleosomes (N = 1,037,801). I obtained 960,991 significant matches (FDR = 5%) and selected TFs with at least 1,000 matches. With this, I obtained 9 TFs (ARNT, E2F1, NHLH1, CEBPB, TCFL5, ASCL2, HES7, CEBPD, CEBPE) which were either human or murine with a human orthologue, with 950,308 matches in total

(99% of matches of all 59 TFs with palindromic motifs). I considered gap sizes from 70-79 bp compatible with the gyre-spanning binding (Fig. 2.7). However, nearly all abundant motif matches for the selected TFs lay outside of this range (NHLH1, CEBPB, ASCL2, HES7, CEBPD, CEBPE), or the abundance of the matches only weakly depended on the gap size (ARNT, TCFL5). The only exception was a motif with a 76-bp gap for the transcription factor E2F1 (41,765 matches), but nearly 96% of its matches lay in repeat elements.

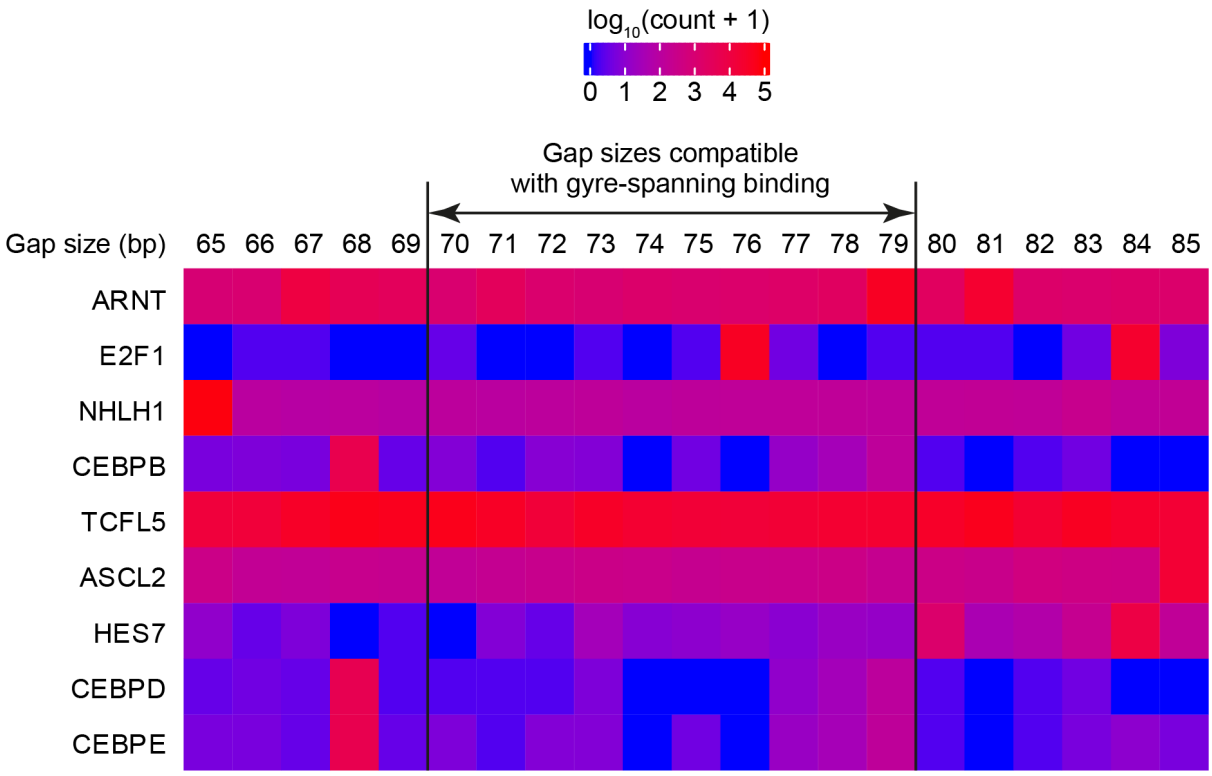


Fig. 2.7. Log10-scaled numbers of the GTM matches of selected TFs in strong nucleosomes. The numbers of matches are presented per gap size (bp).

Overall, based on my analysis, I concluded that gyre-spanning TF binding either is not present in the human genome *in vivo* or is extremely rare.

2.4. Discussion

In this chapter, I described my computational search for the putative gyre-spanning nucleosomal binding of TBXT in the human genome *in vivo*. Based on the *in vitro* results by (Zhu et al., 2018), I hypothesised that the gyre-spanning binding of TBXT could stabilise nucleosomes and make them strongly positioned. Therefore, I scanned gapped TBXT motifs (GTMs) in more than one million strong nucleosomes found previously in the human genome and analysed the TBXT ChIP-seq signal at the matched sequences. I found that the human genome does contain GTMs compatible with the gyre-spanning binding. However, nearly all of the GTM matches resided in Alu repeats, which questioned the functional relevance of the GTMs. Additionally, the only ChIP-seq dataset for TBXT in the human mesoderm progenitors that was available at the time did not show the TBXT binding signal at the found GTMs. Finally, I searched the strong human nucleosomes for gapped motifs of other dimerising TFs but the obtained matches either had gap sizes incompatible with the gyre-spanning binding or resided in repeats. Therefore, I concluded that if the gyre-spanning nucleosomal binding of TFs occurs in the human genome, it must be extremely rare.

Importantly, even if I had discovered the TBXT binding signal in the GTMs, I would have had to prove that this is indeed the *gyre-spanning* binding. Specifically, I would have had to demonstrate that the two TBXT molecules occupying the two halves of the gapped motif dimerise across DNA gyres, as some T-box factors, such as TBX1, TBX3 and TBX5, have been shown or suggested to be able to bind DNA without dimerisation (Papaioannou, 2014).

When I finished the project, I contacted Professor Jussi Taipale, the principal investigator in the (Zhu et al., 2018) study, to share my results. In his response, Professor Taipale told me that their laboratory also searched for various gapped motifs in the genome and also found them only in repeats.

Nevertheless, Professor Taipale suggested that some gapped motifs occurring in repeats could indeed bind TFs, repressing the repeat transcription or condensing the repetitive DNA. These possibilities are consistent with the fact that the presence of a specifically positioned nucleosome in an Alu element coincides with the repression of its transcription *in vitro* (Englander et al., 1993) and *in vivo* (Tanaka et al., 2010). He also reasonably suggested that the lack of an enrichment of the TBXT binding signal in the whole genome is not conclusive, because only a few binding sites could be enough for a specific function. Indeed, in my

analysis, I found that four TBXT ChIP-seq peaks overlapped with GTM77 or GTM79, but I have not analysed these peaks further.

Consequently, further research into the possible presence and functional role of the gyre-spanning nucleosomal TBXT binding in the human genome may be beneficial. First, a TBXT binding motif with a variable gap size could be searched for in the genomic DNA using a profile hidden Markov model (Durbin et al., 2002) with emission probabilities defined by the motif position weight matrix (PWM) and transition probabilities learnt from known gapped motif matches. One such model would replace a family of PWMs with fixed gap sizes and, potentially, allow finding additional, weaker, motifs that, nevertheless, could be functional.

Moving away from the motif search to direct *in silico* binding predictions, one could use a convolutional neural network (CNN) trained on experimentally obtained genomic TBXT binding sites from uniquely mappable regions to predict TBXT binding to Alu elements which have gapped TBXT motifs but are problematic in terms of short read mapping. For instance, (Avsec et al., 2021) developed a CNN called BPNet to predict DNA binding of several TFs depending on the distance between the TF motifs and found a novel putative binding pattern of Nanog with a 10.5-bp periodicity, manifested in synthetic sequences. As I outlined above, in the case of TBXT, one could apply BPNet to Alu elements. Additionally, to predict likely *nucleosomal* binding events, TBXT binding predictions could be integrated with the nucleosome binding predictions obtained using, for instance, ChromWave (Cakiroglu et al., 2021).

BPNet or a similar model could be trained on CUT&RUN data obtained by the laboratory of Professor Steven Henikoff for the TBXT binding in the human genome at four time points during the differentiation of the embryonic stem cell line H1 into the endoderm (Meers et al., 2019). This dataset, published after I finished my project, could be used to investigate the possible gyre-spanning binding of TBXT with a higher resolution than is possible with the ChIP-seq data.

If these CUT&RUN data show the TBXT binding at GTMs directly or via *in silico* predictions, then one could use DNase-seq to profile the nucleosome occupancy at the TBXT-bound GTMs in two human cell lines, one of which expresses TBXT while the other does not. If the gyre-spanning binding of TBXT occurs at GTMs and stabilises nucleosomes, one would expect to see a difference in the nucleosome occupancy profiles between the two cell lines, with the TBXT-expressing cell line showing greater occupancy at the TBXT-bound GTMs. However, with this approach, one still would not be able to discriminate between the nucleosomal binding and binding immediately next to the nucleosome in the linker DNA,

between the binding of one and two TBXT molecules and between the separate binding of two TBXT molecules and their dimerisation. Nevertheless, I can speculate that microscopy may be able to disentangle these possibilities, as various microscopy techniques have been successfully used to track individual TFs in living cells (Izeddin et al., 2014) and to demonstrate TF-chromatin binding (Garcia et al., 2021).

However, when mapping any short-read sequencing data, such as CUT&RUN or DNase-seq datasets, to Alu elements, one has to address the problem of the variable mappability of these repeats. Specifically, evolutionarily young (and hence, less degraded) Alu elements have poorer mappability than evolutionarily old (and hence, more diverse) Alu repeats (Li and Luscombe, 2020). This problem could be solved with an approach based on the one used by (Li and Luscombe, 2020). Specifically, one would need, first, to generate the mappability profile across all Alu elements for the reads of the required length with Genome Mappability Analyzer (Lee and Schatz, 2012) or GenMap (Pockrandt et al., 2020). Secondly, map the chosen sequencing data to the reference genome and study directly the Alu elements with the mappability above a certain predefined cutoff ("highly mappable" elements). Finally, study all the remaining Alu elements (with the mappability below the cutoff) using their consensus sequences. In order to do so, one would need to project the coordinates of reads mapped to these Alu elements onto the consensus sequences that represent the corresponding Alu subfamilies using the element-to-consensus mapping available in the Dfam database (Storer et al., 2021). Although in this way one would miss an enrichment present only in *particular* Alu elements with low mappability, this approach would still allow finding an enrichment present in *the majority* of the elements of a certain Alu subfamily.

Chapter 3. Production of alternative isoforms by human transcription factors

3.1. Introduction

Here, I describe the main types of non-DNA-binding domains in TFs, the general mechanisms of alternative isoform production and the current knowledge about alternative isoforms of TFs.

3.1.1. Non-DNA-binding domains of TFs

Apart from DNA-binding domain (DBDs), TFs can have domains that mediate their activity, or **effector domains**. These domains comprise a great variety which I summarise below.

Protein-binding domains are a group of effector domains that facilitate interactions of TFs with other proteins. First of all, these domains can serve for **dimerisation** of transcription factors. A well-studied example is the leucine zipper dimerisation domain which, for instance, facilitates the heterodimerisation of TFs Myc and Max (Lavigne et al., 1998). Also, protein-binding domains can facilitate **interactions of TFs with cofactors**, proteins that mediate TF activity. For instance, the phosphorylated kinase-inducible-domain (pKID) of the TF CREB binds the coactivator CBP and in this way mediates transcription activation (Radhakrishnan et al., 1997). Finally, some protein-binding domains mediate **interactions between TFs and general TFs (GTFs)** which are a particular set of TFs necessary for the site-specific transcription initiation by RNA Polymerase II (Thomas and Chiang, 2006). These interactions also mediate TF activity. For instance, the TF E2F1 interacts with GTFs, including TFIIA, TFIIB, TFIIH, the TATA-binding protein (TBP) and the TBP-associated factors (TAFs) and in this way activates transcription (Frietze and Farnham, 2011). In general, domains that mediate TF activity are called **transcription regulation domains** (Walhout, 2006).

Another class of effector domains is **ligand-binding domains (LBDs)**. Examples of ligands include hormones and small lipid molecules. For instance, through LBDs nuclear receptors bind hydrophobic **signalling molecules** derived from fatty acids, cholesterol, vitamins or hormones, xeno- or antibiotics (Sladek, 2011). Binding of a signalling molecule to an LBD of a nuclear receptor leads to a conformational switch of the LBD and ultimately – to changes in transcription regulation in response to the received signal (Pardee et al., 2011).

In general, TF domains can combine different functions, therefore the domain classes described above substantially overlap. Apart from zinc fingers which can bind DNA or proteins, another example are LBDs which often have functions other than ligand-binding. For instance, in nuclear receptors, LBDs mediate homo- and heterodimerisation and in this way ensure the correct spacing of DBDs. Additionally, LBDs may facilitate binding of cofactors, depending on the presence of a ligand (Pardee et al., 2011).

Overall, domains of different types are crucial for TF functionality, and, consequently, their differential presence or mutation can dramatically affect TF activity. However, apart from domains, proteins often contain regions that do not have a defined structure and hence are called **intrinsically disordered regions (IDRs)** (van der Lee et al., 2014). A plethora of studies suggest the functional importance of IDRs, especially in TFs. Indeed, TFs are enriched in IDRs (Liu et al., 2006a; Minezaki et al., 2006), and IDRs can regulate TF-TF interactions (Chong et al., 2018; Maiti et al., 2019; Xhani et al., 2020), TF-cofactor interactions (Boija et al., 2018; Kumar and Litwack, 2009) and, in general, are also known as RNA binders (Calabretta and Richard, 2015). Additionally, sequences around various DBDs were found to be significantly disordered (Guo et al., 2011), which may influence TF-DNA/chromatin binding. Furthermore, modelling studies by (Shoemaker et al., 2000) and (Vuzman et al., 2010) suggested that IDRs can accelerate the search for DNA binding sites by TFs, while recent experimental work showed that in particular TFs IDRs located outside DBDs facilitate sequence-specific DNA binding (Brodsky et al., 2020). Consequently, any changes in the protein sequence between well-structured domains may affect the function of the TF gene product.

3.1.2. Alternative isoforms of TFs

A key mechanism that allows one coding gene to produce multiple mRNA isoforms and, potentially, protein isoforms with different amino acid sequences is alternative splicing. **Splicing** is a step in the regulation of eukaryotic gene expression in which certain continuous fragments (**introns**) are removed from a nascent RNA molecule, while the other fragments (**exons**) are retained. Often, the same nascent RNA molecule can be spliced differently, and this phenomenon is called **alternative splicing** (Lee and Rio, 2015) (Fig. 3.1).

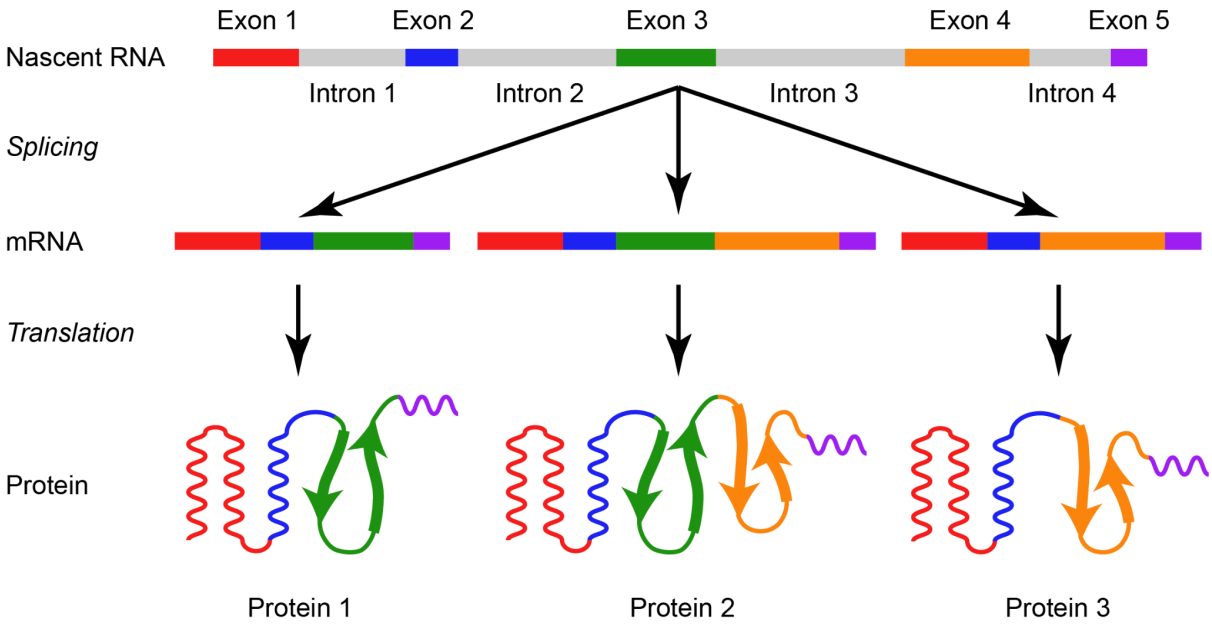


Fig. 3.1. Schematic of the alternative splicing of a nascent RNA molecule to three different mRNA isoforms translated into three different proteins. Only the mRNA that encodes Protein 2 contains all five exons and therefore produces the longest Protein 2, while the other two mRNAs lack either Exon 4 or Exon 3 and hence produce the shorter Protein 1 or Protein 3 of different amino acid sequences and structures. The 5' and 3' untranslated regions (UTRs) are not shown. Based on a work from the Wikimedia Commons (public domain): https://upload.wikimedia.org/wikipedia/commons/0/0a/DNA_alternative_splicing.gif.

In human, about 95% of all genes can be spliced alternatively (Pan et al., 2008; Wang et al., 2008). Although the overall scale of the translation of alternative mRNA isoforms has been actively debated (Agosto et al., 2019; Pickrell et al., 2010; Reixachs-Solé and Eyras, 2022; Tapial et al., 2017; Tress et al., 2017; Weatheritt et al., 2016), many studies show that alternative splicing allows the production of differentially acting protein isoforms and therefore plays a major role in the regulation of the cell functioning. Specifically, among its general functional roles, alternative splicing regulates protein-protein interactions (Buljan et al., 2012; Yang et al., 2016), protein-DNA interactions (Gabut et al., 2011; Raj et al., 2011), cellular signalling (Marti-Solano et al., 2020) and the level of mRNA translation (Floor and Doudna, 2016).

The variety of mRNA and protein isoforms is additionally increased by the **alternative transcription initiation (ATI)** and **alternative transcription termination (ATT)**. In (Shabalina et al., 2014), the authors showed that in mammals the total contribution of the ATI

and ATT to the mRNA and protein diversity is even greater than the contribution of alternative splicing. Moreover, (Reyes and Huber, 2018) demonstrated that in human the ATI and ATT generate the most part of differences between RNA isoforms transcribed in different tissues. The databases containing the human genome annotation, such as Ensembl (Cunningham et al., 2022) or RefSeq (O'Leary et al., 2016), give the opportunity to study the whole set of currently predicted alternative isoforms of coding and non-coding genes, independently of a particular mechanism of isoform production.

The study of alternative isoforms of TFs is especially interesting due to the key role that these proteins play in transcription regulation. Alternative isoforms of particular TFs have been studied in a wide variety of organisms, including plants (Chen et al., 2020; Seo et al., 2011, 2013), worms (Ram et al., 2004), insects (Goeke et al., 2003; Shin et al., 2002), and mammals (Choi et al., 1997; Dolfini et al., 2012; Hahm et al., 1994; Hemesath et al., 1994; Kozmik et al., 1997; Li et al., 1992; Lucena et al., 2016; Nishiyama et al., 2000). Regarding the latter, alternative TF isoforms have also been studied in human, and in section 3.3.2 I present a detailed account of known alternative isoforms of human TFs.

Apart from single TFs or small groups of related TFs, the production of alternative isoforms by TFs has also been studied on the genome-wide scale. First, in (Resch et al., 2004), the authors described the "Alternatively Spliced Protein forms" (ASP) database (not available anymore) that they constructed for 4,422 human genes by predicting alternative protein isoforms from a *de novo* transcriptome assembly. They found that alternative splicing predominantly removed certain functional types of domains from proteins, including TFs, and that it must have profoundly affected protein-protein interactions. Among the domains that underwent preferential alternative splicing were some domains typical for TFs, including KRAB, C4 type zinc fingers, SANT, BRIGHT and IRF. Also, among the examples of proteins functionally affected by alternative splicing, (Resch et al., 2004) listed a number of TFs, including KRAB zinc finger factors with affected protein-protein interactions and Pbx2 with an affected transcription repression domain. Later on, (Yura et al., 2006) confirmed that one of the major groups of human proteins affected by alternative splicing were zinc finger TFs.

Although these results clearly showed the importance of alternative splicing for the function of some human TFs and paved the way for the subsequent genome-wide studies of alternative protein isoforms, these results were obtained in the early days of human genomics and hence may not reflect the characteristics of all alternative TF isoforms known or predicted to date.

Next, (Taneri et al., 2004), based on their transcript annotation, reported that 62% of TFs were alternatively spliced in mice, compared to 29% of all genes. Later, (Talavera et al., 2009) counted experimentally-validated protein isoforms of TFs and non-TF genes in human and mouse and found that 29.4% of human TFs and 26.1% of mouse TFs underwent alternative splicing, in comparison to 26.1% and 17.6% of all genes in the corresponding species. Additionally, both articles reported that in human and mouse, TFs produce more alternative isoforms than non-TF genes. These results show the prevalence of alternative isoform production in TF genes, compared to non-TF genes. Additionally, (Taneri et al., 2004) found that alternative splicing predominantly affected DBDs and that TFs, similar to all murine genes, expressed different isoforms in different tissues. Also, (Talavera et al., 2009) reported that in 28% of human TFs at least one isoform was different from others and that the homeodomain, HOLI (a ligand-binding domain of hormone receptors), HLH (helix-loop-helix dimerisation domain) and PHD (protein-binding plant homeodomain) were most frequently affected by alternative splicing in both human and mouse.

However, in both articles, the authors did not attempt to predict possible functions of alternative TF isoforms. Additionally, these results were obtained more than a decade ago, and the human genome annotation, including the transcript annotation, has developed considerably since that time (Frankish et al., 2019).

Much more recently, (Soto et al., 2022) manually compiled a comprehensive database of 924 experimentally validated effector domains (specifically, transcription activation and repression domains) from 594 human TFs (TFRegDB, https://tfregdb.bu.edu/tfregdb/). The database allows searching for transcription regulation domains in the included alternative TF isoforms or to scan for these domains in an external amino acid sequence. The authors found that transcription regulation domains were affected by alternative splicing, ATI or ATT in a higher proportion of protein isoforms than DBDs. Additionally, transcription regulation domains of the Forkhead TFs were the least affected by indels, while transcription regulation domains in TFs from the bZIP and HMG/Sox families were the most affected. Overall, the database developed by (Soto et al., 2022) is a prominent step forward in the study of the transcription regulation domains in the alternative isoforms of human TFs; however, the authors do not attempt to predict functions of the alternative isoforms and do not include ~1,100 TFs in their database (likely, because of the lack of experimental evidence for the transcription regulation domains in those TFs).

Additionally, (Wang et al., 2021) developed a computational method Iso-Net to predict functions of alternative isoforms by their co-expression with genes that are differentially

expressed between cell types and by the GO term enrichments of these genes. The authors applied Iso-Net to 18 human TFs that take part in myeloid cell differentiation and studied the co-expression of their isoforms with the genes that are differentially expressed between human macrophages, neutrophils and monocytes. However, for the alternative TF isoforms that they studied, the authors predicted only a spectrum of biological processes in which the isoforms may take part, rather than particular functions of the isoforms. Additionally, by the design of the method, these predictions were indirect, based not on the features of the isoforms but on the GO terms associated with the co-expressed genes. Therefore, although this method could be applied to alternative isoforms of all human TFs, it would not be able to directly take into account any functional features encoded by alternative isoforms and to predict particular functions of these isoforms.

Finally, known functional differences between alternative isoforms of some TFs were summarised for developmental processes (López, 1995) and various types of cancer (Belluti et al., 2020) but not for the normal human physiology.

Overall, a genome-wide study of alternative isoforms of human TFs in normal physiology is currently lacking. Therefore, I devised a project to computationally characterise the differences between alternative TF isoforms in human with regards to all types of domains, possible isoform functions and the tissue-specificity of the isoform expression. In the current chapter, I compare the alternative isoform production by the TF and non-TF human genes and present a compendium of the human TFs with studied alternative isoforms. This compendium serves as a basis for my subsequent analysis. In Chapter 4, I analyse the differential presence of DBDs in alternative TF isoforms, as well as the expression levels, tissue specificity and possible functions of isoforms lacking a DBD (DBD- isoforms). I predict possible functions of the DBD- isoforms using a manually curated list of functionally characterised non-DBD domains. Finally, in Chapter 5, I formally define and study the cross-tissue switches of the top-expressed (major) DBD+ and DBD- isoforms of human TFs and predict functional consequences of these switches based on the differential presence of functionally characterised domains.

3.2. Methods

Scripts for the described analysis can be found at https://github.com/ComputationalRegulatoryGenomicsICL/human-tf-isoforms-thesis.

Below, I included the scanning of TF isoforms for known domains and the manual curation of the results (sections 3.2.1 and 3.2.2) because the presence of relevant DBDs affected the final selection of TFs that I used in my analysis.

3.2.1. Preprocessing of domain matches in TF isoforms

I selected all human TFs from the HumanTFs database (Lambert et al., 2018), version 1.01 (N = 1,639). For these genes, I selected coding TF transcripts from the Ensembl database (Cunningham et al., 2022) release 99 and downloaded their protein sequences (N = 7,491). Using InterProScan 5 (Jones et al., 2014), version 5.39-77.0, I searched for domains in these protein sequences and obtained InterPro entry matches in 6,622 protein TF isoforms (24,004 InterPro entry matches with 123,105 signature matches). Next, I processed these matches to obtain the final set of non-overlapping domains:

1. For each InterPro entry match, I calculated its integrated coordinates [*start*; *stop*] which are the outermost coordinates of signatures that comprise the entry (Fig. 3.2). In this procedure, I did not consider PIRSF and PANTHER signatures, if present, because those designate whole proteins rather than particular domains (Mi et al., 2017; Thomas et al., 2003; Wu et al., 2004).

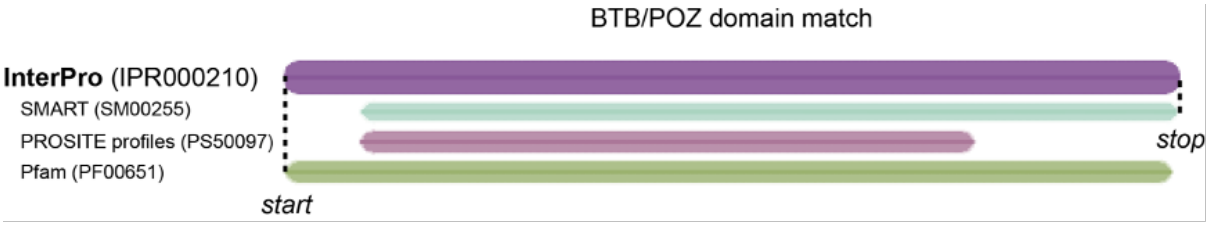


Fig. 3.2. Integrated coordinates of a domain match. The example shows the integrated coordinates [*start*; *stop*] of a domain match (InterPro entry IPR000210) which consists of the following signatures: SMART SM00255, PROSITE profiles PS50097, and Pfam PF00651. This schematic was adapted from the InterProScan online service (https://www.ebi.ac.uk/interpro/).

2. As the matches of InterPro entries that represent the same type of a domain often considerably overlap, I defined "specific" and "general" matches of InterPro entries to eventually remove the redundant ("general") matches from my analysis. I called a match of an InterPro entry *E* in an isoform *I* **general** if *E* is of the Domain or Family type (related entries of each of these two types form a hierarchy) and its child entries matched in *I* and overlapped the match of *E* by at least one amino acid (Fig. 3.3A). Otherwise, I called a match **specific** (Fig. 3.3B-C). I found general and specific InterPro entry matches using custom Bash and Python scripts and the following R packages: *yaml* (Garbett et al., 2022) version 2.3.5; *data.tree* (Glur, 2020) version 1.0.0. In the next several steps, I processed general and specific matches in the same way. Having general matches in addition to specific ones allowed me to better understand functions of domains present in TF isoforms. Out of 24,004 InterPro entry matches, 22,487 were specific and 1,517 were general.

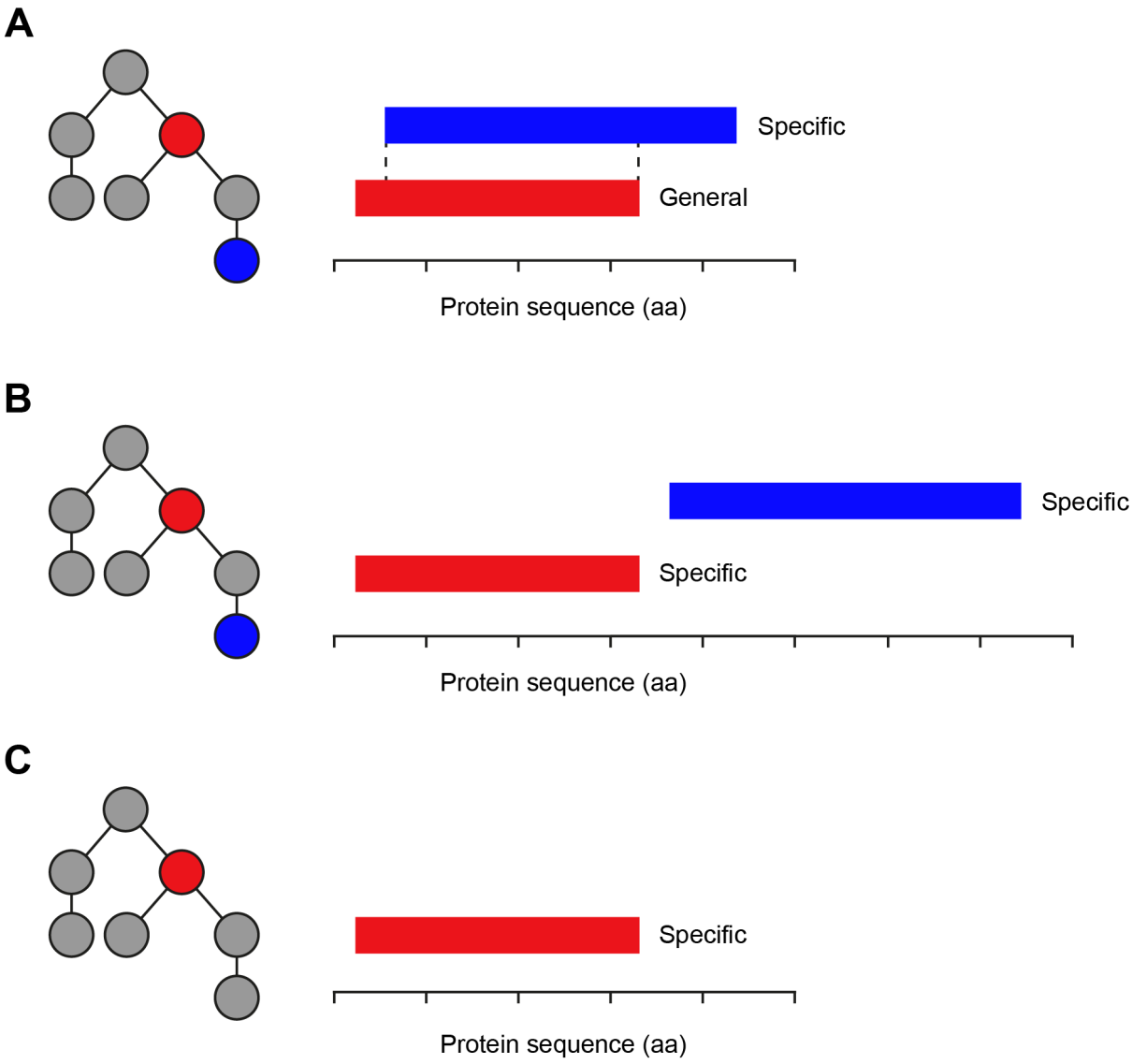


Fig. 3.3. General and specific InterPro entry matches. The hierarchy of entries on the left shows that the entry marked blue is a descendant of the entry marked red. **(A)** The red match is general, because its InterPro entry is ancestral to the blue entry that matched in the same TF isoform and overlapped the red match. The blue match is specific, because it does not have any child entries matched in the same isoform. **(B)** Both matches are specific, because, although the red entry is ancestral to the blue one and they match in the same isoform, their matches do not overlap. **(C)** The only present match is specific, because its entry does not have any child entries matched in the same TF isoform.

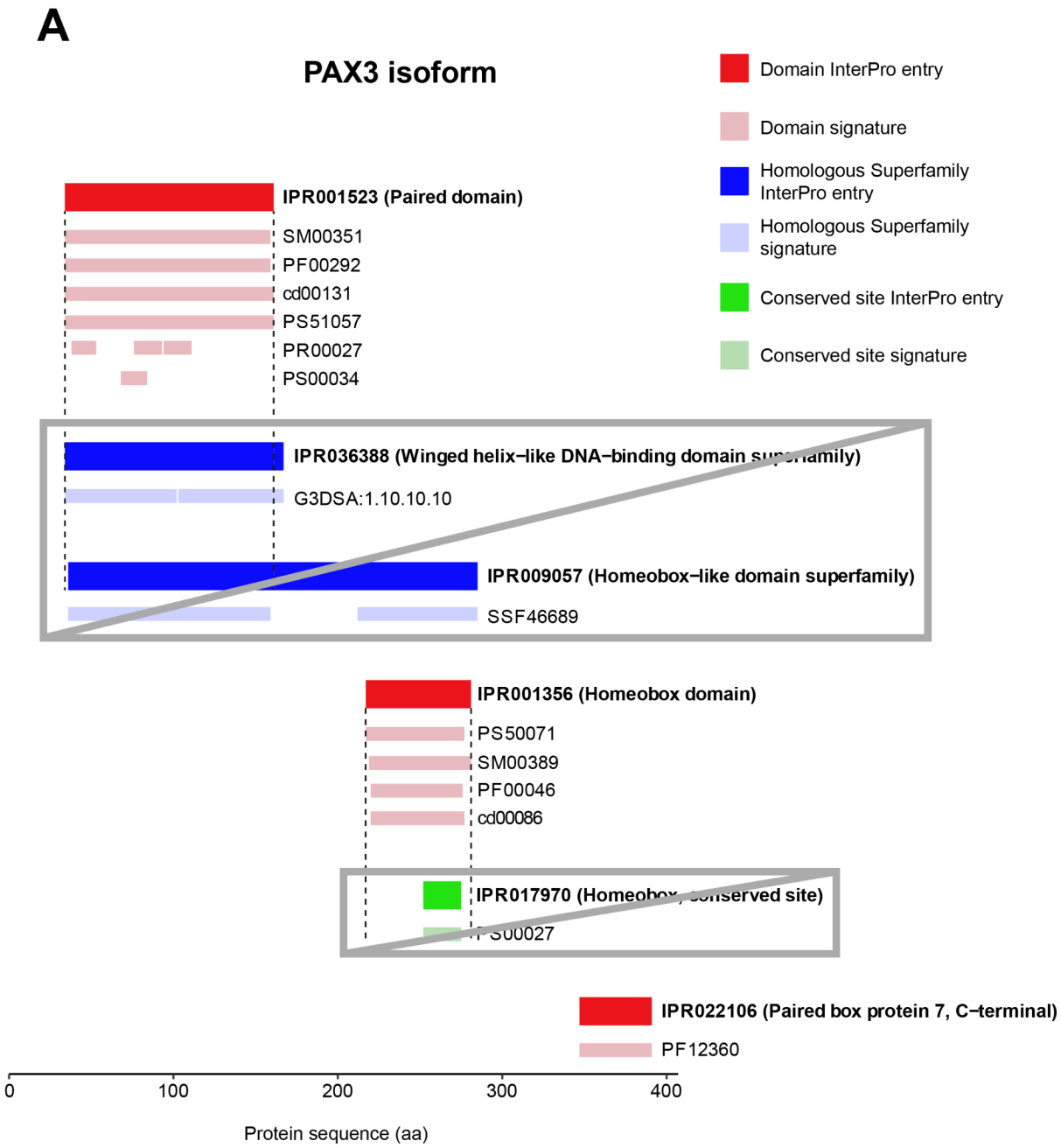
3. I ranked InterPro entry matches by the entry type in the order of decreasing accuracy of domain description:
    a. Domain.

b. Family matches with domain signatures, such as Pfam (Mistry et al., 2021), SMART (Letunic and Bork, 2018; Letunic et al., 2021), PrositeProfiles (Sigrist et al., 2013), CDD (Lu et al., 2020). I defined a separate *Domain-Family* type for these matches.
c. Homologous Superfamily.
d. Site.
e. Repeat.
f. Family (other matched InterPro entries of the Family type).

This ranking is based on the fact that the InterPro entries of the Domain and Family-Domain types describe domains by definition. Next, InterPro entries of the Homologous Superfamily type always consist of the CATH-Gene3D (Lewis et al., 2018; Sillitoe et al., 2021) and/or SUPERFAMILY (Gough et al., 2001) signatures which also represent domains. Finally, Site and Repeat matches represent short functional peptides that may reside within a bigger domain, in this way giving evidence for a domain if a domain match is missing, or facilitate a certain function on their own and, consequently, be as important as domains (https://www.ebi.ac.uk/training/online/course/interpro-functional-and-structural-analysis-protei/what-interpro-entry/interpro-entry-types). Of note, because the Domain, Family-Domain and Homologous Superfamily matches all represent domains, I could have given all of them the equal first priority in the ranking; however, as I use this ranking to filter overlapping matches (see below), I had to give these entry types some order of preference to maintain consistency in retained types.

4. Next, using the integrated coordinates of the InterPro entry matches, I excluded the matches of a lower rank that overlapped with the matches of a higher rank. Specifically, I excluded all non-Domain matches that overlapped the Domain ones, then I excluded all non-Family-Domain matches that overlapped the Family-Domain matches, and so on. Fig. 3.4 illustrates this procedure in the isoform ENST00000336840 of PAX3 (panel A) and in the isoform ENST00000177694 of TBX21 (panel B). In this way, I retained matches that likely describe domains in the most accurate way and also ensured that matches of InterPro entries of different types do not overlap. Additionally, before filtering the remaining InterPro entries against the entries of a given type, I deduplicated the overlapping specific matches of that type. In this

deduplication, I manually selected InterPro entries that were likely more specific, based on their descriptions in InterPro. I did this deduplication for all types of InterPro entries, apart from the Domain signatures which I processed later on. Fig. 3.5 shows the number of InterPro entry matches by type, before and after the filtering. Thus, the total number of InterPro entry matches decreased from 24,004 to 11,660, and the vast majority of the matches retained after the filtering are of the type Domain, which is expected. Additionally, the number of matched signatures decreased from 123,105 to 95,769.



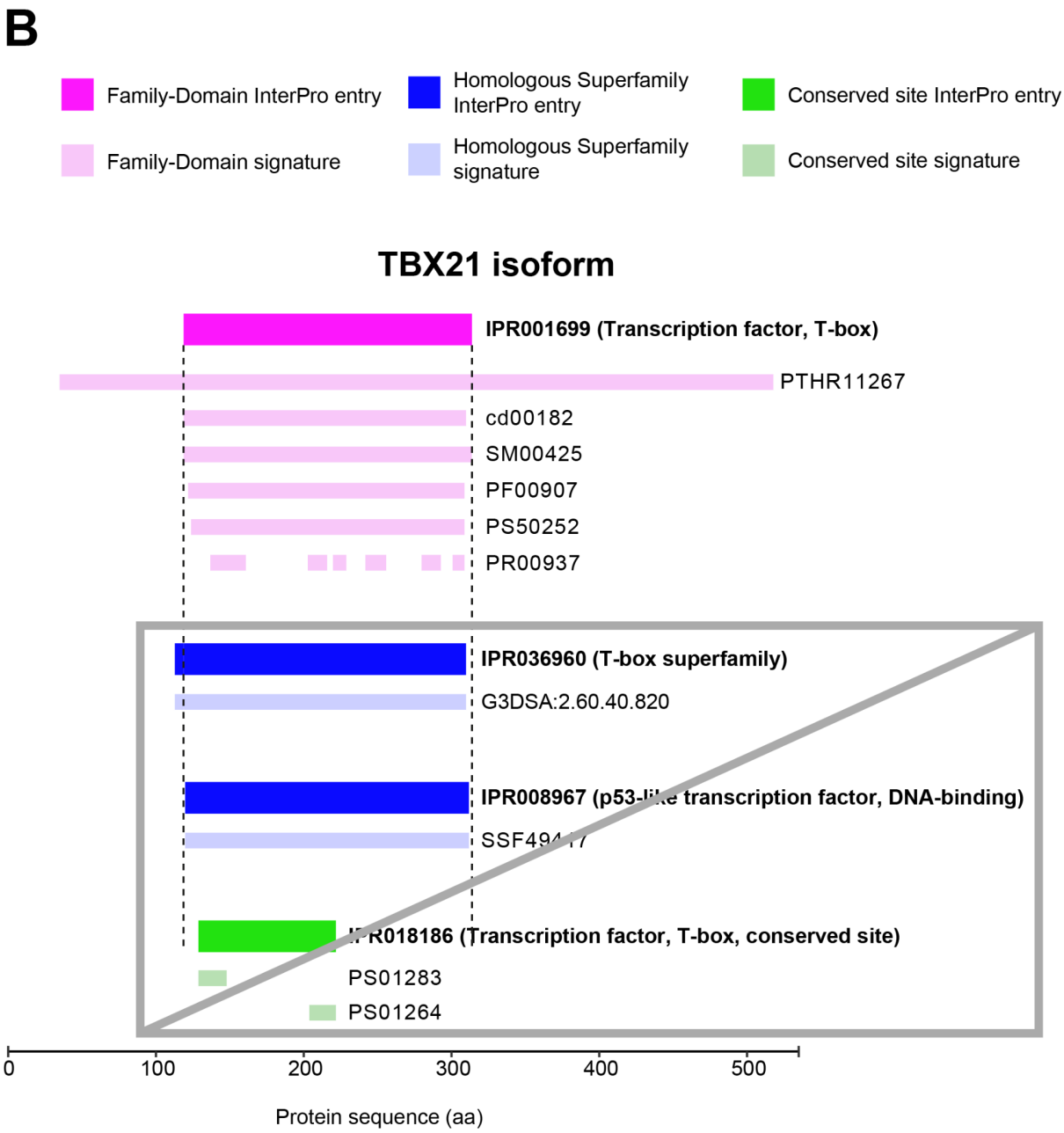


Fig. 3.4. Examples of the rank-based filtering of InterPro entry matches in TF isoforms. Dashed lines mark overlaps of higher-ranked InterPro entry matches with lower-ranked matches that are removed. Removed matches are marked by crossed grey boxes. **(A)** Isoform ENST00000336840 of PAX3. Domain (top-ranked) matches IPR001523 and IPR001356 overlap a Homologous Superfamily and a Conserved Site matches which are, consequently, removed. **(B)** Isoform ENST00000177694 of TBX21. The Family-Domain match IPR001699 (the top ranked match in this isoform) overlaps all other matches which are, consequently, removed.

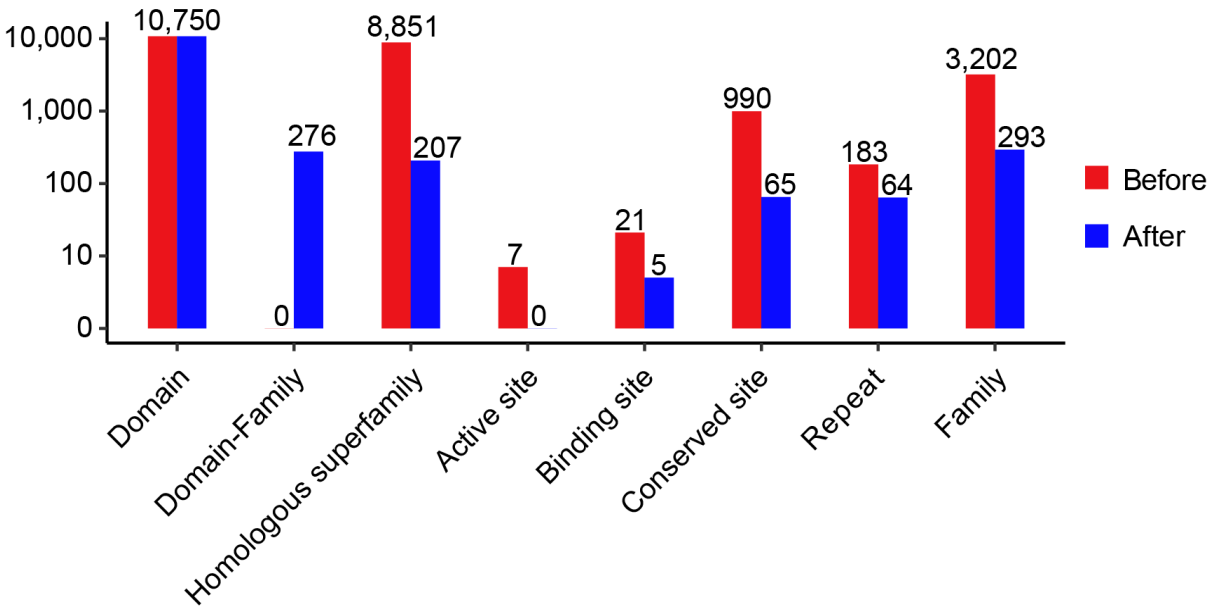


Fig. 3.5. The numbers of InterPro entry matches. The numbers of matches are presented by type, before (red) and after (blue) the rank-based filtering by overlaps.

5. As a match of an InterPro entry may contain several signatures that represent domains with a variable precision or do not represent domains at all, I selected a representative signature within each retained InterPro entry match to be able to accurately account for the presence and number of domains in each TF isoform. To this end, I ranked signatures in the following way, starting from the top priority: Pfam, SMART, ProSiteProfiles, CDD, SUPERFAMILY, Gene3D, PRINTS, ProSitePatterns, PIRSF, PANTHER. Of note, at the time of doing this analysis, I erroneously did not account for signatures of the other two types that were present in my scanning results, namely TIGRFAM (Haft et al., 2001, 2013; Li et al., 2021) and HAMAP (Pedruzzi et al., 2013, 2015). However, I later checked that the coordinates of matches of these signatures corresponded very well to the coordinates of Pfam, CDD and PrositeProfiles matches. Hence, the omission of the TIGRFAM and HAMAP signatures should not have affected my analysis. In the ranking described above, I prioritised signatures that denote domains: Pfam, SMART, ProSiteProfiles, CDD, SUPERFAMILY, Gene3D. I placed Pfam first, as from my experience its matches look the most accurate, and ranked the other domain signatures in the order of the decreasing number of their matches: SMART, ProSiteProfiles, CDD, SUPERFAMILY, Gene3D.

Next, I placed PRINTS and ProSitePatterns; they designate small conserved motifs which often mark a domain or a conserved site. Finally, I placed PIRSF and PANTHER at the end of the ranking because they do not designate domains but still help to recognise a protein as a member of a certain family. After obtaining this signature rank, I selected a representative signature from each InterPro entry match as a top one from the rank. For instance, if a matched entry contained matches from SMART, PRINTS and CDD, I chose the SMART signature as representative.

6. Next, I manually corrected coordinates of some retained matches to choose a signature that better represents them. For instance, upon a visual inspection, SMART had a more adequate representation of C2H2 ZFs and CCCH ZFs, than Pfam. Also, if a match was represented by a sequence of signatures from Gene3D, PRINTS or PROSITE patterns and was not a C2H2 zinc finger array, I merged these signature matches into one match by taking the outermost signature coordinates (Fig. 3.6). In this way, I obtained the coordinates of domains represented by a series of signatures. The total number of signature matches decreased from 95,769 to 29,544. Fig. 3.7 shows the numbers of signature matches, per signature type, before and after the selection of the representatives.

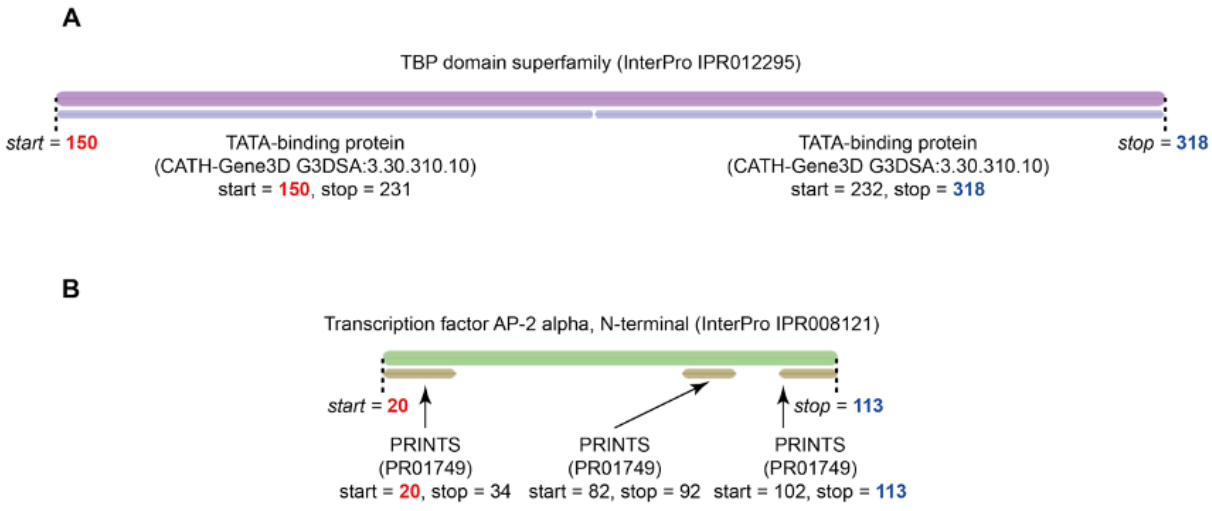


Fig. 3.6. Examples of merged signatures from Gene3D and PRINTS in non-C2H2 zinc finger matches. (A) Match *TBP domain superfamily* (InterPro entry IPR012295) is represented by two signatures from Gene3D database with coordinates [150; 231] and [232;

318]. These signatures matched two symmetrical halves of the TBP DBD; however, for my analysis, I need the TBP DBD represented by one match. Hence, I merged these signatures into one signature with coordinates [150; 318]. **(B)** Match *Transcription factor AP-2 alpha, N-terminal* (InterPro entry IPR008121) is represented by three signatures from PRINTS with coordinates [20; 34], [82; 92] and [102; 113]. Hence, I merged these signatures into one signature with coordinates [20; 113]. The schematics were adapted from the InterProScan online service (https://www.ebi.ac.uk/interpro/).

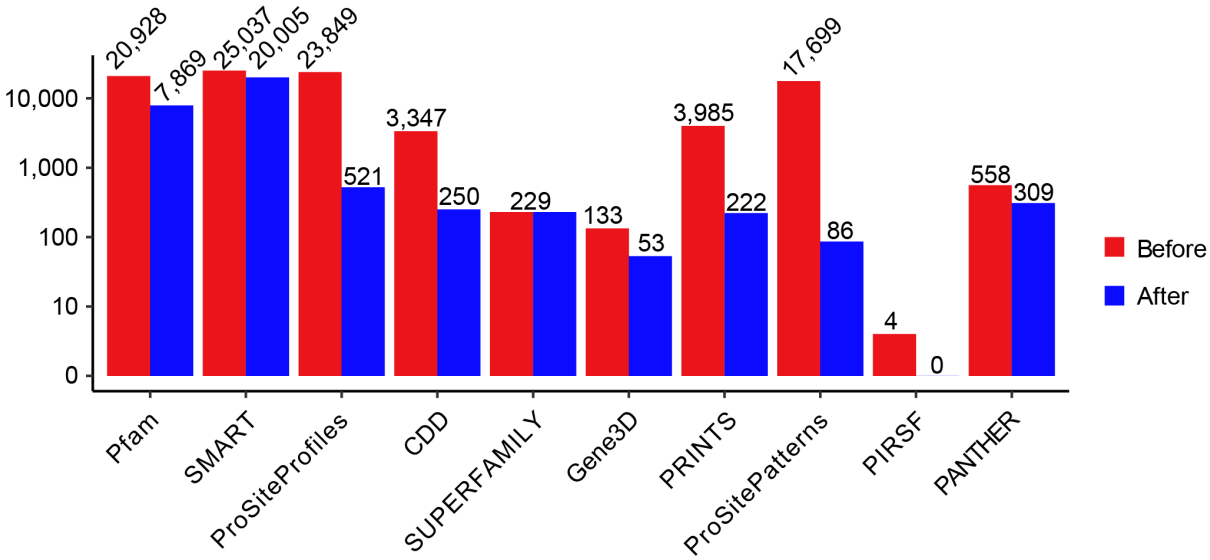


Fig. 3.7. The numbers of signature matches. The matches are presented by signature type, before (red) and after (blue) the selection of representative signatures.

3.2.2. Preliminary classification of matched domains and definition of TF families

Using the final set of domain matches and a custom Python script, I automatically classified the matched domains into two categories: (1) Domains without the word "DNA" in their InterPro entry name, description or associated GO terms (*Non-DBD* category), and (2) Domains with at least one mention of the word "DNA" in their InterPro entry name, description or associated GO terms (*DNA-related* category). Thus, using this initial classification, I manually split the *DNA-related* category into the *DBD*, *Other*, *Unknown* and *Unclear* categories (Fig. 3.8):

1. *DBD*: DNA-binding domain, according to its InterPro description (the total number of matches: 6,162).

2. *Other*: Domain with a function different from DNA binding, according to the InterPro description of the domain. For example, the description of the entry IPR001965 ("Zinc finger, PHD type") contains the word "DNA," because it includes a brief introduction to zinc finger domains in general, but the PHD zinc finger is not a DBD.
3. *Unknown*: Domain with an unknown function, as stated in its InterPro description. For instance, the description of the entry IPR003654 ("OAR domain") contains the word "DNA," because it lists several possible functions proposed for this domain, but then it states that the function of the domain is yet unknown.
4. *Unclear*: Domain whose function is not mentioned in its InterPro description. For example, one of the GO terms associated with the entry IPR029309 ("Calcium-responsive transcription factor"), namely GO:0003700 ("DNA-binding transcription factor activity"), contains the word "DNA," but the function of the domain is not mentioned in the entry description.
5. *Non-DBD*: The *Non-DBD* category that was formed automatically.

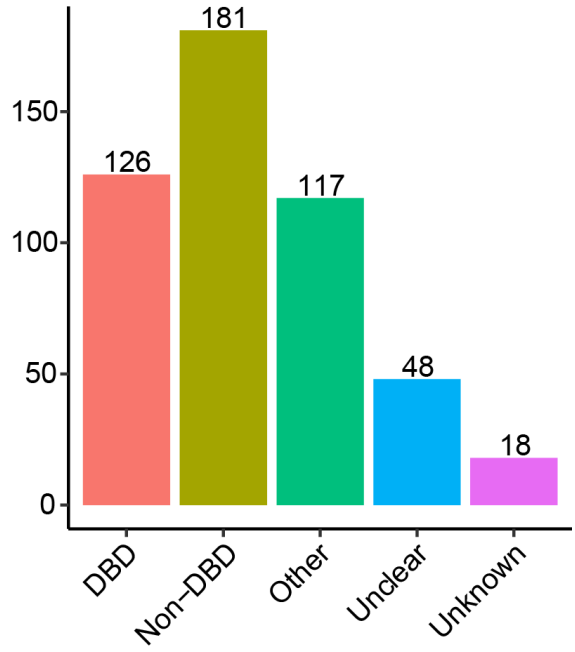


Fig. 3.8. The numbers of unique InterPro entries, by type.

The aim of this classification was to separate DBDs from all other domains and to prepare a list of non-DBDs (as represented by technical categories *Other* and *Non-DBD*) for further classification.

After obtaining the set of matched InterPro entries that correspond to DBDs (N = 126, matched 6,162 times in total), I further refined the set of DBDs and defined TF families based on the DBD types:

1. First, I manually corrected the coordinates of some DBD matches:
    a. I removed the remaining PANTHER matches as they did not represent domains, judging from their lengths.
    b. I removed C2H2 zinc finger ribbons which are two (IPR041661) or three (IPR031799) C2H2 zinc finger motifs matched as one domain, as well as similar dual matches of C4-type zinc fingers (IPR042153) and triple matches of CCCH-type zinc fingers (IPR041686). The reason for the removal was that these combined matches almost always coincided with separate matches of single zinc-finger motifs, and those combined matches that did not coincide with single motif matches likely represented partial zinc fingers.
    c. I removed the POU domain (IPR013847), as it always overlapped with a Homeobox and/or a POU-specific domain (as expected).
    d. I removed matches of the C-terminal part of the Homeobox engrailed domain (IPR019549), as this C-terminal match always coincided with a whole Homeobox match.
    e. In very short isoforms ENST00000509158 and ENST00000514807 of RBPJ from the CSL family, I removed the only two matches present there, “p53-like transcription factor, DNA-binding” (IPR008967) and “RBP-J/Cbf11/Cbf12, DNA-binding domain superfamily” (IPR037095), because these matches were quite general, and the CSL (LAG-1) DBD (“RBP-J/Cbf11/Cbf12, DNA binding,” IPR015351) did not match in these isoforms.
    f. In a short isoform ENST00000488107 of TFDP2 from the E2F family, I removed the “Winged helix-like DNA-binding domain superfamily” (IPR036388) match, because the entry that represents the E2F DBD (IPR003316, “E2F/DP family, winged-helix DNA-binding domain”) did not match in this isoform, and the IPR036388 match did not exactly correspond to a would-be E2F DBD fragment, according to matches present in longer isoforms.

g. In a short isoform ENST00000587646 of STAT5A (ENST00000587646) from the STAT family, I removed the "p53-like transcription factor, DNA-binding" (IPR008967) match, because it is quite general and does not correspond to a would-be match of the STAT DBD ("STAT transcription factor, DNA-binding," IPR013801), according to matches present in longer isoforms (for example, in STAT5A-201, ENST00000345506).

h. I joined disrupted DBD matches in several isoforms.

i. I restored relevant DBD matches lost during hierarchical filtering, namely: AT hook (IPR017956, SMART) in BAZ2A (isoforms ENST00000379441, ENST00000549884, ENST00000551812), C2H2C ZF (IPR002515, Pfam) in ST18 (isoform ENST00000276480) and C2H2C ZF (IPR002515, Pfam) in MYT1L (isoform ENSG00000186487).

j. I excluded InterPro entries and signatures that were integrated to InterPro by mistake (as I clarified with InterPro support team), namely: (1) IPR037220 which is not a "Zinc finger BED domain" and should have been removed from InterPro entries in version 82.0 (and in version 90.0 it is indeed not present); (2) Signature SM00614 from IPR003656 ("Zinc finger, BED-type") which does not correspond to a BED ZF, because, according to the InterPro support team, it "overhits" many proteins, in comparison to other signatures integrated in this entry. Signature SM00614 should have been removed from InterPro entries in version 82.0 (although, in version 90.0 it is still present).

2. Secondly, I excluded general DBD matches, because all of them corresponded to specific DBD matches with clear functions. However, I still retained both general and specific non-DBD matches for further analysis (see below).
3. Next, in each set of overlapping specific DBD matches, I retained the longest match, in this way removing redundant DBD matches. After the three refinement steps described above, the number of DBD InterPro entry matches decreased from 6,162 to 5,467.
4. After the refinement of the DBD InterPro entries, I grouped entries that corresponded to the same type of a DBD (for example, a Homeodomain or a C2H2 zinc finger) and, based on this grouping, I formed TF families. Usually,

these families are defined according to the DBDs present in the constituent TFs (Weirauch and Hughes, 2011). Consequently, I defined a **TF family** as a set of TFs where each TF has the same set of DBDs. For example, a **monotypical** Homeodomain family is a set of TFs whose only DBD is a Homeodomain, while a **combined** "C2H2 ZF; Homeodomain" family is a set of TFs with at least one C2H2 zinc finger and at least one homeodomain, but no other DBDs. Additionally, I removed genes with an uncertain TF status in the HumanTFs database; also, I removed genes without a clear DBD.

3.2.3. Selection of TF and non-TF isoforms

After the domain preprocessing described above, I formed the final set of protein TF isoforms for further analysis. To this end, I concatenated a list of protein TF isoforms with at least one domain match with a list of isoforms without domain matches. Next, I filtered the resulting list of isoforms by the quality of isoform annotation in Ensembl which is summarised by transcript flags (https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html). Specifically, I retained only isoforms with the two highest Transcript Support Levels (TSLs) 1 and 2, isoforms with an undefined TSL (TSL = NA) and isoforms with any value of TSL but with a defined MANE Select flag. The reason for retaining isoforms with an undefined TSL was the fact that such isoforms are not necessarily speculative but simply were not considered for the TSL analysis. These isoforms are annotated in pseudogenes, or arise from human leukocyte antigens, immunoglobulin genes or T-cell receptors, or contain only one exon. The reason for retaining isoforms with the MANE Select flag independently of the TSL was the fact that, by definition, such isoforms are identical in two human genome annotations that are made independently – Ensembl and NCBI RefSeq – and hence are likely to be real and correctly annotated (https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html).

After the isoform selection step, I obtained the final list of 4,134 isoforms produced by 1,512 TFs. See the domain scanning results for these TF isoforms in Supplementary Table 1. The table additionally contains 265 isoforms of 98 TFs classified into the technical families "Excluded" (TFs that I excluded from my analysis during the curation procedures described above) and "Unknown" (genes that do not have a known DBD, according to both the HumanTFs database and my analysis). Therefore, in total, Supplementary Table 1 contains the scanning results for the 4,399 isoforms of 1,610 TFs.

This isoform selection step and the domain preprocessing described in sections 3.2.1 and 3.2.2 allowed me to analyse the production of alternative TF isoforms, which I describe in the current chapter, and to analyse DBDs across TF isoforms in Chapter 4.

Finally, I selected coding isoforms (N = 48,116) produced by non-TF genes (N = 19,631) using the same filtering procedure based on the transcript quality flags. I used these non-TF isoforms to compare the production of alternative isoforms by the TF and non-TF human genes.

3.3. Results

3.3.1. Transcription factors produce more alternative isoforms than other coding genes

In (Talavera et al., 2009; Taneri et al., 2004), the authors showed that the proportion of genes producing alternative isoforms is higher among TFs, than non-TFs, and that TFs tend to produce more alternative isoforms. Because these results were obtained more than a decade ago, I decided to evaluate if they still hold with the current human genome annotation. Also, neither of the two articles accounted for possible confounding factors that may affect the rate of alternative splicing in TF and non-TF genes, such as the gene length (Grishkevich and Yanai, 2014) and the number of exons. I therefore supposed that the longer the gene and the more exons it has, the more isoforms it could produce. Hence, I evaluated if the previous findings about alternative isoform production by TF genes still hold when accounting for the gene length and the number of exons.

For this and further analyses, I selected 1,512 TFs from the HumanTFs database (Lambert et al., 2018) that were also annotated in the Ensembl database (Cunningham et al., 2022) release 99. These TFs had 4,134 protein-coding isoforms with a high-quality annotation. To compare the production of alternative isoforms by the TF and non-TF genes, I also selected 19,631 non-TF protein-coding genes annotated in Ensembl release 99. The selected non-TFs had 48,116 coding isoforms with a high-quality annotation.

Counting TFs and non-TFs with more than one isoform, I found that significantly more TFs (66.5%), than non-TFs (62.5%), produced alternative isoforms (p = 0.0024, two-sided Fisher's exact test), which supports the findings by (Taneri et al., 2004) and (Talavera et al., 2009).

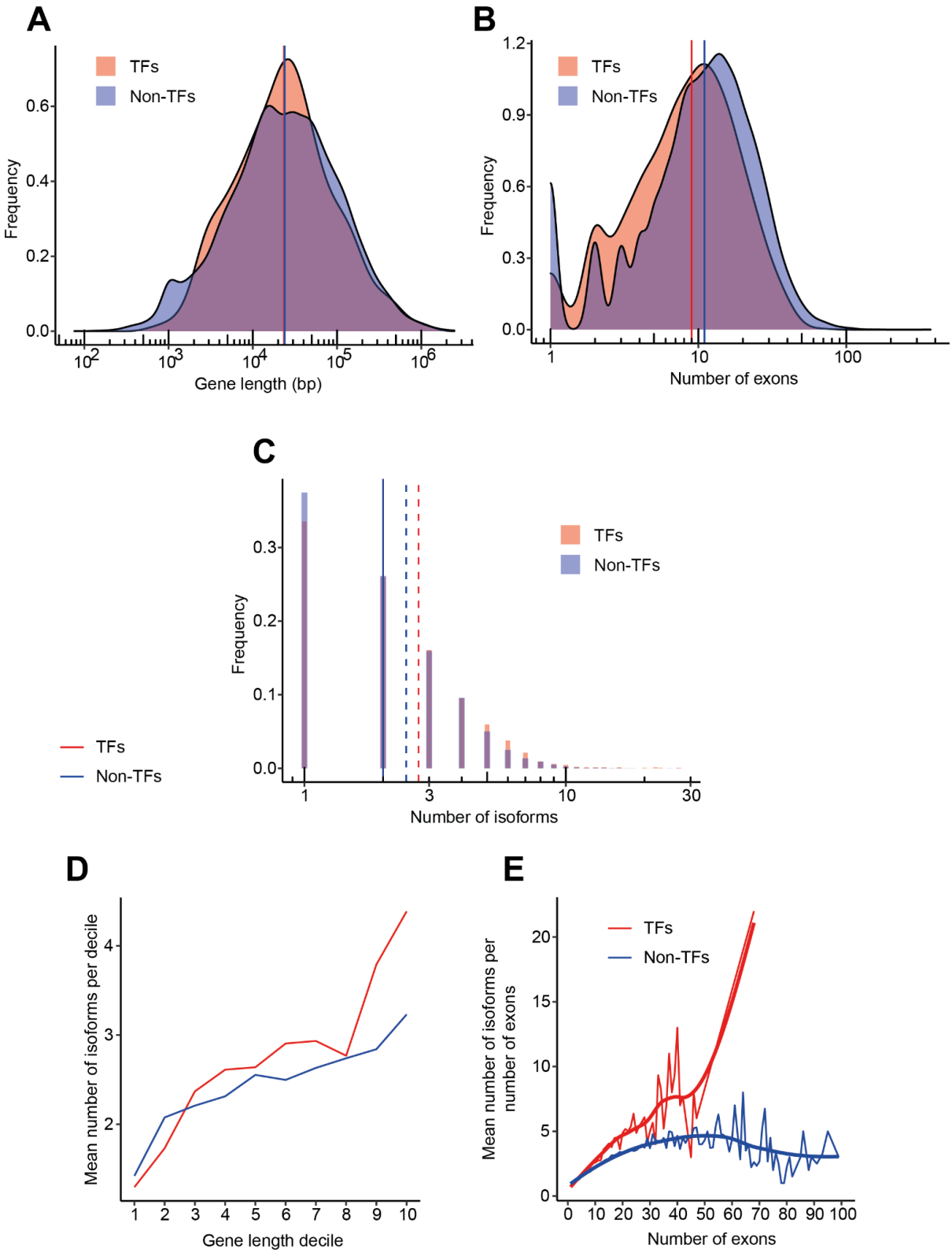


Fig. 3.9. TFs produce more isoforms than non-TF coding genes. Log-scaled distributions of the gene length **(A)** and the number of exons **(B)** in TFs (red) and non-TFs (blue). Vertical lines denote the medians of the distributions (the red line for TFs and the blue line for non-TFs). **(C)** The log-scaled distributions of the number of isoforms produced by TFs (red) and

non-TFs (blue). Vertical lines of the corresponding colours denote the distribution medians (the overlapping solid lines) or means (the dashed lines). **(D)** The mean number of isoforms per gene length decile for TFs (in red) and non-TFs (in blue). **(E)** The mean number of isoforms for every possible number of exons, up to 100, for TFs (in red) and non-TFs (in blue). The plots in (E) were LOESS-smoothed.

Next, I tested if the average number of isoforms produced by TFs is greater than the average number of isoforms produced by non-TF coding genes. For this, I first studied the distributions of the gene length and the number of exons in the two groups of genes. I observed similar gene length distributions for TFs and non-TFs (Fig. 3.9A) with indistinguishable medians. Specifically, the TF median gene length was 23.4 kb, while the non-TF median gene length was 23.9 kb. In contrast, the distributions of the number of exons for TFs and non-TFs were significantly shifted relative to each other ($p < 2.2 \cdot 10^{-16}$, Wilcoxon rank sum test with continuity correction), as their medians indicated (Fig. 3.9B). Namely, the median number of exons in TFs was 9, while the median number of exons in non-TFs was 11.

Consequently, as the median gene length was the same in TFs and non-TFs and the median number of exons in TFs was smaller, I expected TFs to have a smaller median number of isoforms. However, Fig. 3.9C shows that the median number of isoforms produced by a TF is exactly the same as the median number of isoforms produced by a non-TF coding gene and equals 2. Importantly, the two distributions are significantly different ($p = 3.5 \cdot 10^{-5}$, Wilcoxon rank sum test with continuity correction), and their shapes demonstrate that a single isoform is more frequently produced by non-TFs, while 5 or more isoforms are more characteristic of TFs. This is reflected by the fact that the mean number of isoforms produced by a TF (2.73) is greater than the mean number of isoforms produced by a non-TF (2.45; see Fig. 3.9C).

To find if the production of a greater number of isoforms by TFs depended on the gene length, I grouped TFs and non-TFs into gene length deciles and calculated the mean number of isoforms per TF and non-TF in each decile (Fig. 3.9D). I found that from decile 3 onwards, TFs produced on average more isoforms than non-TFs, and deciles 1 and 2 (consisting of very short genes) were the only ones where non-TFs produced on average more isoforms than TFs (Fig. 3.9D). In other words, TFs produced more isoforms than non-TFs for any gene length, apart from a very small one. Additionally, with the increasing gene length, the mean number of isoforms grew faster in TFs, than in non-TFs (Fig. 3.9D).

Next, I investigated if the production of a greater number of isoforms by TFs held for any number of exons. To this end, I counted isoforms produced by TFs and non-TFs with every

possible number of exons and limited the analysis to 100 exons per gene (Fig. 3.9E), as only 13 coding genes out of 21,143 have more exons than that. I observed that for the vast majority of exon counts shared between TFs and non-TFs (from 1 to ~70), TFs produced on average more isoforms, and the number of isoforms grew considerably faster for TFs, than for non-TFs, with the increasing number of exons (Fig. 3.9E). Interestingly, after ~50 exons, the smoothed mean number of isoforms produced by non-TFs ceased to grow and started to decrease (Fig. 3.9E). Finally, of note, there is a TF with >20 isoforms and ~70 exons in Fig. 3.9E. It is TCF4, which was previously shown to co-express many different isoforms across many adult human tissues; some of its isoforms were shown to be expressed more broadly than others (Grajkowska et al., 2017; Sepp et al., 2011).

In total, my results support and extend previous findings that the proportion of alternatively spliced TFs is higher than the proportion of alternatively spliced non-TFs and that TFs produce more isoforms than non-TFs. Motivated by these results, in the next section I survey literature to form the basis for a genome-wide computational study of human TF isoforms.

3.3.2. Compendium of human transcription factors with studied isoforms provides a basis for genome-wide predictions

To study isoforms of human TFs at the scale of the whole genome, I compiled from the literature a collection of 24 human TFs whose alternative isoforms were studied in detail experimentally and were shown to have particular functions explained by the presence of certain domains (Fig. 3.10; Tables 1-2).

First, I considered functionally characterised examples of differences in DBDs between alternative isoforms of the same TF. First, IKZF1 can change the number of its DBDs (C2H2 zinc fingers), while CREM swaps one DBD sequence for another (both are basic domains) between its isoforms. Secondly, PAX6, PAX8 and FOXP1 express alternative isoforms with partially different DBD sequences (Fig. 3.10A). All these differences lead to changes in the DNA-binding specificity of the respective TFs (hereafter, for each example TF, see the references to the supporting literature in Tables 1-2).

Next, I found three examples of TFs, namely IKZF1, IKZF2 and HOXA1, that express DBD+ and DBD- alternative isoforms, both with a dimerization domain (Fig. 3.10B). When only a DBD+ isoform is expressed, the monomers of these TFs can dimerize, bind DNA and regulate transcription; on the contrary, when both a DBD+ and a DBD- isoforms are expressed, they also dimerize but cannot bind DNA due to the lack of a DBD in one of the two monomers.

Hence, the DBD- isoform sequesters the DBD+ one from the chromatin and in this way reverts the transcription regulation mode of the DNA-bound dimer (Fig. 3.10B).

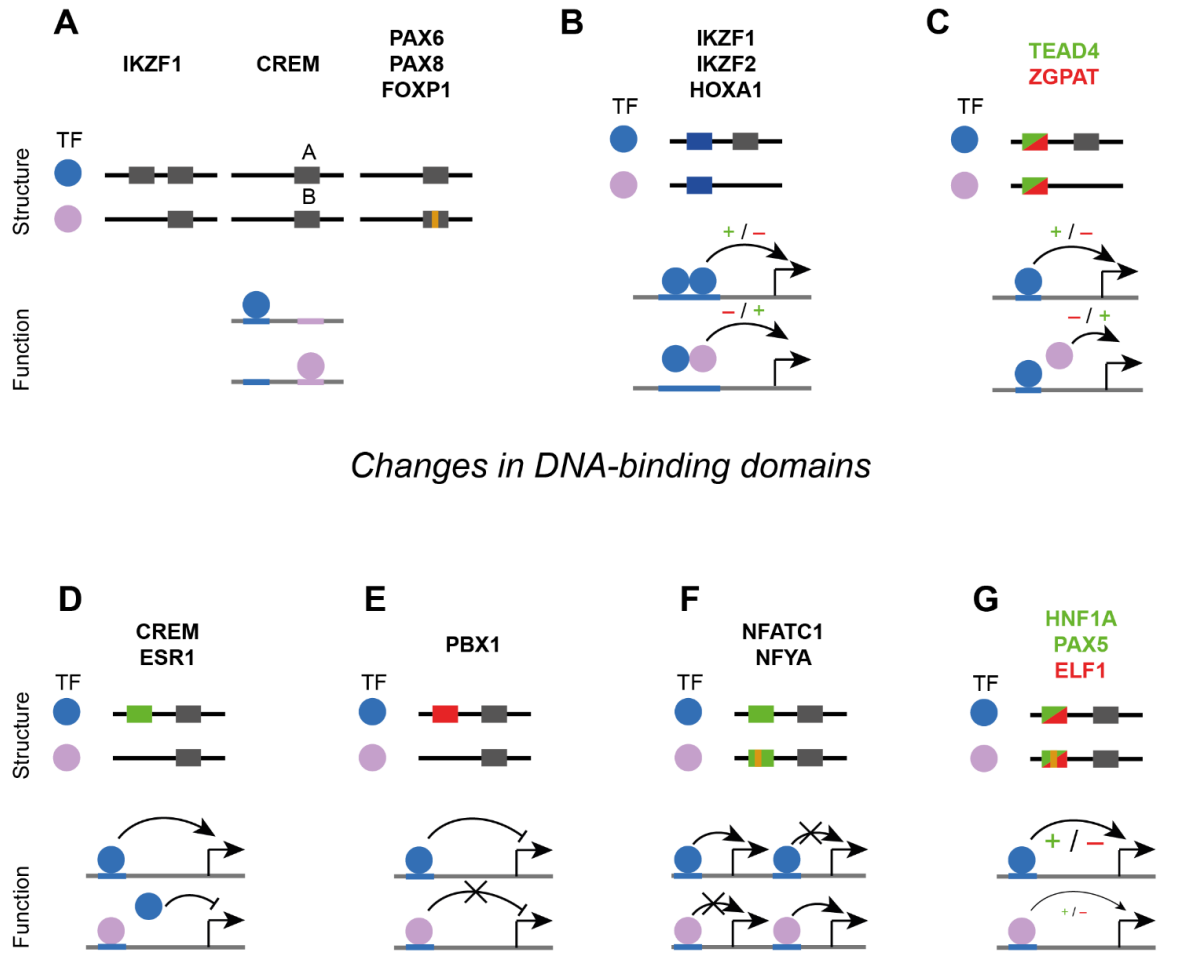


Fig. 3.10. Experimentally studied functional differences in alternative isoforms of human TFs. Part 1. The schematics show how differences in DBDs (**A-C**) or transcription regulation domains (**D-G**) between alternative isoforms of the same TF affect the function of the TF. CREM may not co-express both isoforms, but the one lacking the transcription activation domain (ICER) still represses the transcription. The DBD- isoforms of IKZF2 were shown to heterodimerize with IKZF1 and also to attenuate transcription repression exerted by the DBD+ isoforms of IKZF2 (Zhao et al., 2016). Hence, I assume that this attenuation of the transcription repression exerted by the IKZF2 DBD+ isoforms also comes from the dimerization with the IKZF2 DBD- isoforms. Additionally, the IKZF2 DBD- isoforms bind components of the histone deacetylase complex Mi-2β/NuRD (Zhao et al., 2016) and consequently may perturb transcription repression also in this way. Hence, I could additionally depict IKZF2 on panel C.

Finally, two TFs of the opposite modes of action, an activator TEAD4 and a repressor ZGPAT, also can express a DBD+ and a DBD- isoform, and both isoforms would have a transcription activation (or repression) domain (Fig. 3.10C). Consequently, when only a DBD+ isoform is expressed, these TFs can bind DNA and activate (or repress) transcription; however, when a DBD+ and a DBD- isoform are co-expressed, the DBD- isoform, which cannot bind DNA, sequesters coactivators (or corepressors) from the chromatin and in this way precludes the transcription activation (or repression) by the DNA-bound DBD+ isoform (Fig. 3.10C). Hence, a complete loss of a DBD in the presence of a dimerization or transcription regulation domain can switch the transcription regulation mode of a TF from activation to repression and vice versa.

Apart from differences in DBDs, alternative isoforms of TFs may have differences in transcription regulation domains, and such examples have also been experimentally studied. Specifically, CREM and ESR1 express DBD+ isoforms with or without a transcription activation (TA) domain (Fig. 3.10D). As a consequence, when these TFs express only the TA domain-containing (TA+) isoform, it binds DNA and activates transcription; however, when a TA+ and a TA- isoform are co-expressed, the TA- isoform competes with the TA+ isoform for DNA binding and hence precludes transcription activation (Fig. 3.10D). Similarly, when PBX1 expresses a DBD+ isoform without a transcription repression domain, it binds to the DNA but cannot exert transcription repression (Fig. 3.10E). Additionally, partial differences in the sequence of a transcription regulation domain can play a functional role (Fig. 3.10F-G). Specifically, transcription activators NFATC1 and NFYA can express isoforms with alternative fragments within their TA domains. Consequently, the resulting isoforms activate

different sets of genes (Fig. 3.10F). Similarly, transcription activators HNF1A and PAX5 and a repressor ELF1 encode a peptide that is alternatively included in their transcription regulation domains; those isoforms that contain the peptide exert an attenuated effect on transcription in comparison to isoforms that lack the peptide (Fig. 3.10G). In sum, the differential presence or sequence of a transcription regulation domain can facilitate various functional differences between respective alternative TF isoforms: the reversal, attenuation or removal of the regulation effect, as well as a change in the set of regulated genes.

In the examples described above, differences between alternative isoforms of TFs occurred in the most common types of TF domains: DBDs and transcription regulation domains. However, in other instances, these differences may lie in more rare domains, namely a ligand-binding, degradation-promoting, DBD inhibition or activation domain, and such examples have also been studied experimentally (Fig. 3.11). First of all, the glucocorticoid receptor, NR3C1, can express isoforms with or without a ligand-binding domain (LBD) (Fig. 3.11A). Consequently, when it expresses only an LBD+ isoform, it homodimerises and binds to the DNA upon the binding of two molecules of a glucocorticoid. In contrast, when NR3C1 co-expresses an LBD+ and an LBD- isoforms, they heterodimerise and bind DNA, but only one of them binds the ligand, which is not enough for transcription regulation (Fig. 3.11A).

IRF5 presents another example of an alternative region outside DNA-binding or transcription regulation domains. Specifically, it codes for a domain that promotes the degradation of an isoform that contains it (Fig. 3.11B). Thus, when this region is included into an isoform, the isoform undergoes rapid degradation and hence does not take part in transcription regulation; in contrast, when the degradation-promoting domain is omitted, the protein product of IRF5 is stable and activates transcription (Fig. 3.11B). Two more examples of differences outside a DBD or a transcription regulation domain are found in TCF3 and MITF. Thus, TCF3 alternatively encodes a peptide that inhibits the DNA binding of the isoform that contains it (Fig. 3.11C). Consequently, when the peptide is included, the corresponding TCF3 homodimer cannot bind DNA; whereas, when the DBD inhibition domain is omitted, the respective homodimer is able to bind DNA (Fig. 3.11C). Similarly, MITF encodes a domain that promotes DNA binding, and hence when the domain is included the binding is stable, and vice versa (Fig. 3.11D).

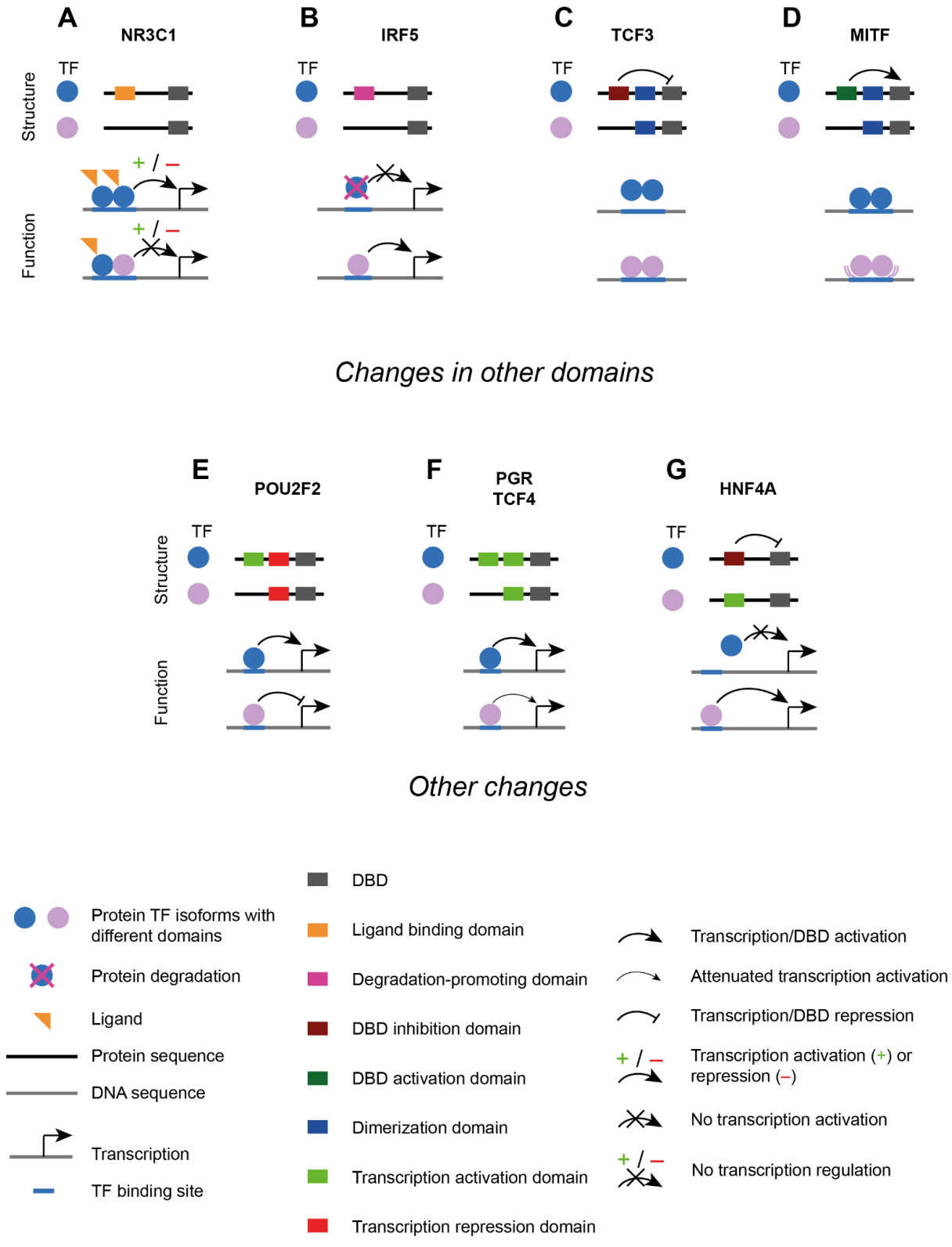


Fig. 3.11. Experimentally studied functional differences in alternative isoforms of human TFs. Part 2. Schematics showing how differences in other domains (**A-D**) or more complex differences in transcription regulation domains (**E-G**) between alternative isoforms of the same TF affect the function of the TF.

Finally, I found several examples of more complex differences involving transcription regulation domains (Fig. 3.11E-G). First, POU2F2 encodes both a transcription activation and a repression domain, but the activation domain is stronger (Fig. 3.11E). Consequently, when the activation domain is included, the TF is an activator, but when it is excluded, the TF becomes a repressor (Fig. 3.11E). Similarly, PGR and TCF4 encode two transcription activation domains, but one of them is included alternatively (Fig. 3.11F). Hence, an isoform that contains both domains activates transcription stronger than an isoform that contains only one of the two domains (Fig. 3.11F). Finally, HNF4A swaps a transcription activation domain for a DBD inhibition one, in this way switching the transcription activation function of the TF on and off (Fig. 3.11G).

Overall, these examples of experimentally validated functions of alternative TF isoforms will help me to predict the functions of *all* TF isoforms genome-wide. In addition, in Table 3 I compiled a collection of another 68 human TFs or groups of TFs (as some publications study small TF subfamilies) with multiple isoforms. I did not use alternative isoforms of these TFs for my functional predictions either because the connection between the domains and the functions of the isoforms was not clear, or because I found the corresponding publications later on.

Because functional data about the alternative isoforms of human TFs is scattered in the literature, I hope that this total collection of 92 TFs and groups of TFs serves as a convenient reference point for future studies of human TF isoforms. Moreover, the fact that tens of human TFs have been shown to produce functional alternative isoforms underscores the importance of studying alternative isoforms of *all* human TFs. In the following sections, I address this problem computationally, using this compendium.

Table 1. Experimentally studied human TF isoforms with functional differences explained by domain differences. I use this information in the analysis in Chapters 4 and 5. I apologise to those authors whose work on alternative isoforms of human TFs I do not cite because I did not find it when compiling the table.

Human orthologue	Figure	Domain change	Functional effect	Publication
CREM	3.10A	Exchange of a DBD for another DBD	Changed specificity of DNA binding	(Laoide et al., 1993)
CREM	3.10D	Loss of a transcription activation domain in presence of a DBD	Transcription repression	(Molina et al., 1993)
ELF1	3.10G	Change in the sequence of a transcription repression domain in presence of a DBD	Changed strength of transcription repression	(Nishiyama et al., 2000)
ESR1	3.10D	Loss of a transcription activation domain in presence of a DBD	Transcription repression	(Figtree et al., 2003)
FOXP1	3.10A	Change in a DBD sequence	Changed specificity of DNA binding	(Gabut et al., 2011)
HNF1A	3.10G	Change in the sequence of a transcription activation domain in presence of a DBD	Changed strength of transcription activation	(Bach and Yaniv, 1993)
HNF4A	3.11G	Exchange of a DBD inhibition domain for a transcription activation domain in presence of a DBD	Transcription activation	(Lambert et al., 2020)

HOXA1	3.10B	Loss of a DBD in presence of a dimerisation domain	Reverse transcription regulation	(Fernandez and Gudas, 2009)
IKZF1	3.10A	Change in the number of DBDs	Changed specificity of DNA binding	(Hahm et al., 1994)
IKZF1	3.10B	Loss of a DBD in presence of a dimerisation domain	Reverse transcription regulation	(Sun et al., 1996)
IKZF2	3.10B	Loss of a DBD in presence of a dimerisation domain	Reverse transcription regulation	(Zhao et al., 2016)
IRF5	3.11B	Loss of a degradation domain in presence of a DBD	Transcription activation	(Lazzari et al., 2014)
MITF	3.11D	Loss of a DBD activation domain in presence of a DBD	Unstable DNA binding	(Hemesath et al., 1994)
NFATC1	3.10F	Change in the sequence of a transcription activation domain in presence of a DBD	Changed set of activated genes	(Lucena et al., 2016)
NFYA	3.10F	Change in the sequence of a transcription activation domain in presence of a DBD	Changed set of activated genes	(Dolfini et al., 2012; Li et al., 1992)
NR3C1	3.11A	Loss of a ligand-binding domain in presence of a DBD	Loss of transcription regulation	(Oakley et al., 1999)
PAX5	3.10G	Change in the sequence of a transcription activation domain in presence of a DBD	Changed strength of transcription activation	(Robichaud et al., 2004)
PAX6	3.10A	Change in a DBD sequence	Changed specificity of DNA binding	(Epstein et al., 1994)

PAX8	3.10A	Change in a DBD sequence	Changed specificity of DNA binding	(Kozmik et al., 1997)
PBX1	3.10E	Loss of a transcription repression domain in presence of a DBD	Loss of transcription repression	(Asahara et al., 1999)
PGR	3.11F	Loss of a transcription activation domain in presence of other transcription activation domains and a DBD	Lowered strength of transcription activation	(Takimoto et al., 2003)
POU2F2	3.11E	Loss of a strong transcription activation domain in presence of a weak transcription repression domain and a DBD	Switch from transcription activation to transcription repression	(Lillycrop et al., 1994)
TCF3	3.11C	Loss of a DBD inhibition domain in presence of a DBD	Gain of DNA binding	(Sun and Baltimore, 1991)
TCF4	3.11F	Loss of a transcription activation domain in presence of other transcription activation domains and a DBD	Lowered strength of transcription activation	(Sepp et al., 2011)
TEAD4	3.10C	Loss of a DBD in presence of a transcription activation domain	Transcription repression	(Qi et al., 2016)
ZGPAT	3.10C	Loss of a DBD in presence of a transcription repression domain	Transcription activation	(Yu et al., 2010)

Table 2. Experimental details about TFs from Table 1 and their literature-based isoform expression patterns.

Human orthologue	Figure	Studied TF	Organism	Expression model	Relevance to human	Literature-based isoform expression pattern
CREM	3.10A	Crem	Mouse	Cell lines	Two long isoforms of human CREM, ENST00000354759 and ENST00000342105, have different DBDs and in this way correspond to murine Crem isoforms alpha and beta.	(Laoide et al., 1993) showed that murine Crem did not swap its mRNA isoforms across examined cell lines but rather co-expressed two or three different mRNA isoforms in some cell lines. The set of co-expressed isoforms always included a full isoform, with the P site, activation domains and a DBD. The full isoform is the major one, judging from the bands in their fig. 1B. (Stehle et al., 1993) found that in rat, at night, the pineal gland co-expresses mRNAs of two ICERs (DBD+ isoforms lacking all other domains) with two different DBDs, and that during the night it does not express any other Crem isoforms. They additionally showed that these mRNAs are translated in the pineal gland.
CREM	3.10D	Crem	Mouse; Rat	Cell line and primary tissues	Isoforms with the activation domain and a DBD and isoforms with just a DBD (the latter are called ICERs) are also expressed in the human; see, for example, (Gellersen et al., 1997).	(Molina et al., 1993) showed that mRNAs of ICERs (DBD+ isoforms of CREM that lack all non-DBDs) are expressed in several different cell lines. Additionally, (Stehle et al., 1993) found that in rat, the pineal gland co-expresses ICER proteins with two different DBDs at night, and at that time it does not express any other

						CREM isoforms. Also, in mouse and human, pre-meiotic germ cells express only ICER mRNA, while post-meiotic germ cells switch to expressing only activating CREM isoforms (Foulkes et al., 1992).
ELF1	3.10G	Elf1	Rat	Cell lines	(Nishiyama et al., 2000) found the same ELF1 isoforms in human.	(Nishiyama et al., 2000), using a panel of human cDNA libraries with two independent libraries for some tissues, showed that at least two ELF1 mRNA isoforms encoding different N termini were co-expressed in various human tissues and that the same tissue could express different major ELF1 isoforms.
ESR1	3.10D	ESR1	Human	Cell lines	Human TF	(Figtree et al., 2003) showed that the MCF7 cell line expressed both the full and the short protein isoforms of ESR1, while the EA.926 cell line expressed only the short protein isoform. This isoform lacks the transcription activation domain.
FOXP1	3.10A	FOXP1	Human	Cell lines	Human TF	(Gabut et al., 2011) showed that the human FOXP1 switches its major mRNA isoform with one DBD sequence for another major mRNA isoform with a different DBD sequence when human ESCs start to differentiate into neural progenitor cells (NPCs).
HNF1A	3.10G	HNF1A	Human	Cell lines	Human TF	(Bach and Yaniv, 1993) showed that the

						adult human liver and thymus co-express three HNF1A mRNA isoforms encoding different sequences of the C-terminal transcription activation domain. The authors also demonstrated that the longest mRNA isoform is the major one in all adult human tissues that they examined and where the gene was expressed (namely, in the liver, kidney, intestine and thymus). Additionally, (Harries et al., 2006) showed that all the three human HNF1A mRNA isoforms are co-expressed in liver, kidney, pancreas and isolated islets. However, HNF1A switched its major mRNA isoforms between two groups of tissues: (1) liver and kidney, where the longest isoform was the major one, and (2) adult pancreas and isolated islets where a shorter isoform became major. Interestingly, they also showed a switch of HNF1A major mRNA isoforms between the fetal pancreas (where the longest isoform is the major one) and the adult pancreas (where a shorter isoform is major).
HNF4A	3.11G	HNF4A	Human	Cell lines	Human TF	(Lambert et al., 2020) suggest that three human mRNA isoforms transcribed from the promoter P1 (α1, α2 and α3) have a transcription activation domain, while another three isoforms transcribed from the same promoter (α4, α5 and α6), due to

						alternative splicing, swap this domain for a DBD inhibition domain. As this work is quite recent, literature still does not have data on differential expression of the first trio of isoforms versus the second trio in human tissues. Previous literature either compared protein isoforms transcribed from the first and the second promoters of the murine or human gene (Chellappa et al., 2016; Vuong, 2014) or, in addition, did compare different mRNA isoforms of the human gene that were produced from promoter P1 but the N terminus of these isoforms was annotated differently (Harries et al., 2008).
HOXA1	3.10B	Hoxa1	Mouse	Cell lines	(Hong et al., 1995) found that human HOXA1 also has a long DBD+ and a short DBD- isoforms; however, they do not mention dimerization.	(Murphy and Hill, 1991) showed that both the DBD+ and the DBD- mRNA isoforms are expressed in the murine embryogenesis.
IKZF1	3.10A	Ikzf1	Mouse	Cell lines	(Molnár et al., 1996) found that IKZF1 isoforms are highly conserved between human and mouse.	(Hahm et al., 1994) observed Ikzf1 protein isoforms with different numbers of C2H2 ZFs and different DNA binding specificities co-expressed in B and T lymphoid murine cell lines.
IKZF1	3.10B	Ikzf1	Mouse?	Cell lines	(Molnár et al., 1996) found that IKZF1 isoforms are highly conserved between human and mouse.	DBD+ and DBD- mRNA isoforms of IKZF1 are co-expressed in murine B and T cell progenitors and also in some subpopulations of myeloid cells (Klug et

						al., 1998); they are also co-expressed in human cord blood and bone marrow (Payne et al., 2001).
IKZF2	3.10B	IKZF2	Human	Cell lines and primary tissues	Human TF	(Zhao et al., 2016) showed that IKZF2 has DBD- mRNA isoforms that are specific to primary human T cells and myeloid leukemic cells.
IRF5	3.11B	IRF5	Human?	Cell lines and primary tissues	Human TF?	(Stone et al., 2013) showed that all four human IRF5 mRNA isoforms are co-expressed in monocytes both in normal physiology and systemic lupus erythematosus and, if translated, give rise to proteins of different stability and transcription activation potential.
MITF	3.11D	Mitf	Mouse	*In vitro* experiments	(Hemesath et al., 1994) showed that the DBD activation domain is a 6-amino acid peptide located closely upstream of the bHLH and encoded by an alternative exon in murine Mitf. Later on, (Amae et al., 1998) showed that human MITF encodes two different isoforms, MITF-A and MITF-M, that, among other differences, have an indel of 6 amino acids (ACIFPT) closely upstream of the bHLH. Finally, from (Steingrímsson et al., 1994), fig. 5b in their article,	MITF has several isoforms. Isoform MITF-M has a DBD activation site, while MITF-A/H do not have it (Yasumoto et al., 1998). (Amae et al., 1998) showed that human MITF switches major mRNA isoforms between (1) melanoma cells where MITF-M is the major isoform and (2) HeLa and A-172 cell lines and kidney primary tissue where MITF-A/H are the major isoforms. Additionally, (Hershey and Fisher, 2005) demonstrated that the same switch of MITF mRNA isoforms occurs between melanoma cells and primary human osteoclasts. (Bharti et al., 2008) showed the co-expression of

					one can see that the alternative murine exon encodes exactly the same peptide (ACIFPT) via using the last nucleotide of the previous exon as the start of the first triplet (for alanine). Hence, I suppose that the murine Mitf isoform with the peptide corresponds to the human isoform MITF-M with the same peptide, while the murine Mitf isoform without the peptide corresponds to the human isoform MITF-A without the peptide.	different murine MITF protein isoforms in developing choroid and retina.
NFATC1	3.10F	Nfatc1	Mouse	Cell lines	(Lucena et al., 2016) found the same two isoforms in human Burkitt lymphoma samples.	(Sherman et al., 1999) showed that NFATC1 mRNA isoforms with different transcription activation domains are co-expressed in murine T cells and mast cells and that one of them (“alpha”) is upregulated in response to cell activation. They also showed that the other mRNA isoform (“beta”) is expressed in the murine heart, spleen and kidney and that in the spleen it is co-expressed with “alpha” which is minor in that tissue.
NFYA	3.10F	Nfya	Mouse	Cell lines	(Li et al., 1992) found that both human NFYA and murine Nfya genes encode a short isoform (lacking a 28-amino acid	(Li et al., 1992) showed that the long and the short mRNA isoforms of murine NFYA (isoforms differing by the length of the N-terminal transcription activation

					optional exon) and a long isoform (containing the optional exon). Hence, I suppose that this is the same 28-amino acid optional peptide that (Dolfini et al., 2012) found to be a part of a transcription activation domain in murine Nfya.	domain) are switched between various cell lines and normal tissue types. They also showed that the corresponding protein isoforms are switched between lymphoid and non-lymphoid cell lines. Additionally, (Dolfini et al., 2012) demonstrated that the short isoform (both in its mRNA and protein forms) is the major one in murine ESCs, while the long isoform (both in its mRNA and protein forms) becomes major after mESC differentiation.
NR3C1	3.11A	NR3C1	Human	Cell lines	Human TF	(Hollenberg et al., 1985) showed that HeLa and various human and mouse lymphoma cell lines express only the longer, LBD+, protein isoform. Additionally, (Bamberger et al., 1995; de Castro et al., 1996; Oakley et al., 1997; Pujols et al., 2002) demonstrated that multiple human tissues and cell lines co-express the LBD+ and LBD- mRNA (and possibly protein) isoforms, while (Oakley et al., 1996; Pujols et al., 2002) showed that in several human tissues the LBD+ mRNA isoform is the major one.
PAX5	3.10G	PAX5	Human	Cell lines	Human TF	(Robichaud et al., 2004) showed that PAX5 switches its major mRNA isoforms between different normal and malignant human CD19+ B cells. Additionally, (Bharti and Mishra, 2015) demonstrated the presence of four murine PAX5 mRNA

						isoforms, differing by encoded C termini and by the presence of the paired domain, in the normal spleen. Two of the four isoforms differed only by encoded C termini.
PAX6	3.10A	PAX6	Human	Cell lines and primary tissues	Human TF	(Azuma et al., 2005) showed that PAX6 swaps its major mRNA isoforms encoding different DBDs during retina development in chicken. (Pinson et al., 2005) demonstrated that the two mRNA isoforms are co-expressed in different parts of the developing murine brain; the ratio of the two isoforms changes but one of them stays major.
PAX8	3.10A	Pax8	Mouse	Cell lines and primary tissues	(Kozmik et al., 1997) found that human PAX8 had the same splicing event changing the DBD as the murine Pax8.	(Kozmik et al., 1997) showed that PAX8 co-expresses two mRNA isoforms with different DBDs at the same ratio (5:1) throughout the murine embryogenesis and in adult ovary and kidney.
PBX1	3.10E	PBX1	Human	Cell lines	Human TF	(Asahara et al., 1999) demonstrated (reproducing, as they note, previous results by other authors) that the two DBD+ protein PBX1 isoforms, with and without a transcription repression domain, are switched between the endocrine and the exocrine compartments of the adult pancreas. Additionally, (Linares et al., 2015) showed that PBX1 switches the two mRNA isoforms when murine ESCs start

						to differentiate into NPCs.
PGR	3.11F	PGR	Human	Cell lines	Human TF	(Mangal et al., 1997) demonstrated that in the human uterus, throughout the menstrual cycle, PGR co-expresses two protein isoforms that were previously shown to differentially activate gene expression. The authors showed that although the ratio of the two isoforms changes during the cycle, one of them was always major. Later on, (Goldman and Shalev, 2006) demonstrated that PGR switches its major protein isoforms between early and late first-trimester trophoblast cells in human.
POU2F2	3.11E	POU2F2	Human?	Cell lines	Human TF?	(Lillycrop and Latchman, 1992) showed that POU2F2 switched its major mRNA isoform between (1) murine B cells and (2) rat brain tissue and murine neuronal cells.
TCF3	3.11C	TCF3	Human	*E. coli* and *in vitro* experiments	Human TF	(Yamazaki et al., 2018) showed that the hESC, HeLa, 293T and U87 cell lines co-express both TCF3 mRNA isoforms, E12 and E47. However, while E12 that contains a DBD inhibitory domain is the major mRNA isoform in hESCs, E47 that lacks the inhibitory domain becomes the major mRNA isoform in HeLa and U87 cells. Additionally, (Pfurr et al., 2017) showed the co-expression of E12 and E47 mRNA isoforms in murine embryonic neural stem

						cells (NSCs).
TCF4	3.11F	TCF4	Human	Cell lines and primary tissues	Human TF	(Sepp et al., 2011) showed that TCF4 co-expresses many different mRNA isoforms across many adult human tissues and that some isoforms are expressed more broadly than others, but they do not comment explicitly on major mRNA isoforms. However, they suggest the existence of switches of major protein isoforms between the following groups of tissues: (a) frontal cortex, hippocampus and cerebellum (where the major isoform should have only one transcription activation domain), (b) testis (where the major isoform should have two transcription activation domains), and (c) lung (where the major isoform again should contain only one transcription activation domain). Additionally, (Grajkowska et al., 2017) demonstrated that two mRNA isoforms of TCF4, a longer one with all transcription activation domains and a shorter one lacking one transcription activation domain, are co-expressed in human and murine plasmacytoid dendritic cells. They also showed that the shorter mRNA isoform becomes major in two other types of murine dendritic cells and in murine B cells where the longer isoform is not expressed.

TEAD4	3.10C	TEAD4	Human	Cell lines, cancer and normal primary tissues	Human TF	(Qi et al., 2016) showed that TEAD4 co-expresses a DBD+ and a DBD- mRNA and protein isoforms in many adult human tissues. The DBD+ mRNA isoform is always the major one, but the levels of expression of protein DBD+ and DBD- isoforms are comparable in some tissues (with the DBD+ still being major): namely, in the spleen, skeletal muscle and liver.
ZGPAT	3.10C	ZGPAT	Human	Cell lines	Human TF	(Yu et al., 2010) showed that the DBD+ and DBD- protein isoforms of ZGPAT are co-expressed in several cell lines that they studied, the DBD+ isoform being the major one. In their previous work (Li et al., 2009), they found that the shorter (presumably, the DBD-) mRNA isoform of ZGPAT was co-expressed with the DBD+ mRNA isoform(s) in the liver and kidneys; in kidneys, the shorter isoform was definitely major.

Table 3. Other human TFs or small groups of TFs with studied alternative isoforms. I apologise to those authors whose work on alternative isoforms of human TFs I do not cite because I did not find it when compiling the table.

TF	References	TF	References
BACH1	(Kanezaki et al., 2001)	PBX3	(Milech et al., 2001)
CEBPG	(Chang et al., 2018)	PHD3	(Cervera et al., 2006)
CREB	(Girardet et al., 1996)	PHF19	(Jain et al., 2020)
CTCF	(Li et al., 2019; Pugacheva et al., 2020)	POU5F1	(Gao et al., 2010)
DPF3	(Cui et al., 2016; Lange et al., 2008; Mignon et al., 2022)	PPARGC1A	(Zhang et al., 2009)
ELF5	(Piggin et al., 2016; Zhou et al., 1998)	PRDM2	(Cheedipudi et al., 2015)
ESR1	(Chantalat et al., 2016; Herynk and Fuqua, 2004; Zhang et al., 2016)	PRRX1	(Norris et al., 2000)
FOXJ3	(Cabezas-Wallscheid et al., 2014)	PSIP1	(Blokken et al., 2017; Brown-Bryan et al., 2008; Pradeepa et al., 2012; Turlure et al., 2006)
FOXO4	(Lee et al., 2008)	RARA	(Parrado et al., 2001)
HNF1B	(Bach and Yaniv, 1993)	REST	(Raj et al., 2011)
HNF4A	(Eeckhoute et al., 2003)	RFX genes	(Sugiaman-Trapman et al., 2018)
HOXB6	(Shen et al., 1991)	RUNX1	(Komeno et al., 2014)
HSF genes	(Pirkkala et al., 2001)	SALL2	(Hermosilla et al., 2017, 2018)

HSF4	(Tanabe et al., 1999)	SLC2A11	(Scheepers et al., 2005)
IKZF3	(Caballero et al., 2007)	SMAD4	(Pierreux et al., 2000)
IRF3	(Karpova et al., 2001; Li et al., 2011)	SMRT	(Goodson et al., 2005)
KDM2A	(Lađinović et al., 2020)	SREBF1	(Amemiya-Kudo et al., 2002; Horton, 2002; Hua et al., 1995)
KLF genes	(Camacho-Vanegas et al., 2013)	SREBF2	(Wang et al., 2006a)
KLF6	(Narla et al., 2005a, 2005b)	STAT3	(Maritano et al., 2004; Ng et al., 2012; Park et al., 2000; Shao et al., 2001)
LEF1	(Hovanes et al., 2000)	TBX3	(Fan et al., 2004)
MEF2 genes	(Yu et al., 1992)	TFAP2A	(Berlato et al., 2011)
MEF2C	(McDermott et al., 1993)	TFEB	(Kuiper et al., 2004)
MEF2D	(Breitbart et al., 1993)	TFEC	(Kuiper et al., 2004)
MEIS1	(Crist et al., 2011; Noro et al., 2006; Xiong et al., 2009)	TP63	(Candi et al., 2007; Petitjean et al., 2005, 2008; Yang et al., 1998)
MEIS2	(Yang et al., 2000)	TP73	(Zaika et al., 2002)
MZF1	(Brix et al., 2020)	WT1	(Bickmore et al., 1992; Ullmark et al., 2017)
NFAT genes	(Serfling et al., 2012; Vihma et al., 2016)	ZFX	(Schneider-Gädicke et al., 1989)
NFAT2	(Lucena et al., 2016)	ZNF197	(Li et al., 2003)

NFIB	(Chen et al., 2014)	ZNF202	(Langmann et al., 2003; Porsch-Özcürümez et al., 2001; Wagner et al., 2000)
NR1H4	(Huber et al., 2002)	ZNF215	(Alders et al., 2000)
NR2F2	(Rosa and Brivanlou, 2011)	ZNF415	(Cheng et al., 2006)
PAX3	(Tsukamoto et al., 1994; Wang et al., 2006)	ZNF655	(Houlard et al., 2005)
PAX6	(Zhang et al., 2010)	ZNF746	(Al Chiblak et al., 2019)
PAX genes	(Thompson et al., 2021)	ZNF76	(Zheng and Yang, 2006)

3.4. Discussion

In this chapter, I compared the production of alternative isoforms by TF and non-TF coding genes and found that the proportion of TF genes that produce multiple isoforms is greater than the corresponding proportion of non-TF genes, and also that TF genes produce, on average, more alternative isoforms than non-TF genes. These results support previously published findings (Talavera et al., 2009; Taneri et al., 2004) and also extend them, as I found that my results hold for any number of exons and any gene length, apart from the very small gene lengths from the first two length deciles. Importantly, by selecting isoforms from the full human genome annotation, I automatically included isoforms that are produced not only by alternative splicing, but also via the usage of alternative transcription start or polyA sites. In total, my results suggested the importance of alternative isoforms for human TFs, and hence – the importance of studying human TF isoforms genome-wide.

Consequently, as a basis for such a study, I compiled a collection of 24 human TFs whose alternative isoforms have been characterised experimentally. Moreover, I found another 68 human TFs (and small groups of closely related TFs, as some publications study small TF subfamilies) whose alternative isoforms were studied only to some extent. Overall, I hope that this collection of 92 TFs and small groups of TFs becomes a convenient resource for the future functional studies of human TF isoforms.

Chapter 4. Alternative isoforms without a DNA-binding domain

4.1. Introduction

According to an up-to-date and comprehensive overview (Lambert et al., 2018), the human genome encodes ~1,650 TFs from more than 60 families defined by single DBD types or combinations of different DBD types (http://humantfs.ccbr.utoronto.ca/allTFs.php). The biggest TF families are C2H2 zinc fingers ($N \approx 750$), Homeodomains ($N \approx 200$) and bHLH ($N \approx 110$). Hereafter, if a TF gene encodes more than one DBD type, I will denote its family by a combination of its DBD types in the alphabetical order. For instance, "Homeodomain; POU" or "CUT; Homeodomain."

By definition, each TF gene encodes at least one DBD, and a protein expressed by a TF gene and containing a DBD can function as a TF. However, some TFs, including human ones, also express isoforms lacking a DBD (DBD- isoforms) that cannot serve as TFs but may nevertheless have other functions. For example, TEAD4 can express two isoforms, a full-length isoform that includes a DBD and a cofactor-binding domain (the DBD+ isoform) and a short DBD- isoform that still contains the cofactor-binding domain. These two isoforms are co-expressed in the human tissues, with the DBD+ isoform being the major one. The DBD+ isoform provides a final point of the Hippo-YAP signalling pathway that controls cell proliferation. Specifically, via binding the cofactor YAP and the cognate motifs in promoters, the DBD+ isoform of TEAD4 activates the expression of its target genes. However, the DBD- isoform competes with the DBD+ isoform via binding YAP outside the chromatin and transferring it to the cytoplasm. In this way, the DBD- isoform disrupts the Hippo-YAP pathway and attenuates the transcription activation by the DBD+ isoform (Qi et al., 2016).

The DBD+ isoforms of the human TFs that encode several DBDs may contain more than one DBD of the same type or different types. A combination of DBDs defines the DNA binding affinity and specificity of an isoform in a more complex way than a single DBD. A classical example of TFs with multiple DBDs are C2H2 zinc finger factors. Their DBD+ isoforms contain arrays of C2H2 zinc finger domains that bind DNA with unique specificities (Najafabadi et al., 2015; Schmitges et al., 2016). Previous studies established that C2H2 zinc finger factors need at least two zinc finger domains to bind DNA (Wolfe et al., 2000) and that

the neighbouring C2H2 zinc fingers highly affect each other's DNA binding specificity (Garton et al., 2015).

A canonical example of TFs producing DBD+ isoforms with DBDs of more than one type are "POU; Homeodomain" (POU-domain) factors. Their bipartite DBD consists of an N-terminal POU-specific domain and a C-terminal POU homeodomain. The two domains are connected by a linker of a variable length (Herr et al., 1988). POU-domain TFs bind variable 8-bp motifs, and each of the two domains binds a half of the motif (Kristie and Sharp, 1990; Sturm and Herr, 1988; Takeda et al., 1994; Verrijzer et al., 1990). The variability of the recognised motifs is likely explained by the effect that differences in the DNA sequence bound by one domain exert on the DNA binding specificity of the other domain (Aurora and Herr, 1992). Consequently, the alternative isoforms of a TF with different DBD combinations are likely to have different DNA binding affinities and specificities.

Apart from the differential presence of a DBD, alternative splicing may cause changes within the amino acid sequence of a DBD. For example, two DBD+ isoforms of PAX8 that differ by one serine residue in their DBD sequences have drastically different DNA binding specificities. Namely, the isoform that contains the additional serine residue is not able to bind the paired domain recognition sequence but instead binds a 5aCON DNA sequence that consists of four consensus sites for the binding of the C-terminal part of the paired domain (Kozmik et al., 1997). Another well-studied example of functional changes in the DBD sequence is presented by two DBD+ isoforms of FOXP1. Specifically, one of them maintains the pluripotency of embryonic stem cells (ESCs), while the other promotes ESC differentiation. The difference between the two isoforms lies in the Forkhead DBD, as alternative splicing swaps two exons inside the DBD and hence produces two variants of the domain with different DNA-binding specificities (Gabut et al., 2011).

Therefore, in the current chapter I analyse the DBD- isoforms of human TFs, study the expression and tissue specificity of these isoforms and predict their functions based on effector domains.

4.2. Methods

Scripts for the described analysis can be found at https://github.com/ComputationalRegulatoryGenomicsICL/human-tf-isoforms-thesis.

Refer to sections 3.2.1, 3.2.2 and 3.2.3 for the detailed procedures for the selection of the human TF genes and isoforms, for the scanning of the TF isoforms for known domains and for the subsequent manual curation of the scanning results. Procedures described in sections 4.2.1 and 4.2.2 below are based on the sets of TF genes, isoforms and domains formed using the procedures described in sections 3.2.1-3.2.3.

4.2.1. Annotation of non-DNA-binding domains

To predict functions of DBD- isoforms, I selected unique combinations of non-DBDs present in DBD- isoforms and manually predicted functions that these combinations may confer. I based my predictions on examples of functional DBD- isoforms that I collected from the literature (see the Results section 4.3.3). However, prior to that, I manually annotated functions of matched non-DBD InterPro entries (N = 262) using the descriptions provided in these entries and via extensive literature search (Supplementary Table 2). Next, I deduplicated all non-DBDs in TF isoforms. Specifically, I found non-DBD InterPro entry matches that overlapped by ≥70% of their length with at least one other non-DBD InterPro entry match and then manually selected matches to retain based on the following rules:

1. If overlapping matches represented the same domain, I kept the longer match (Fig. 4.1A-B). In this way, I deduplicated pairs of overlapping specific matches and pairs of matches where one match was general and the other was specific (see section 3.2.1 for the definitions of general and specific matches). For instance, in the isoform ENST00000318003 of CC2D1A, the match of IPR000008 overlapped by 100% of its length the match of IPR037772, and both matches represented the same C2 domain (Fig. 4.1A). Similarly, in the isoform ENST00000322349 of EEA1, the match of IPR017455 overlapped by 100% of its length the match of IPR000306, and both matches represented the same FYVE-type zinc finger (Fig. 4.1B). While in the first case, IPR037772 was a descendant of IPR000008 in the InterPro Domain hierarchy, in the second case, IPR017455 and IPR000306 were not related through the InterPro hierarchy. Nevertheless, in both cases I retained the longer match. In principle, I could

exclude general matches (as I did for DBDs) and then filter the remaining overlaps of specific non-DBD matches by length.

2. If overlapping matches represented different domains (or at least it was unclear if they represented the same domain), then:
    a. If they overlapped reciprocally by ≥70% of their lengths, I retained the match of a more biologically relevant or better described InterPro entry. For instance, in the isoform ENST00000318524 of NFX1, matches of IPR019787 ("Zinc finger, PHD-finger") and IPR001841 ("Zinc finger, RING-type") overlapped reciprocally by ≥70% of their lengths but represented two different types of a zinc finger. I retained the match of IPR001841 because, according to UniProt (The UniProt Consortium, 2021), this locus matches to an atypical RING-type zinc finger (Fig. 4.1C).
    b. If they did not overlap reciprocally by ≥70% of their lengths, I kept both matches, as they may represent a combination of different smaller domains. For example, in the isoform ENST00000235790 of KDM5B, the matches of IPR019787 and IPR003347 overlapped, but the latter covered considerably less than 70% of the length of the former. I retained both matches, as the signature of IPR003347 matched in between the two signatures of IPR019787 (Fig. 4.1D).

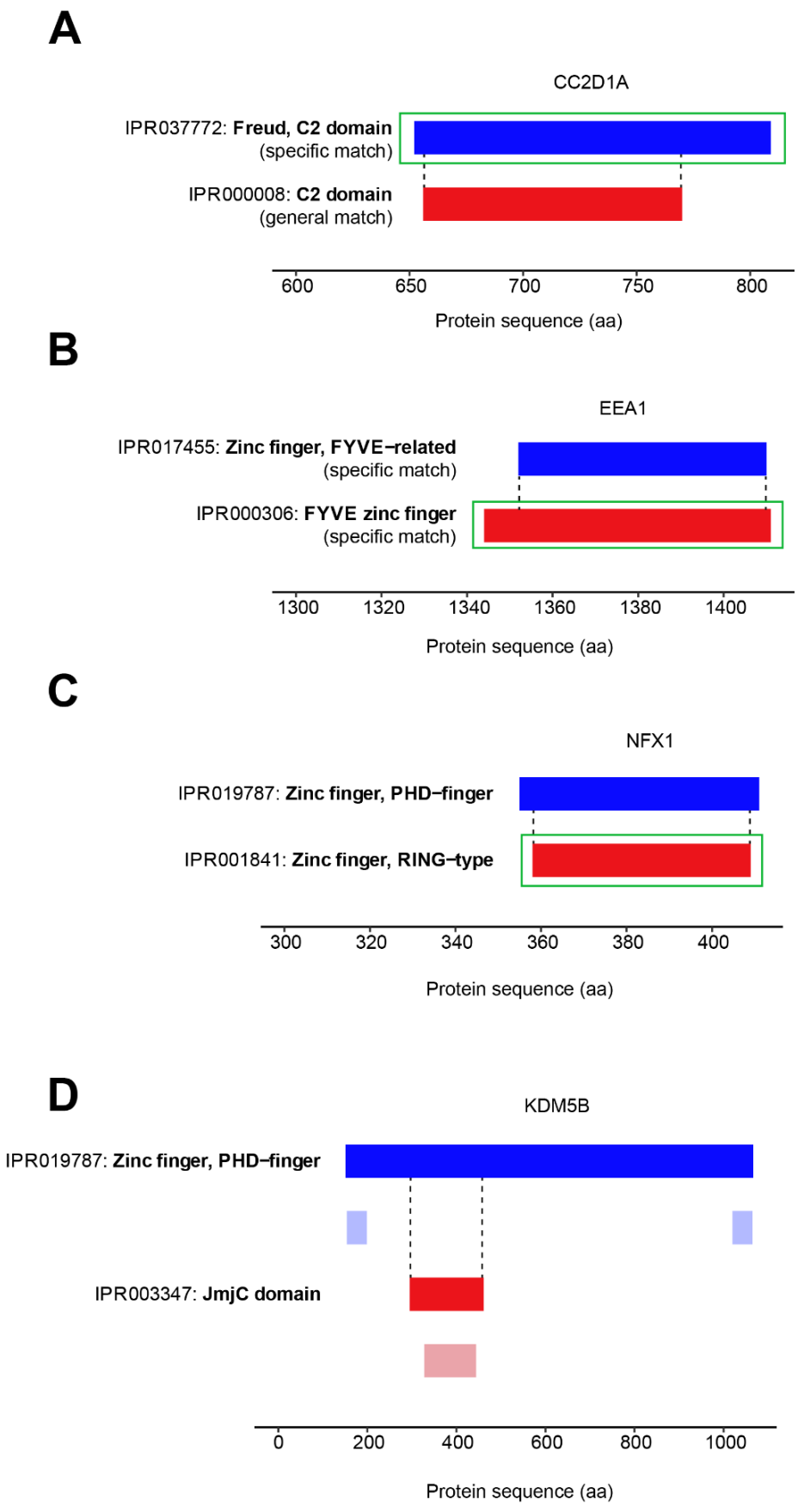

A
CC2D1A
IPR037772: Freud, C2 domain
(specific match)
IPR000008: C2 domain
(general match)
600
650
700
750
800
Protein sequence (aa)
B
EEA1
IPR017455: Zinc finger, FYVE−related
(specific match)
IPR000306: FYVE zinc finger
(specific match)
1300
1320
1340
1360
1380
1400
Protein sequence (aa)
C
NFX1
IPR019787: Zinc finger, PHD−finger
IPR001841: Zinc finger, RING−type
300
320
340
360
380
400
Protein sequence (aa)
D
KDM5B
IPR019787: Zinc finger, PHD−finger
IPR003347: JmjC domain
0
200
400
600
800
1000
Protein sequence (aa)

Fig. 4.1. Examples of the deduplication of overlapping non-DBD matches. (A) An overlap of a general and a specific match. **(B)** An overlap of two specific matches. In both cases, the overlapping matches represent the same domain, and I retained the longer match (framed in a green box). **(C)** A reciprocal ≥70%-overlap between matches of two different types of zinc fingers. The retained zinc finger is framed in a green box. **(D)** An overlap, which is not 70%-reciprocal, between matches of two different InterPro entries signifies the position of a domain of one type in between two domains of another type. These domains are signatures within InterPro entries, and corresponding signatures and entries are marked by a shade of the same colour. In this case, I retained both InterPro entry matches.

4.2.2. Analysis of TF isoform expression

For the analysis of TF isoform expression in adult human tissues, I chose the Analysis Freeze dataset from the Genotype-Tissue Expression (GTEx) database version 8 (Lonsdale et al., 2013) and selected 46 human tissues that were represented by ≥100 samples. From these, I excluded "Cells – Cultured fibroblasts" and "Cells – EBV-transformed lymphocytes," because they are cell cultures and not primary tissues; I also removed "Whole Blood" as a very heterogeneous "tissue". Next, for each of the selected tissues, I chose 50 samples at random. In this way, I obtained 2,300 samples for further analysis.

Next, I selected coding TF isoforms which I used in the domain analysis and which were available in GTEx version 8 ($N = 4{,}089$). For each TF isoform in each sample, I calculated the value of $\log_2(\text{TPM} + 1)$ to reduce the cross-sample expression variation of highly expressed isoforms. Here, "TPM" denotes the expression level of an isoform in Transcripts per kilobase Per Million (TPMs), and the addition of 1 accounts for the cases when TPM = 0. Further, I calculated the expression level of each TF isoform in each tissue as the median expression level of the isoform across all samples from that tissue. Finally, to compare the expression levels of the DBD+ and DBD- isoforms, Dr. Koustav Pal calculated the buffered expression level of each isoform across all selected human tissues. We defined the **buffered expression level** of an isoform as the mean of the top 20% of its expression levels across all tissues.

To quantify the **tissue specificity** of TF isoform expression, I used the **tau index** (Yanai et al., 2005) which was previously shown to be one of the best measures of tissue specificity (Kryuchkova-Mostacci and Robinson-Rechavi, 2017):

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1}, \quad \hat{x}_i = \frac{x_i}{\max_{1\le j\le n} x_j},$$

where x_i is the expression level of an isoform in tissue i and n is the number of tissues. Isoforms with $\tau = 0$ are expressed in all tissues at the same level, while isoforms with $\tau = 1$ are expressed in only one tissue.

Next, from the GTEx version 8 Analysis Freeze dataset, I extracted expression quantification data for the TF genes (N = 1,488) whose isoforms I processed above. I log-scaled the expression of the selected genes across samples and calculated the tissue specificity of TF genes according to the same procedures that I used for TF isoforms.

To calculate the buffered expression levels of TF genes and the tissue specificity of TF genes and isoforms, we used the log-scaled expression values; to calculate the numbers of expressed TF genes and isoforms, I used a cutoff of 1 log-scaled TPM to count a gene (isoform) as expressed.

Finally, for each TF expressed in adult human tissues, I found its major isoforms. I defined the **major isoform** of a TF in a tissue as its highest-expressed isoform in that tissue. As the major isoform of a TF is defined per tissue, one TF may express more than one major isoform.

This preprocessing allowed me and Dr. Koustav Pal to analyse the expression levels and tissue specificity of DBD+ and DBD- isoforms across adult human tissues and also to study the switches of major TF isoforms (see the Results sections 4.3.2, 5.3.1, 5.3.2).

4.3. Results

4.3.1. Loss of DNA-binding domains is widespread among human transcription factors

A DNA-binding domain (DBD) is a hallmark of a TF, and isoforms produced by TF genes but lacking DBDs cannot function as TFs. Consequently, I began the genome-wide characterisation of human TF isoforms by considering a loss of a DBD. Using the 4,134 coding isoforms of the selected 1,512 human TFs, I found that 27.5% (415) of all human TFs lose a DBD in at least one of their mRNA isoforms and that 17% (705) of all human TF isoforms do not have a DBD. I denoted such isoforms "DBD-" (Fig. 4.2A).

Next, I calculated in each TF family the proportion of TFs that produced DBD- isoforms. I found that this proportion varied from 0 to 1 across families and that in the majority of families it was between the two extremities (Fig. 4.2B). It took the extreme values mostly in families consisting of only one or two TFs, such as "ARID/BRIGHT; C2H2 ZF; RFX," "AT hook; C2H2 ZF," or "AT hook; MBD"; however, there were families with considerably more TFs (for instance, "Homeodomain; POU," "ARID/BRIGHT," or "Grainyhead") that produced only DBD-containing (DBD+) isoforms (Fig. 4.2B).

Interestingly, some TF families of comparable sizes (for instance, of sizes that differ by no more than two TFs) had considerably different proportions of TFs producing DBD- isoforms. For example, in the "C2H2 ZF; Homeodomain" family (13 TFs) this proportion was 38%, while in the "THAP finger" family (12 TFs) it equalled 25% and in the "CENPB; Pipsqueak" family (11 TFs) it was zero (Fig. 4.2C). Families "CxxC ZF" (11 TFs), "Rel" (10 TFs) and "IRF" (9 TFs) with the corresponding proportions of 54%, 30% and 22% provided another example (Fig. 4.2D).

Furthermore, apart from losing/gaining a DBD in alternative isoforms, a TF may change its DBD sequence. Hence, I considered TFs encoding at most one DBD (single-DBD TFs), which comprised 47% of all TFs (Fig. 4.3A), and stratified them by the number of unique DBD sequences they expressed (Fig. 4.4A). I found that only 15% of single-DBD TFs encoded different DBD sequences and that the majority of such TFs encoded just two different sequences (Fig. 4.4A). Additionally, Fig. 4.4B shows the proportion of single-DBD TFs expressing different numbers of DBD sequences in each TF family. Importantly, the stratification of TFs in each family by the maximum number of DBDs (Fig. 4.3B) showed that all single-DBD TFs encoded only one *type* of a DBD, as follows from their family names, "AP-

2" to "C2H2 ZF," none of which is combined. Consequently, single-DBD TFs do not change their DBD *type* between alternative isoforms, although 15% of them change the DBD sequence.

Finally, I considered TFs with more than one DBD and studied the way they change DBD combinations across DBD+ isoforms. I defined a **DBD combination** as a set of DBDs in a particular isoform. Consequently, two DBD combinations may differ by the number and/or types of DBDs. I grouped the selected TFs into those changing their DBD combinations and those retaining the same combination in all DBD+ isoforms. I found that 30% of the selected TFs have different DBD combinations in different isoforms (Fig. 4.5A). Interestingly, while TFs from families such as "Myb/SANT," "GATA," "AT hook," or "CENPB; Pipsqueak" always preserved a DBD combination across their DBD+ isoforms, high proportions of TFs from other families (for instance, "C2H2 ZF" or "Homeodomain; POU") changed their DBD combinations (Fig. 4.5B).

Overall, I found that 27.5% of human TFs lose DBDs in alternative isoforms. Because DBD- isoforms cannot function as TFs, they may be non-functional or have non-TF functions. Consequently, I attempt to predict possible functions of DBD- isoforms in section 4.3.3, while in the next section I study the expression of DBD- isoforms across adult human tissues to estimate the possible functional importance of these isoforms.

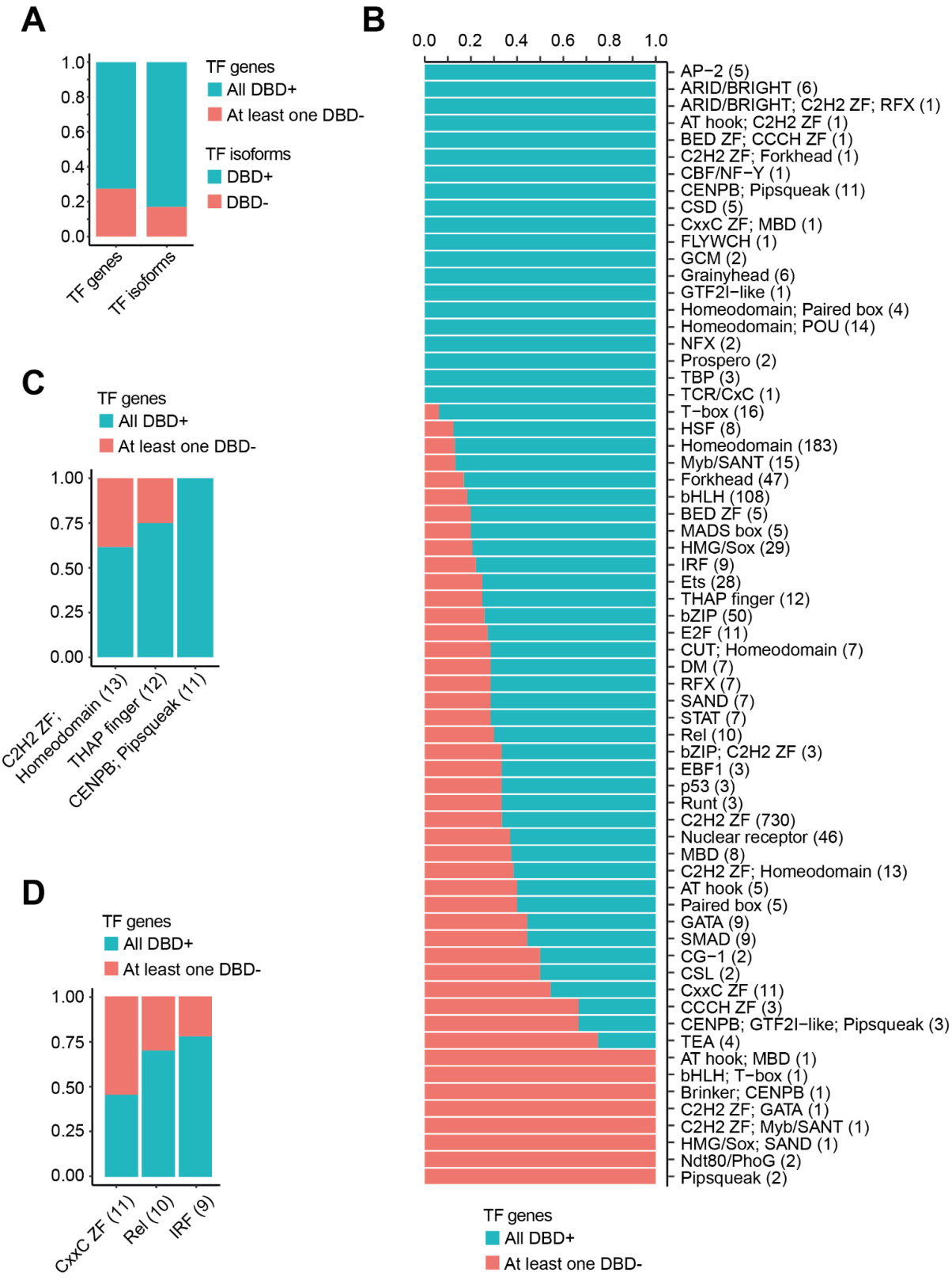


Fig. 4.2. TF isoforms without a DBD. (A) The proportion of TFs that lose a DBD in at least one isoform (left) and the proportion of DBD- isoforms among all isoforms of human TFs (right). **(B)** The proportion of TFs that produce only DBD+ isoforms (blue) or at least one

DBD- isoform (red), per TF family. **(C, D)** Examples of TF families which have comparable sizes but drastically different proportions of TFs producing DBD- isoforms. See sections 3.2.1 and 3.2.2 for the domain search and curation procedures and for the definition of TF families.

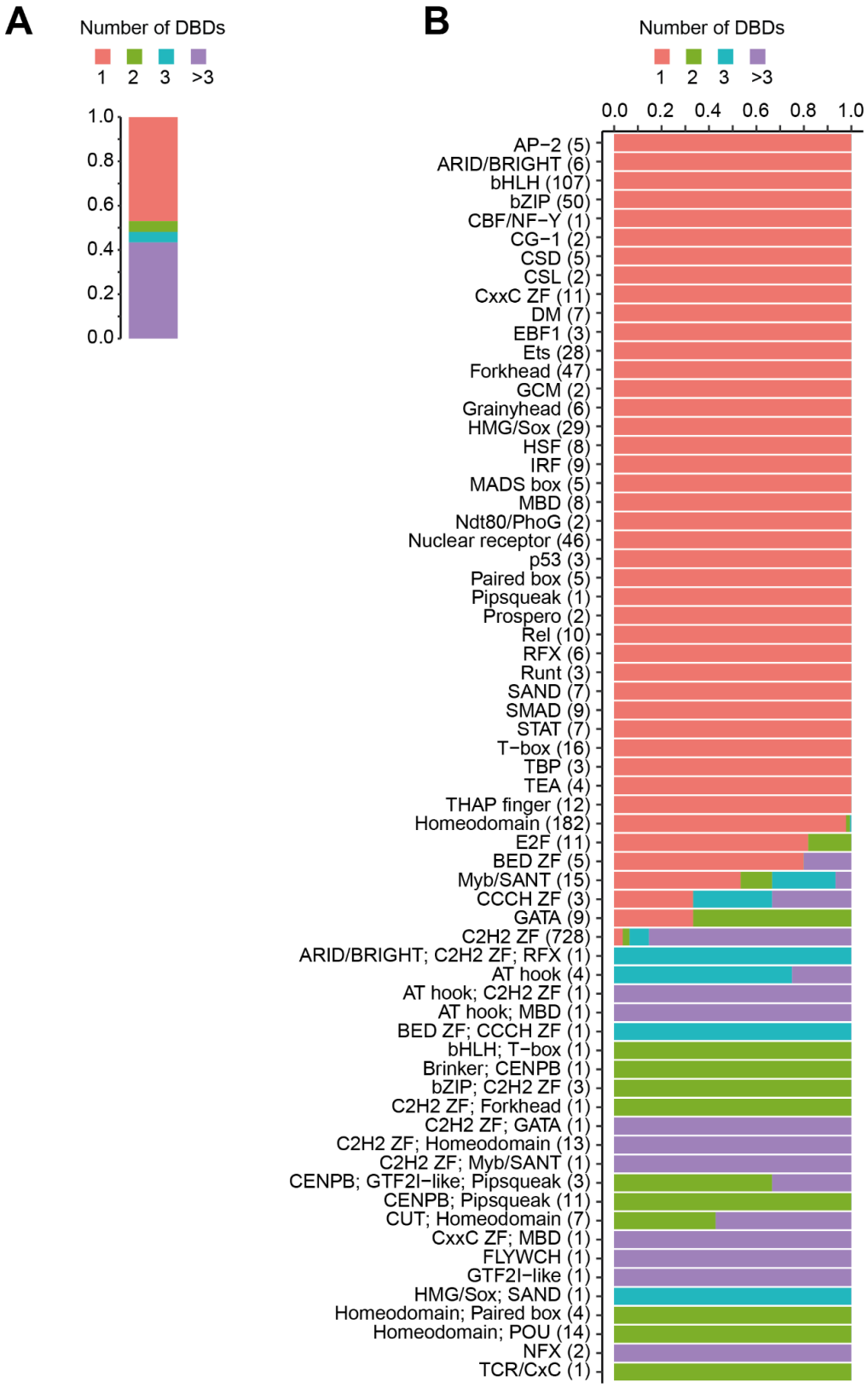


Fig. 4.3. Proportions of TFs with different numbers of DBDs. (A) Overall. **(B)** Per family. See sections 3.2.1 and 3.2.2 for the domain search and curation procedures and for the definition of TF families.

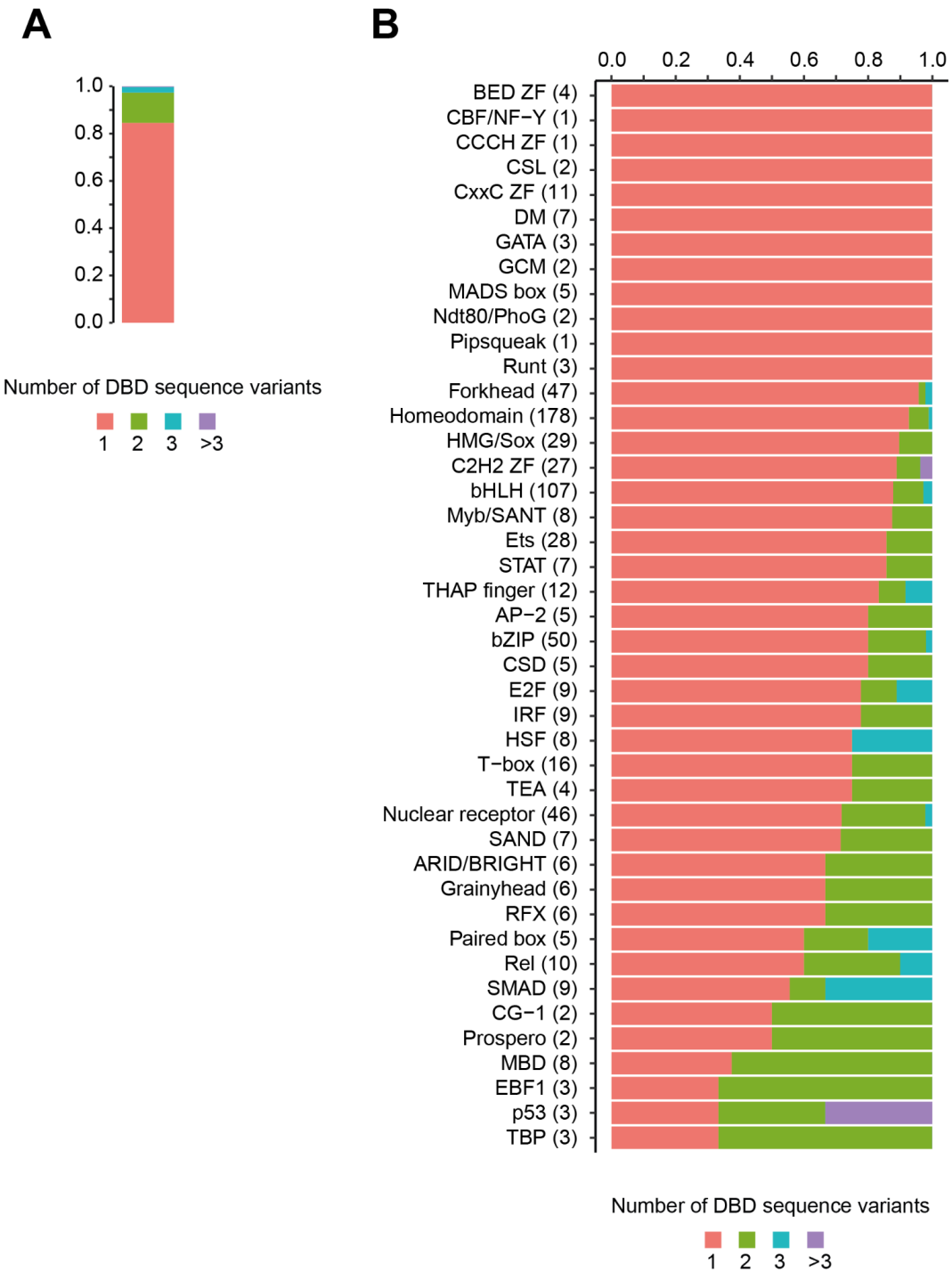


Fig. 4.4. Proportions of single-DBD TFs with different numbers of DBD sequences. (A) Overall. **(B)** Per family. See sections 3.2.1 and 3.2.2 for the domain search and curation procedures and for the definition of TF families.

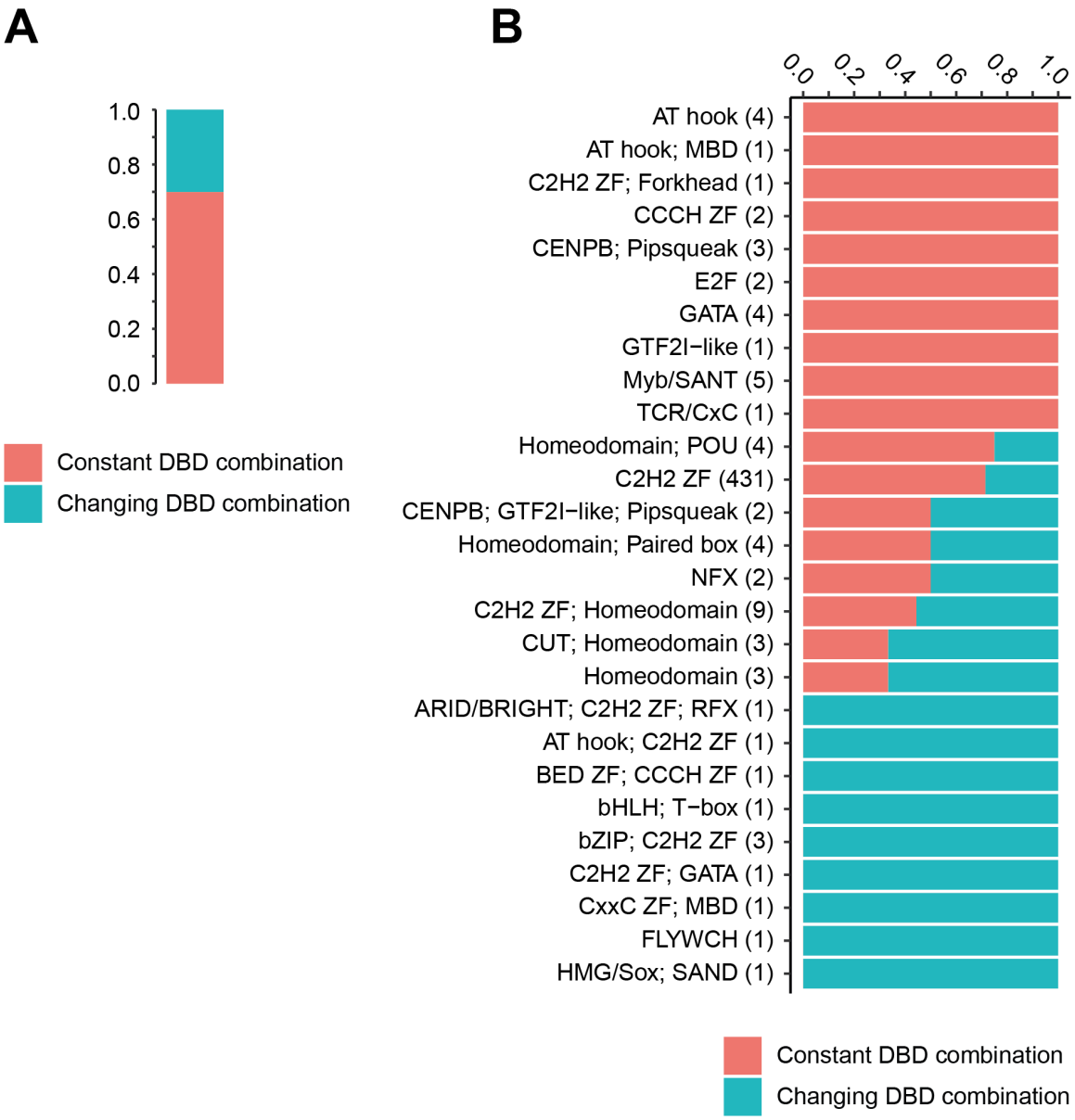


Fig. 4.5. Proportions of multi-DBD TFs with constant or changing DBD combinations. **(A)** Overall. **(B)** Per family. See sections 3.2.1 and 3.2.2 for the domain search and curation procedures and for the definition of TF families.

4.3.2. Expression and tissue specificity of DBD- isoforms suggest their functional importance

To assess the functional importance of DBD- isoforms, I studied their expression across a panel of 2,300 samples from 46 adult human tissues processed by the Genotype-Tissue Expression (GTEx) Consortium (Lonsdale et al., 2013). For this analysis, I used 4,089 coding TF isoforms (of which 3,391 were DBD+ and 698 were DBD-) quantified by the GTEx in the selected samples. I postulated that an isoform is expressed in a tissue if its expression value (in TPMs) is greater than or equal to 1 in that tissue. Among the selected isoforms, 66% (2,707) were expressed in at least one tissue. The selected isoforms were annotated in 1,488 TFs, of which 92% (1,378) were expressed in at least one tissue. Also, I defined the major isoform of a TF in a tissue as the highest-expressed isoform of that TF in that tissue. There were 1,777 major isoforms expressed by 1,378 TFs across all tissues. The fact that the number of major isoforms exceeds the number of expressed TFs means that some TFs exchange their major isoforms between tissues.

Using these definitions, I compared the proportions of expressed and major DBD+ and DBD- isoforms (Fig. 4.6A-B). I found that 70% (2,359) of DBD+ isoforms were expressed in at least one adult human tissue, and that from these, 70% (1,656) became major in at least one tissue (Fig. 4.6A). In contrast, only 50% (348) of DBD- isoforms were expressed in at least one tissue, and among these, 33% (115) became major (Fig. 4.6B).

To exclude the possibility that the majority of expressed DBD- isoforms counted as expressed simply because they randomly showed a low, but sufficient, transcription in a single tissue, my collaborator, Dr. Koustav Pal, implemented an idea by Professor Boris Lenhard and calculated the “buffered” expression of each isoform as the mean of the top 20% of its expression levels across all tissues; then Dr. Pal compared the distributions of the buffered expression levels of DBD+ and DBD- isoforms (Fig. 4.6C). He found that 56% of DBD+ and 33% of DBD- isoforms were still counted as expressed with this stricter approach (namely, had the buffered expression level greater than or equal to 1).

Next, we studied the tissue specificity of DBD+ and DBD- isoforms across adult human tissues. To this end, I calculated the tau index of tissue specificity separately for TF genes and TF isoforms, and then Dr. Pal and I studied these datasets in the following way (Fig. 4.7). We grouped TFs into quantiles by their tau index (where 0 denoted equal expression in all tissues and 1 denoted expression in only one tissue), and in each quantile we calculated the median tissue specificity of isoforms produced by the corresponding TFs. We found that the tissue specificity of DBD+ isoforms correlated well with the tissue specificity of TF genes (red line

in Fig. 4.7), while the specificity of DBD- isoforms diverted from this trend for unspecific TFs (blue line in Fig. 4.7). From this, we concluded that TFs expressed across a broad set of tissues have broadly expressed DBD+ isoforms but considerably more tissue-specific DBD- isoforms.

Furthermore, we studied the tissue specificity and buffered expression of DBD+ and DBD- isoforms within particular TF families (Fig. 4.8). For this analysis, Dr. Pal selected families with more than 10 DBD- isoforms and compared distributions of the tissue specificity index and buffered expression levels of the two types of isoforms in each family. We observed a lower median buffered expression of DBD- isoforms, in comparison to DBD+ isoforms, in all TF families except for "CxxC ZF", where DBD+ and DBD- isoforms had almost the same median expression (Fig. 4.8A). We also observed a higher median tissue specificity of DBD- isoforms in all TF families except for the Forkhead family where the median specificities of DBD+ and DBD- isoforms were comparable (Fig. 4.8B).

Overall, Dr. Pal and I found that a half of DBD- isoforms were expressed in adult human tissues, and from these a third became major in at least one tissue. Along with a higher tissue specificity, this fact suggests a functional role for at least a subset of DBD- isoforms.

Consequently, in the next section, I functionally characterise DBD- isoforms based on their non-DBDs.

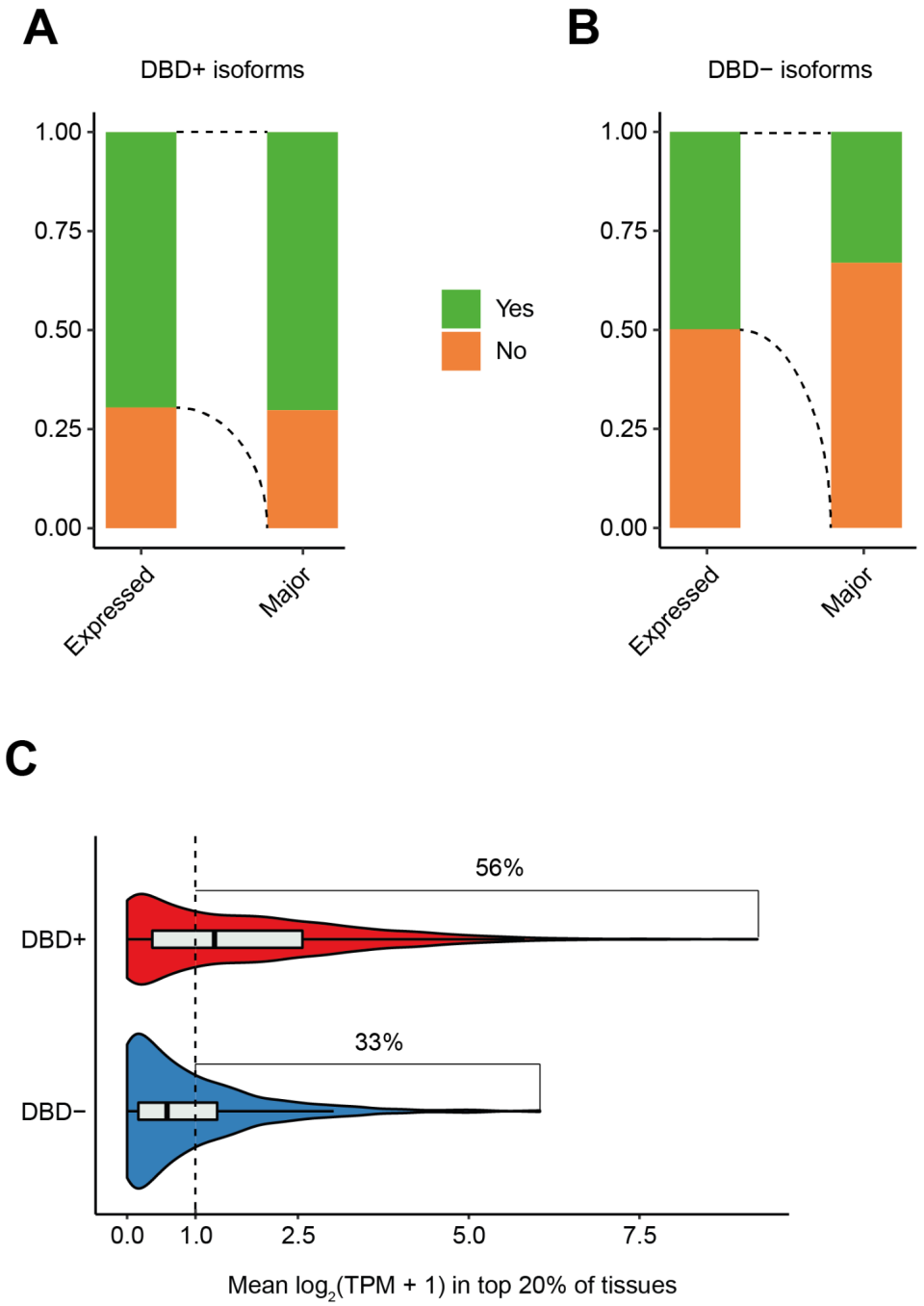


Fig. 4.6. Expression of DBD+ and DBD- TF isoforms in adult human tissues. Proportions of expressed (left) and major (right) isoforms among all DBD+ isoforms **(A)** and DBD- isoforms **(B)**. **(C)** Distributions of the "buffered" expression levels of DBD+ (red) and DBD- (blue) isoforms. The threshold of 1 TPM signifies the postulated minimal expression level. Panel C was generated by Dr. Pal and adapted by me for the present text. See section 4.2.2 for methodological details.

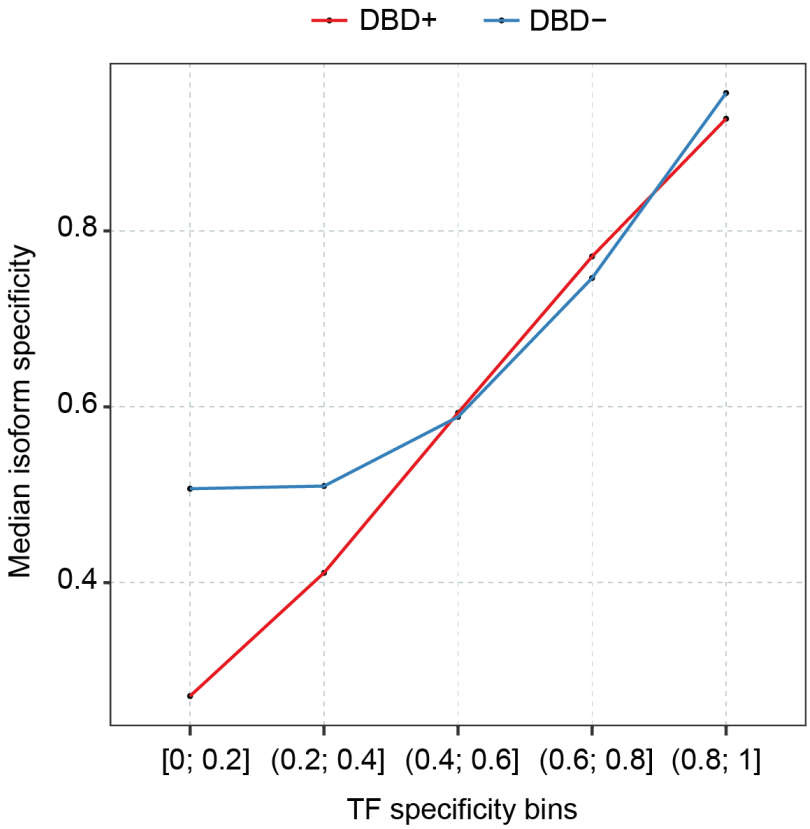


Fig. 4.7. Tissue specificity of DBD+ and DBD- TF isoforms. Quantiles of the tissue specificity of TF genes are shown on the x axis, while the median isoform tissue specificity in each quantile is shown on the y axis. The tissue specificities of DBD+ (red) and DBD- (blue) isoforms are shown separately. The plot was generated by Dr. Pal and adapted by me for the present thesis. See section 4.2.2 for methodological details.

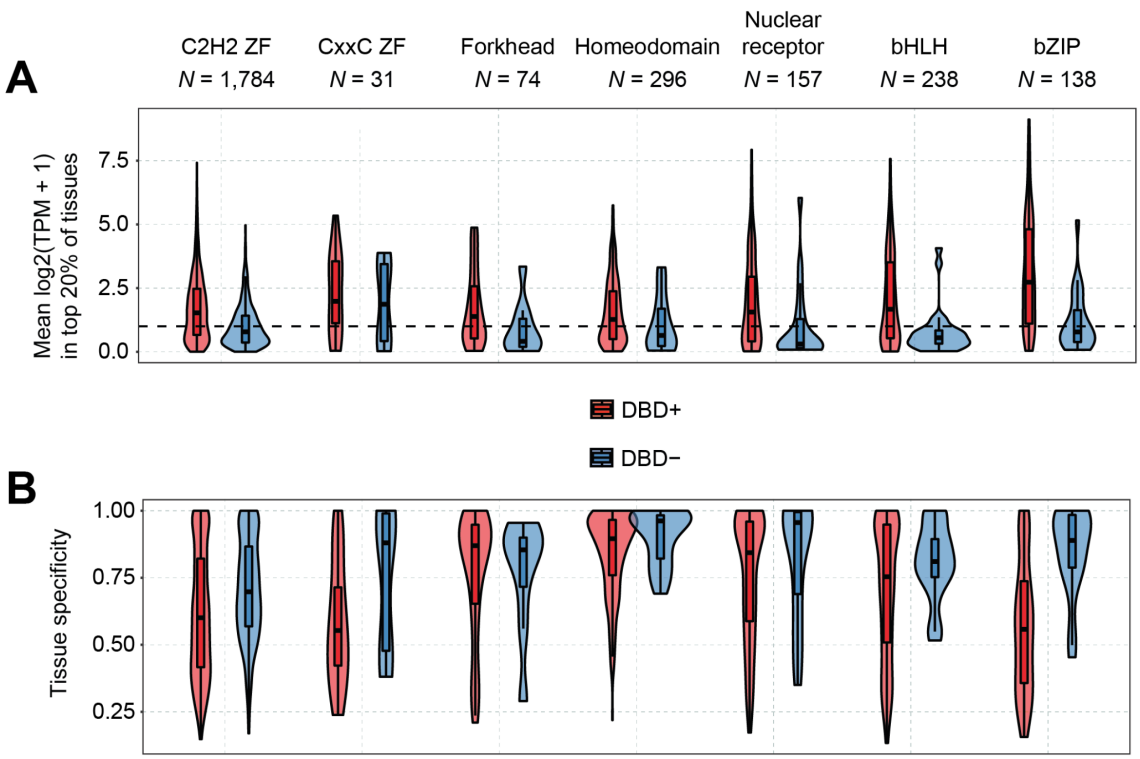


Fig. 4.8. Expression and tissue specificity of DBD+ and DBD- TF isoforms per TF family. Distributions of the "buffered" expression levels **(A)** and tissue specificity values **(B)** of DBD+ (red) and DBD- (blue) isoforms produced by TFs from families with more than 10 DBD- isoforms. The two plots were generated by Dr. Pal and adapted by me for the present text. See section 4.2.2 for methodological details.

4.3.3. Non-DBDs enable functional predictions for a part of DBD- isoforms

To predict possible functions of DBD- isoforms, I studied the non-DNA-binding domains that they contain. To this end, using the descriptions of non-DBDs from the InterPro database and literature, I first manually annotated general functions of 262 unique InterPro entries matched across all human TF isoforms (Supplementary Table 2) and then considered non-DBDs specifically in DBD- isoforms. Because the DBD- isoforms of the C2H2 zinc finger TFs ($N = 430$) were almost two times as numerous as the DBD- isoforms of all other TFs ($N = 268$), I considered these two sets of isoforms separately. I found that 47% of the DBD- isoforms produced by the C2H2 ZF TFs had a transcription repression (KRAB) domain, 11% had an oligomerization domain (SCAN) and ~5% had a dimerization domain (BTB/POZ). Frequencies of all other types of domains were negligible (Fig. 4.9A).

Among the DBD- isoforms of other TFs, ~13% contained a dimerization domain (mainly the one from the interferon regulatory factor 3, or an HSR or a FOXP coiled-coil

domain). Additionally, 7.5% contained a ligand-binding domain (the one from nuclear receptors or a PAS domain) and 5.6% contained a transcription activation domain. Other types of domains with the frequency >5% included general ones, "Protein binding" and "Transcription regulation," for which I could not define a function precisely, and the "Unspecified" domains for which I could not define a function even approximately. Other domain types had negligible frequencies (Fig. 4.9B).

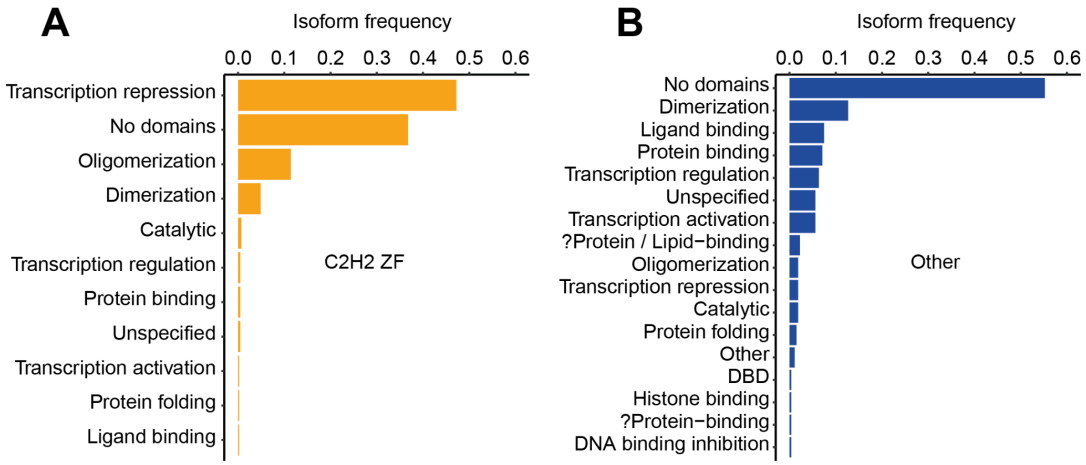


Fig. 4.9. Proportions of DBD- isoforms by the types of non-DBDs that they contain. DBD- isoforms are stratified into those produced by the C2H2 zinc finger TFs **(A)** and other TFs **(B)**. Because a DBD- isoform may contain non-DBDs of more than one type, the proportions may not sum up to 1. See section 4.2.1 for methodological details.

Based on my functional annotation of non-DBDs in DBD- isoforms and also on the examples of DBD- isoforms characterised in literature (Fig. 3.10B-C), I predicted possible functions of DBD- isoforms expressed by the C2H2 zinc finger TFs and other TFs (Fig. 4.10; Table 4). DBD- isoforms of the C2H2 zinc finger TFs were mainly predicted to be activators or "reverse regulators" (transcription regulators whose effect is opposite to the one of DBD+ isoforms; Fig. 4.10A). DBD- isoforms expressed by other TFs were mainly predicted to reverse the regulation of transcription by DBD+ isoforms, to sequester ligands from the chromatin, to bind proteins outside the chromatin or to repress transcription (Fig. 4.10B).

Importantly, 37% of the DBD- isoforms expressed by the C2H2 zinc finger TFs and 62% of the DBD- isoforms expressed by other TFs do not have a predicted function (Fig. 4.10A-B). This is mainly due to the absence of any domains in these isoforms (Fig. 4.10C-D).

Overall, in this section I was able to predict possible functions of a part of DBD- isoforms produced by both C2H2 ZF and other TFs. The most frequent predicted function of

the DBD- isoforms in both TF groups was the reversion of the mode of transcription regulation exerted by the DBD+ isoforms. In the next section, I use these predictions, as well as my functional annotation of all non-DBDs matched in all TF isoforms, to analyse switches of TF isoforms across adult human tissues.

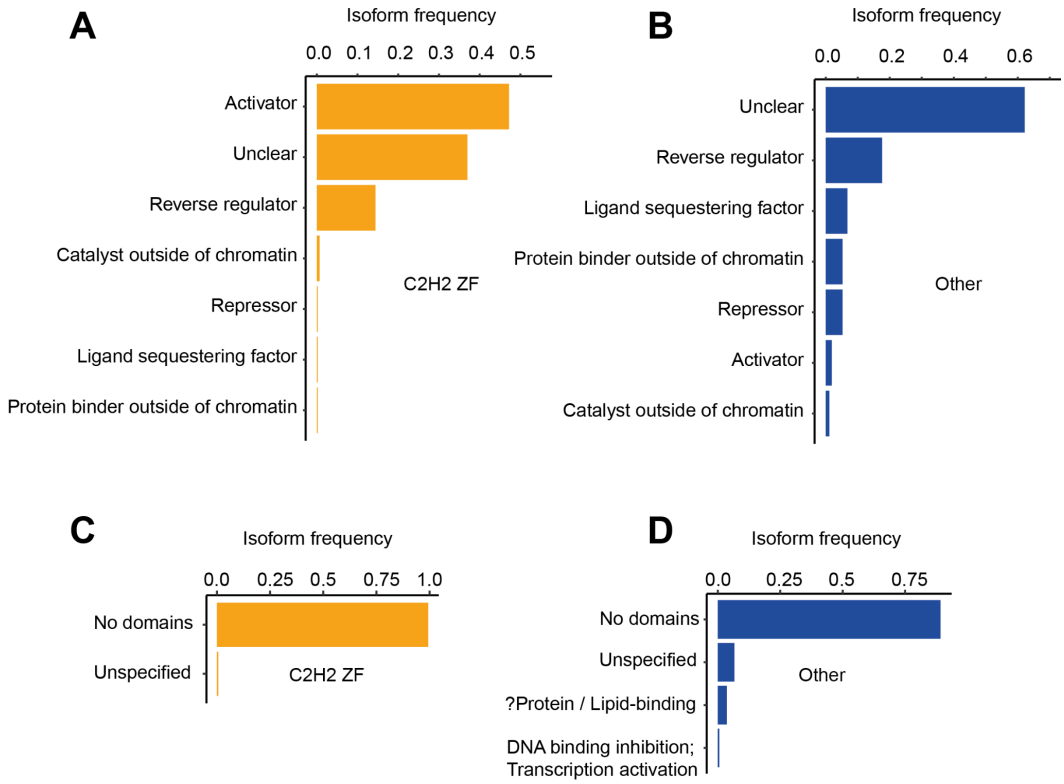


Fig. 4.10. Proportions of DBD- isoforms produced by the C2H2 zinc finger TFs and other TFs. (A) C2H2 zinc finger TFs. **(B)** Other TFs. The isoforms are stratified by the predicted function. **(C, D)** Proportions of the DBD- isoforms with the "Unclear" function by the presence and type of non-DBDs. The majority of these isoforms do not have any domains. See section 4.2.1 for methodological details.

Table 4. Predicted functions of DBD- isoforms. Predictions are based on the combinations of the general predicted functions of non-DBDs.

Combination of non-DBD functions	Functional prediction	Prediction justification
Catalytic	Catalyst outside chromatin	I assume that without a DBD the protein acts as a catalyst outside chromatin.
Catalytic; Oligomerization; Other	Reverse regulator	Fig. 3.10B (I cannot interpret the "other" function but include possible catalytic activity into the reverse transcription regulation.)
Dimerization	Reverse regulator	Fig. 3.10B
Dimerization; Ligand binding; Transcription activation	Repressor	Fig. 3.10B-C (I include possible ligand sequestration by the ligand binding domain into the reverse transcription regulation.)
Dimerization; Protein binding	Reverse regulator	Fig. 3.10B (I include any additional protein sequestration by the protein binding domain into the reverse transcription regulation.)
Dimerization; Transcription activation	Repressor	Fig. 3.10B-C
Dimerization; Transcription regulation	Reverse regulator	Fig. 3.10B-C
Dimerization; Transcription repression	Activator	Fig. 3.10B-C
DNA binding inhibition; Transcription activation	Unclear	Although the transcription activation domain may sequester a coactivator (Fig. 3.10C), the role of the DNA binding inhibition domain is not clear in the absence of the DBD. Consequently, I predicted an unclear function for a DBD- isoform.

Ligand binding	Ligand sequestering factor	I predicted ligand sequestration by analogy with the sequestration of cofactors by transcription regulation domains and the sequestration of partner TFs by dimerisation domains.
Ligand binding; Transcription repression	Activator	Fig. 3.10C (I include possible ligand sequestration by the ligand binding domain into the reverse transcription regulation.)
Ligand binding; Unspecified	Ligand sequestering factor	I predicted ligand sequestration by analogy with the sequestration of cofactors by transcription regulation domains and the sequestration of partner TFs by dimerisation domains. (I cannot interpret an unspecified function of a domain.)
No domains	Unclear	Domains are functional units of TFs, and without domains the function of a protein is unclear.
Oligomerization	Reverse regulator	Fig. 3.10B (In my predictions, I consider an oligomerization domain in the same way as a dimerization domain.)
Oligomerization; Other	Reverse regulator	Fig. 3.10B (I cannot interpret the “other” function.)
Oligomerization; Transcription repression	Activator	Fig. 3.10B-C
Protein binding	Protein binder outside chromatin	By analogy to 3.10B.
Protein binding; Protein folding; Transcription regulation	Reverse regulator	Fig. 3.10C (I do not interpret the presence of the protein folding domain but include any additional protein sequestration by the protein binding domain into the reverse transcription

		regulation.)
Protein binding; Transcription repression	Activator	Fig. 3.10C (I include any additional protein sequestration by the protein binding domain into the reverse transcription regulation.)
Protein binding; Unspecified	Protein binder outside chromatin	By analogy to 3.10B (I cannot interpret an unspecified function of a domain.)
Protein folding; Transcription regulation	Reverse regulator	Fig. 3.10C (I do not interpret the presence of the protein folding domain.)
?Protein / Lipid-binding	Unclear	I cannot interpret a very general domain function.
?Protein-binding; Protein binding	Protein binder outside chromatin	By analogy to 3.10B.
Transcription activation	Repressor	Fig. 3.10C
Transcription regulation	Reverse regulator	Fig. 3.10C
Transcription regulation; Unspecified	Reverse regulator	Fig. 3.10C (I cannot interpret an unspecified function of a domain.)
Transcription repression	Activator	Fig. 3.10C
Unspecified	Unclear	I cannot interpret an unspecified function of a domain.

4.4. Discussion

4.4.1. Summary and interpretation of the results

First, I investigated the loss of DBDs in TF isoforms and found that 27.5% of human TFs may lose DBDs in some isoforms. However, the presence of a DBD in *all* isoforms may be functionally important for TFs from particular families (for instance, "Homeodomain; POU" or "Grainyhead"). Additionally, the fact that TF families of comparable sizes may have quite different proportions of TFs producing DBD- isoforms suggests that different families of TFs evolved different uses of (or at least, tolerance to) DBD- isoforms. Specifically, the abundance of DBD- isoforms in TF families such as "C2H2 ZF; Homeodomain" and "CxxC ZF" may have become beneficial, or at least neutral, during the evolution of these families, while in families such as "CENPB; Pipsqueak" the production of DBD- isoforms for some reason may have been detrimental.

Moreover, differences in DBDs between isoforms, such as variation in the DBD sequence or the number of DBDs, may affect the specificity or stability of DNA-binding. I found that 85% of single-DBD TFs produce just one unique DBD sequence. Consequently, the DBD+ isoforms of these TFs may have the same or very similar DNA binding properties. However, some TFs change their DNA binding specificity depending on an amino acid sequence outside the DBD. For instance, two DBD+ isoforms of WT1 that differ by the presence or absence of a three-amino acid insertion *between* two DNA-binding C2H2 zinc fingers (+KTS and -KTS isoforms, respectively) have different DNA binding specificities (Bickmore et al., 1992).

Additionally, those 15% of single-DBD TFs that produce more than one unique DBD sequence never change the DBD type across their isoforms. Nevertheless, this fact does not necessarily imply close DNA binding properties of the DBD+ isoforms produced by the same TF but with different DBD sequences. Finally, while studying multi-DBD TFs, I observed that 70% of them retain the same DBD combination across their DBD+ isoforms and hence may produce DBD+ isoforms with similar DNA binding characteristics. This does not take into account, however, the possible presence of other amino acid sequences that may affect DNA binding.

Returning to my very first result in the DBD analysis, namely the fact that almost a third of all human TFs lose a DBD in alternative isoforms, Dr. Koustav Pal and I studied the

DBD- isoforms in more detail. We found that a half of DBD- isoforms were expressed in adult human tissues, and from these a third became major in at least one tissue. This fact suggests the existence of a functional role for at least a subset of DBD- isoforms, and we also speculate that at least a part of the DBD- isoforms that are not expressed in adult human tissues could be expressed transiently during development.

Moreover, we found that TFs that are broadly expressed across tissues have broadly expressed DBD+ isoforms but quite tissue-specific DBD- isoforms. This observation further supports the possible functional importance of at least some DBD- isoforms of human TFs. Finally, we found that DBD- isoforms may have different median expression levels or tissue specificities in different TF families, which suggests different roles of these isoforms in different families. In summary, our analysis supports the possible functional importance of DBD- isoforms of human TFs and suggests their further experimental study.

Next, using my functional annotation of non-DBDs, I computationally predicted possible functions of DBD- isoforms produced by human TFs. The most frequently predicted function was the reversion of the mode of transcription regulation by a DBD+ isoform, so that upon the loss of a DBD an activator became a repressor and vice versa. However, I was not able to predict the functions of 37% of DBD- isoforms produced by C2H2 zinc finger TFs and of 62% of DBD- isoforms produced by other TFs. The main reason was the absence of any domains in these isoforms. Consequently, such isoforms could either be non-functional (and, possibly, even untranslated) or, if translated, lack a stable folding (be disordered). In general, disordered proteins are able to bind DNA, RNA or proteins and, in this way, may have a function (Brodsky et al., 2020; Hisaoka et al., 2014; Minezaki et al., 2006).

Importantly, our observations summarised above do not imply a functional role for each alternative TF isoform but are also compatible with the possibility that the alternative isoforms and their expression patterns emerge via neutral evolution. Indeed, sequence motifs that influence the production of alternative isoforms by the same TF gene, such as initiator elements in core promoters, splice sites and polyadenylation sites, are quite short (Lenhard et al., 2012; Marasco and Kornblihtt, 2022; Neve et al., 2017; Vo ngoc et al., 2017), and hence may emerge and disappear in short evolutionary time periods. Consequently, if the production of a particular RNA isoform of a TF becomes likely due to, for instance, the emergence of a specific splice site, and the isoform is not detrimental, then such an isoform will not be eliminated by negative selection; however, at the same time, it will not necessarily attain a specific function. In the future, the analysis of evolutionary conservation, as well as experimental studies, that I briefly

discuss in section 5.4.3 would help to establish the scale of the functional importance of alternative TF isoforms.

4.4.2. Limitations of the analysis

My functional annotation of DBD- isoforms (and of major isoform switches that I found and analysed in Chapter 5) was limited by the domain collection of the InterPro database. For example, in InterPro some classical dimerisation domains, such as a helix-loop-helix and a leucine zipper, are not annotated separately from the basic domain and therefore in my analysis they were considered a part of the DBD. Also, InterPro frequently lacks transcription activation domains and some other types of non-DBDs. For instance, ZGPAT lacks the coiled-coil transcription repression domain, because it is not counted as a domain in InterPro; NFYA, PAX5, ESR1, NFATC1 and TCF4 lack matches of known transcription activation domains; TCF3 lacks a match of a known DBD inhibition domain. Hence, further development of the InterPro database would make the analysis of TF isoforms more accurate.

Of note, the lack of transcription activation domains in InterPro is likely due to the fact that these domains are characterised by intrinsic disorder (Dyson and Wright, 2005; Liu et al., 2006a) and hence, in contrast to other domains, cannot be predicted based on the structural conservation. Consequently, transcription activation domains could be present among IDRs predicted by InterProScan; however, these IDRs lack functional annotation. In future studies, this problem could be mitigated by using TFRegDB (https://tfregdb.bu.edu/tfregdb/) (Soto et al., 2022), a literature-based resource containing evidence about transcription activation domains, and also by using software tools that predict transcription activation domains in a given protein sequence, such as PADDLE (https://paddle.stanford.edu/) (Sanborn et al., 2021) or ADpred (https://adpred.fredhutch.org/) (Erijman et al., 2020).

Additionally, I could not assign a definite function to some domain matches neither based on the InterPro annotation, nor based on literature. Finally, domains may be more versatile than we tend to assume, and hence, the “one domain – one function” annotation that I developed in my analysis may not always be adequate. For example, some homeodomains, apart from DNA binding, can also mediate protein-protein interactions (Lai et al., 1992).

Next, my analysis of DBD- isoforms has two limitations. First of all, I assumed that all DBD+ isoforms are able to bind DNA, while in reality some of them may not be DNA-binding for at least two reasons. One is that the ability of a C2H2 ZF isoform to bind DNA may be lost even when some C2H2 zinc fingers are still present (Beverly and Capobianco, 2003). Another reason is the inability of TFs that dimerise in order to bind DNA, such as the bHLH (Jones,

2004) and bZIP (Bader and Vogt, 2006) factors, to bind DNA as monomers upon the loss of a dimerisation domain. Consequently, I likely underestimated the actual number of alternative TF isoforms that are not able to bind DNA. Secondly, I assumed that DBD- isoforms do not bind chromatin but instead always sequester their binding partners from the chromatin and in this way reverse the mode of transcription regulation exerted by their DBD+ counterparts. While this is true for some well-studied cases (Fernandez and Gudas, 2009; Qi et al., 2016; Sun et al., 1996; Yu et al., 2010; Zhao et al., 2016), DBD- isoforms that bind chromatin and confer the same function as their DBD+ counterparts also exist (Noro et al., 2006).

Finally, in section 5.4.2, I describe several more limitations that pertain to the analysis described both in Chapter 4 and Chapter 5.

4.4.3. Perspective

Here, I outline potential projects that stem from my results described in this chapter.

First, my primary supervisor Professor Nicholas Luscombe noted that if one manages to formally define isoform conservation between species, then certain interesting questions may become accessible for study. Considering the investigation of the functional importance of the DBD- TF isoforms, Professor Luscombe suggested that it would be informative to know how many DBD- isoforms, including domainless ones, are conserved between human and, for instance, mouse or macaque. Which TFs produce them and in which tissues? These questions are interesting, because the conservation of a DBD- isoform would suggest its functional importance. Noteworthy, software tools for the assessment of isoform conservation, such as ThorAxe (Zea et al., 2021), have been emerging, making this analysis possible.

Furthermore, any experimental study of the isoform-specific DNA binding of a TF *in vivo* is complicated by the need to use isoform-specific antibodies which may not be easily available or produced. Nevertheless, I envision a computational project that could help to tackle this problem. Namely, one could develop a pipeline which takes as input a pair of alternative isoforms produced by the same TF but containing different DBDs and searches publicly available RNA-seq data and DNA binding data (such as ChIP-seq, ChIP-exo or CUT&RUN) for pairs of primary human tissues such that one tissue almost exclusively expresses one isoform, while the other tissue almost exclusively expresses the other isoform. Then the DNA binding data found for the same TF in the same pair of tissues would demonstrate differential DNA binding of the two isoforms of this TF. Although this method requires the simultaneous presence of RNA-seq data and DNA binding data for the same TF for at least two different tissues and the transcriptional dominance of just one isoform in each tissue, it would hopefully

become more and more feasible with the growth of the amount of sequencing data from the human primary tissues. Hence, in principle, such a pipeline could be published as a web-service that automatically repeats its search across publicly available databases periodically and, hopefully, generates a growing list of results on the differential DNA binding of TF isoforms, both in healthy and diseased human tissues.

Chapter 5. Switches of major alternative isoforms of human transcription factors

5.1. Introduction

Functional differences between alternative isoforms of TFs become important only when these isoforms are expressed: either co-expressed in one cell (Qi et al., 2016) or differentially expressed between cell types in an adult or developing organism (Gabut et al., 2011). The exchange of one isoform for another across cell types or conditions is often called an **alternative isoform switch**. Thus, FOXP1 switches its isoform that maintains ESC pluripotency for another isoform that promotes cell differentiation (Gabut et al., 2011).

Another well-known example of a developmental switch of alternative TF isoforms is represented by the TF REST. Specifically, REST expresses a full-length DBD+ isoform that contains an array of 9 C2H2 zinc finger domains in neural precursor cells where this isoform represses the neural differentiation. In contrast, at the time when the neural differentiation should start, REST switches its full-length isoform for a truncated one, REST4, that contains only 4 C2H2 zinc finger domains and cannot bind DNA. Consequently, the neural differentiation program is activated (Raj et al., 2011).

Apart from developmental processes, switches of alternative TF isoforms may occur between differentiated cell types. Thus, human IRF3 can produce two different DBD+ isoforms, one with a full-length DBD and the other with a truncated DBD. The isoform with the full-length DBD binds to the IFN-β promoter and activates the transcription of the gene, while the isoform with the truncated DBD cannot bind to this promoter and consequently cannot activate IFN-β. Importantly, IRF3 switches these isoforms between the brain and non-brain tissues. Specifically, the isoform with the truncated DBD is expressed higher in the brain, while in other tissues the isoform with the full-length DBD is predominant. This switch likely leads to a considerable attenuation of the IFN-β activation in the brain (Karpova et al., 2001).

However, the systematic collection and characterisation of the cross-tissue switches of the human TF isoforms are still lacking. Therefore, in this chapter I study switches of the top-expressed (major) isoforms of 277 human TFs across 46 adult human tissues and predict functional consequences of these switches. I also outline a number of especially interesting examples of isoform switches to motivate their experimental validation in future.

5.2. Methods

Scripts for the described analysis can be found at https://github.com/ComputationalRegulatoryGenomicsICL/human-tf-isoforms-thesis.

5.2.1. Analysis of switches of major TF isoforms across adult human tissues

In section 4.2.2, I defined the major isoform of a TF in a tissue as its highest-expressed isoform in that tissue. To analyse the switches of major isoforms of human TFs (defined below), I selected the 4,089 isoforms that were obtained as described in section 4.2.2 and processed them as follows:

1. I set a threshold of 1 log-scaled TPM as a minimal isoform expression level and set to 0 all isoform expression levels that were less than 1.
2. I found major isoforms produced by TFs expressed in adult human tissues, and for each TF I considered exchanges of one of its major isoforms in one set of tissues for another major isoform in another set of tissues, if these exchanges are present. I called such an exchange a **switch** if the difference between the expression levels of the two exchanged major isoforms in both sets of tissues was at least 10%. Schematics in Fig. 5.1 demonstrate exchanges that are switches (Fig. 5.1A-B) or not switches (Fig. 5.1C-D).

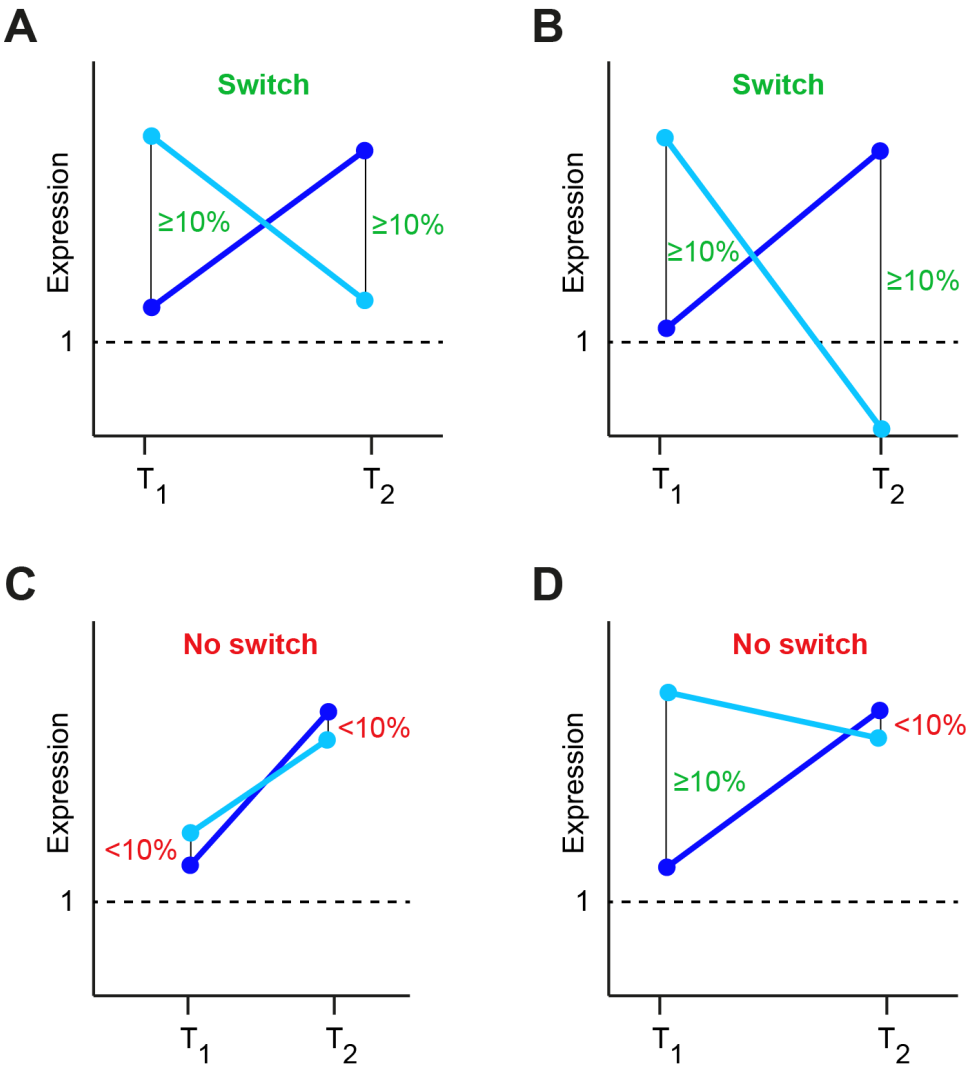


Fig. 5.1. The definition of a major isoform switch. The schematics show the exchange of two major isoforms by the same TF between two tissues, T_1 and T_2. One of the isoforms becomes major in T_1, while the other isoform becomes major in T_2. The expression value of 1 marks the minimal expression cutoff that I set in my analysis. **(A, B)** In both tissues, the difference between the expression levels of the exchanged isoforms is greater than 10%, hence these exchanges are switches. **(C)** In both tissues, the difference between the expression levels of the exchanged isoforms is less than 10%, hence this exchange is not a switch. **(D)** Although in tissue T_1 the difference between the expression levels of the two isoforms is greater than 10%, it is less than 10% in tissue T_2; hence, this exchange is not a switch.

This isoform processing and the non-DBD annotation described in section 4.2.1 allowed me to predict the functional consequences of the switches of major isoforms, including the importance of changes in DBD sequences. It also enabled my analysis of the

overrepresentation of TF families in the functional categories of switches (see the Results section 5.3.1).

For the analysis of changes in DBD sequences, I used the Nucleic acid-Protein Interaction DataBase, NPIDB (Kirsanov et al., 2013; Zanegina et al., 2016), version from 27/12/2021. It allowed me to annotate contacts to nucleotide bases and the sugar-phosphate backbone by residues in DBDs. To align multiple DBD sequences to the corresponding annotated sequence from NPIDB, I used the ClustalOmega method implemented in the *msa* function from the *msa* R package (Bodenhofer et al., 2015) version 1.22.0.

To find the overrepresentation of TF families and GO terms (gene sets) in all switching TFs, their clusters and functional categories (signatures), I used the hypergeometric test (Fig. 5.2) implemented in the function *hypeR* of the R package *hypeR* (Federico and Monti, 2020) version 1.6.0. As a background, I used all TFs that expressed multiple major isoforms. When testing the overrepresentation of any TF family or GO term in the set of switching TFs, I adjusted raw p-values across all tests using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) implemented in the function *hypeR*. When testing the overrepresentation of any TF family or GO term in any cluster or functional category of switching TFs, I adjusted the raw p-values of all pairwise tests using the Benjamini-Hochberg procedure implemented in the function *p.adjust* of the R package *stats* version 4.0.3. For the p-value adjustments, I set the FDR to 5%.

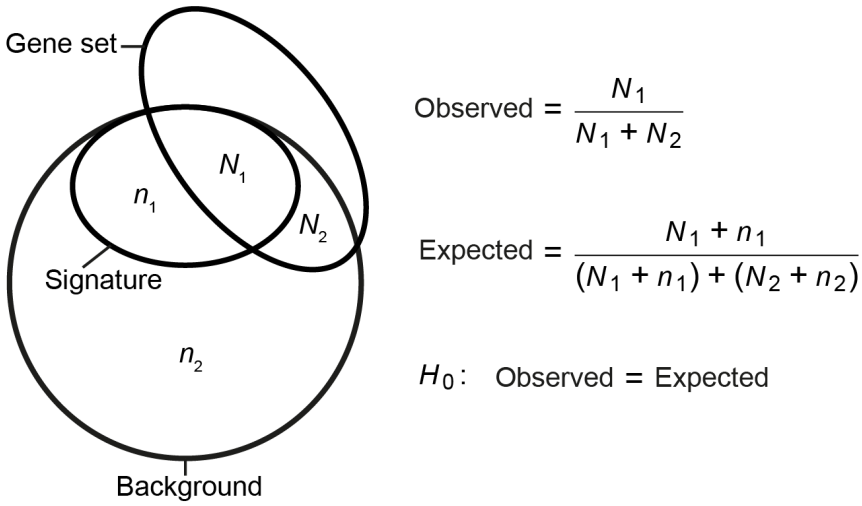


Fig. 5.2. Hypergeometric test of the overrepresentation of a gene set in a signature, in comparison to the background.

5.2.2. Analysis of switch patterns across adult human tissues

Next, to analyse characteristic switch patterns, I did the following preparatory steps:

1. For each switching major isoform, I defined its **tissue list** as a set of tissues where the isoform is major and its expression is ≥10% higher than the expression of at least one isoform with which it switches (its **switching partner**). As an example, the schematic in Fig. 5.3 shows the transcription levels of three isoforms (A, B and C) of the same TF across 10 tissues (T_1, T_2, …, T_{10}). Isoform A switches with isoform B, and the tissue list of isoform A is {T_2, T_3, T_4, T_5}, because in these tissues the expression of the major isoform A is ≥10% higher than the expression of isoform B. In turn, the tissue list of isoform B is {T_6, T_7, T_8}, because in these tissues the expression of the major isoform B is ≥10% higher than the expression of isoform A. Isoform C is expressed in all 10 tissues and becomes major in one of them (T_{10}), but it does not switch with the other two isoforms and, consequently, does not have a tissue list.
2. I defined a *switching TF* as a TF with at least one switch between its major isoforms. For each switching TF, I ranked its switching major isoforms by the decreasing size of their tissue lists and gave sequential ranks to isoforms with equal sizes of tissue lists. For example, if a TF has 4 switching major isoforms and the sizes of their tissue lists are 3, 2, 2, and 1, then these isoforms will have ranks 1, 2, 3, and 4, respectively. The ranking order of isoforms with equal sizes of tissue lists is arbitrary. Fig. 5.3 provides another example of ranking of switching major isoforms. Specifically, isoform A has rank 1, isoform B has rank 2, and isoform C does not have a rank because it does not switch with the other two isoforms. Biologically speaking, this ranking prioritises switching isoforms that dominate other switching isoforms most *frequently* across tissues. Indeed, in Fig. 5.3, isoform A with rank 1 dominates B in 4 tissues, while isoform B with rank 2 dominates A in only 3 tissues.

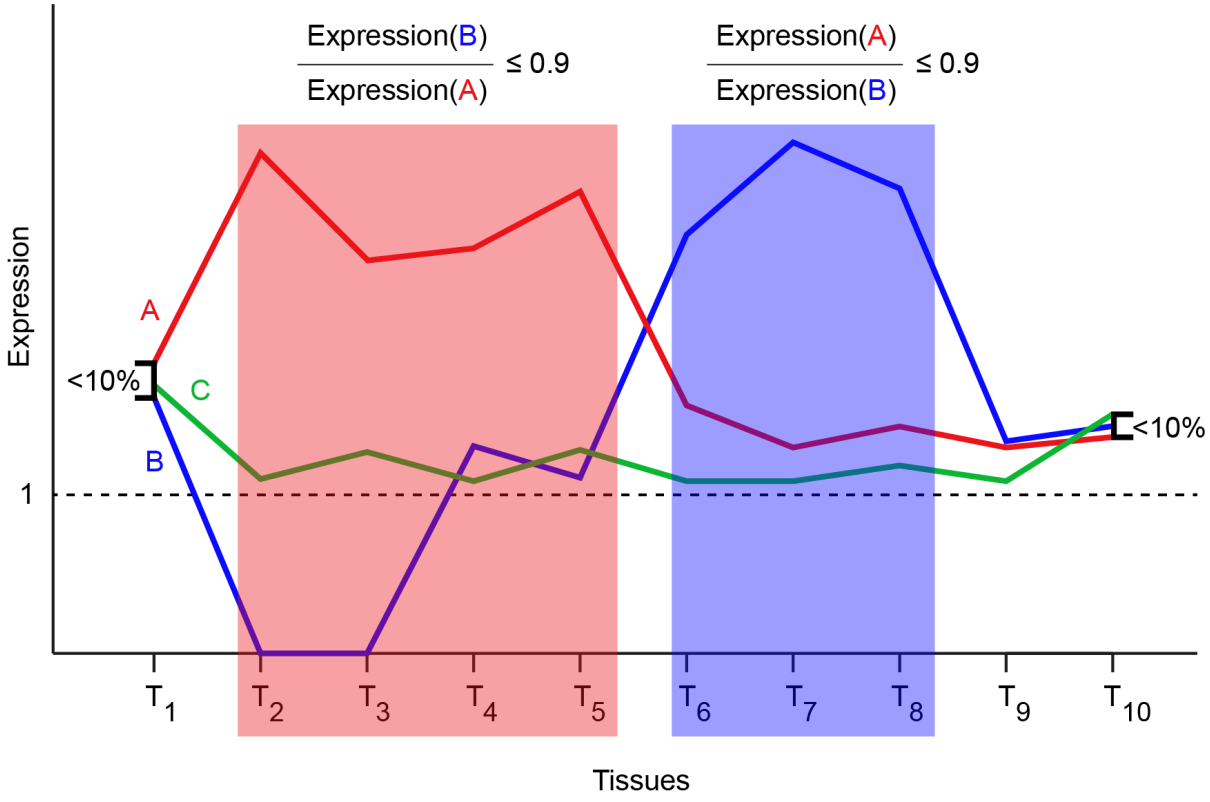


Fig. 5.3. Switches between isoforms of the same TF. Isoform A switches with isoform B, while isoform C does not switch with the other two isoforms. The red rectangle marks the tissue list of isoform A, while the blue rectangle marks the tissue list of isoform B. See the main text for details.

3. To reveal the isoform switching patterns of TFs across adult human tissues, I scored each switching TF in each tissue according to the ranks of switching isoforms that the TF expresses in that tissue and the expression levels of these isoforms. Specifically, let F be a switching TF and T be a tissue. Because F is a *switching* TF, it has at least two switching major isoforms. Let I be a switching isoform of F that is major in tissue T and J be the switching isoform of F that is second highest-expressed in tissue T. For instance, in Fig. 5.3, in tissue T_3, $I \equiv A$ and $J \equiv B$, while in tissue T_7 $I \equiv B$ and $J \equiv A$. Additionally, let $e(I)$ and $e(J)$ be the expression levels of isoforms I and J in tissue T. By definition, $e(I) > e(J)$, and if $e(I) = 0$, then F is not expressed in T. Next, let $r(I)$ and $r(J)$ be the ranks of isoforms I and J. Of note, as I and J are *switching* major isoforms, their ranks are defined. Thus, I calculated score S of each switching TF F in each tissue T in the following way:

$$S_T(F) = \begin{cases} r(I), & \text{if } r(I) \in \{1,2,3\},\ e(J)/e(I) \leq 0.9,\ I \text{ is DBD+}, \\ -r(I), & \text{if } r(I) \in \{1,2,3\},\ e(J)/e(I) \leq 0.9,\ I \text{ is DBD}-, \\ 4, & \text{if } r(I) = 1,\ r(J) = 2,\ e(J)/e(I) > 0.9, \\ 5, & \text{if } r(I) > 3 \text{ and } e(J)/e(I) \leq 0.9, \text{ or } r(I) > 1 \text{ and } e(J)/e(I) > 0.9, \\ 0, & \text{otherwise } (F \text{ is not expressed in } T). \end{cases}$$

Importantly, in the first four cases I assume that F is expressed in T and, consequently, $e(I) > 0$. This scoring formula has the following biological meaning. In the first two cases, the major switching isoform I dominates all other switching isoforms in tissue T (the difference between the expression levels of the major switching isoform and the second-highest expressed switching isoform is ≥10%); hence, the score of TF F in that tissue equals the rank of I, if I is DBD+, or the negated rank of I, if I is DBD-. For example, if in Fig. 5.3 isoform A were DBD+ and isoform B were DBD-, then the TF that expresses these isoforms, would have a score of 1 in tissues T_2, T_3, T_4, T_5 and a score of -2 in tissues T_6, T_7 and T_8. Next, the third case corresponds to a situation when the two top-ranked switching isoforms are tightly co-expressed in tissue T (the difference between their transcription levels is <10%); consequently, neither of these two isoforms dominates in T. For instance, in Fig. 5.3, the TF would have a score of 4 in tissues T_1, T_9 and T_{10}. The fourth case summarises situations which are not informative for my analysis. Namely, F obtains a score of 5 if a low-ranked switching isoform dominates other switching isoforms or if non-top-ranked switching isoforms are tightly co-expressed in tissue T. Finally, F is assigned a score of 0 if it is not expressed in T.

4. After scoring the switching TFs across tissues, I generated a matrix of these scores, with tissues in rows and TFs in columns, and hierarchically clustered rows and columns as vectors of TF scores defined above, using the Hamming distance and the complete linkage. I did the clustering with the function *Heatmap* from the R package *ComplexHeatmap* (Gu et al., 2016) version 2.7.9.1005. I defined the numbers of the TF and tissue clusters manually, based on their interpretability.
5. To visualise the clustering results as a heatmap, I colour-coded the TF scores. Specifically, scores 1, 2 and 3 corresponded to shades of red, from dark to pale,

scores -1, -2 and -3 corresponded to shades of blue, also from dark to pale, score 4 corresponded to a green colour, score 5 corresponded to an orange colour, and score 0 corresponded to a grey colour (Fig. 5.4). I plotted this heatmap using the function *ComplexHeatmap*.

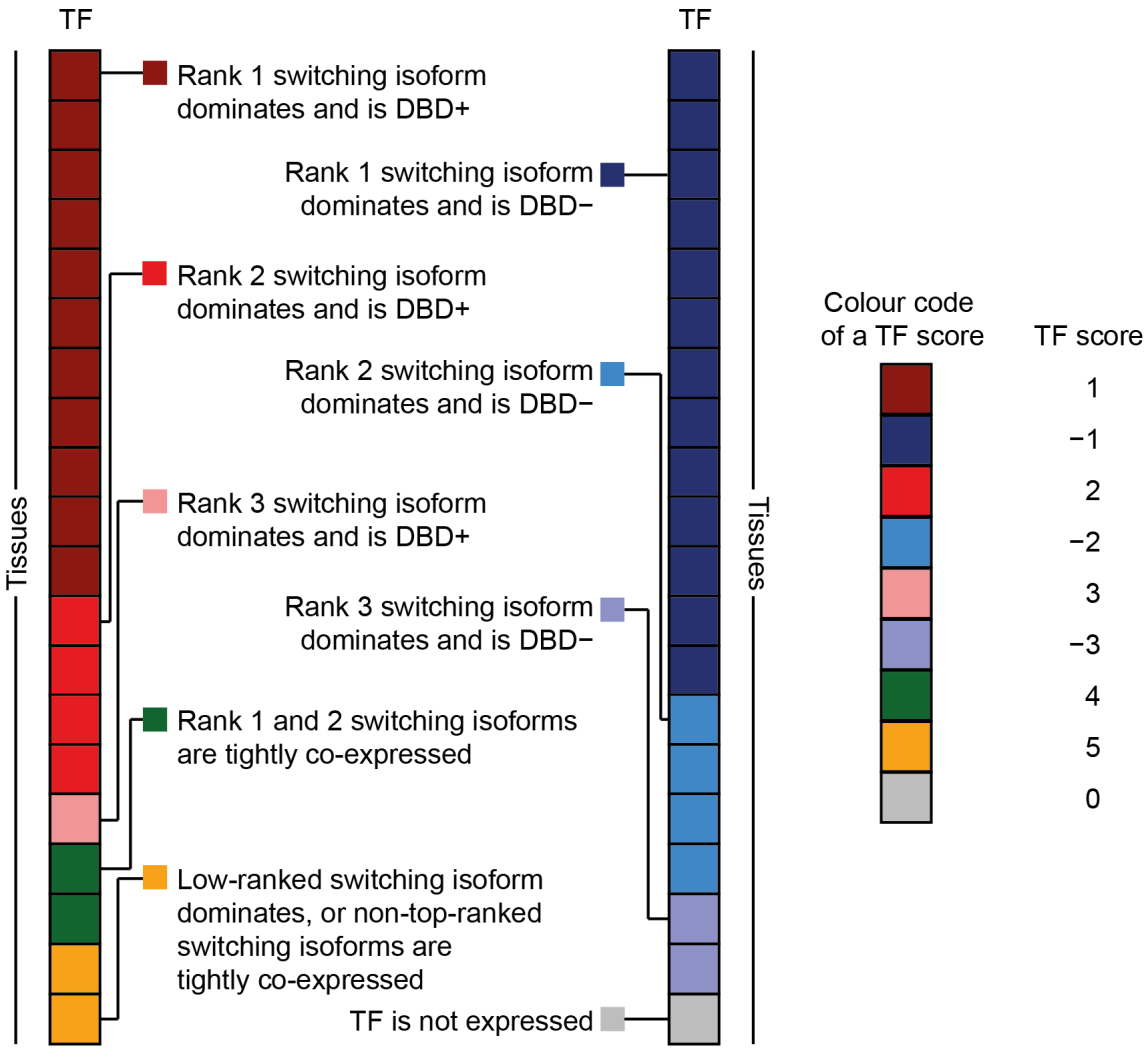


Fig. 5.4. Colour-coding of TF scores. See the main text for details.

These preparatory steps allowed me to analyse the characteristic patterns of switches of major TF isoforms across adult human tissues (see the Results section 5.3.2). For the principal component analysis of the selected tissue samples from GTEx, I used the function *prcomp* from the R package *stats* (R Core Team, 2020) version 4.0.3.

5.3. Results

5.3.1. TFs switch major isoforms across adult human tissues with predictable functional outcomes

In the last part of my analysis, I explored possible functional differences between TF isoforms in the context of their expression quantified in bulk samples of adult human tissues by the GTEx Consortium. Because the quantification of the isoform expression from bulk RNA-seq data averages the expression levels across the whole tissue sample, it does not allow discrimination between the co-expression of several TF isoforms in one cell and their expression in different cells of the same tissue. Hence, I predicted functional consequences of the differential expression, rather than co-expression, of TF isoforms between adult human tissues.

For this, I used the 4,089 coding isoforms produced by the 1,488 TFs that I selected before. Among these, I considered the major isoforms (N = 1,777) and defined a switch of two major isoforms of a TF as an exchange of its major isoform expressed in one group of tissues for another major isoform expressed in another group of tissues. Additionally, I required that the difference (δ) between the expression levels of the exchanged isoforms should be at least 10% (Fig. 5.5A). Of note, major isoforms (N = 1,777) outnumber TFs (N = 1,488) because different isoforms of the same TF may become major in different tissues.

Using these definitions, I calculated the percentage of TFs with different numbers of major isoforms across all tissues and with different numbers of switches (Fig. 5.5B). I found that 327 (22%) of all human TFs in my analysis express more than one major isoform across at least two tissues and hence have a potential to switch them. Indeed, the vast majority of these TFs (277, or 85%) produce at least one switch (Fig. 5.5B; Supplementary Table 3). In total, I found 1,777 major isoforms, and 616 (35%) of them were switching.

Of note, fewer TFs produce two switches than three switches (Fig. 5.5B). It could be explained by the design of my analysis. Indeed, to form two switches, two out of three major isoforms should be expressed at the levels that are less than 10% apart ($\delta < 10\%$), while in the case of three switches, the three switching major isoforms can be expressed at any levels, just different enough from each other ($\delta \geq 10\%$).

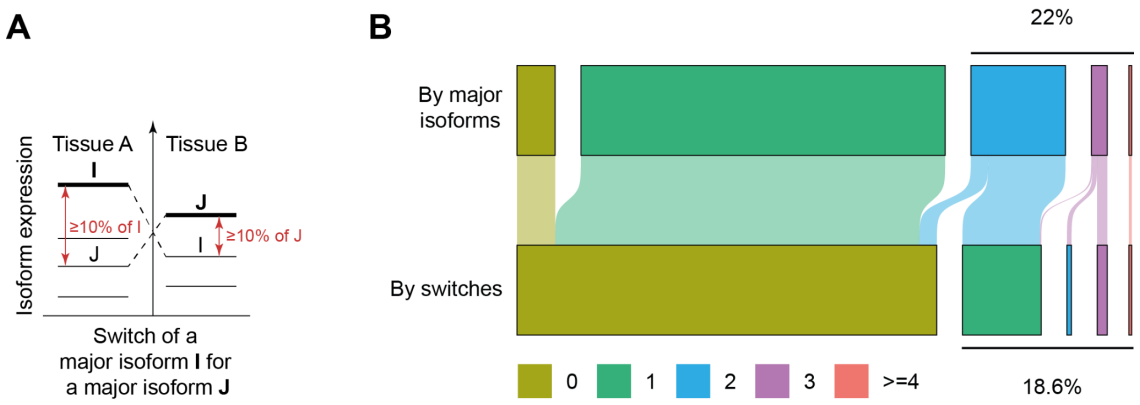


Fig. 5.5. Switching TFs. (A) Schematic of a switch of two major isoforms, I and J, of the same TF between tissues A and B. **(B)** Proportions of TFs by the number of major isoforms (above) and by the number of switches (below). Zero major isoforms mean that a TF is not expressed in adult human tissues. See section 5.2.1 for methodological details.

Because most TFs that produce more than one major isoform switch their major isoforms, I next decided to characterise switches in more detail and predict their functional consequences (Fig. 5.6). I observed a total of 400 switches (unique pairs of exchanging major isoforms with a minimum of 10% difference in expression levels) in all switching TFs (Supplementary Table 3). 75% of these switches were from one DBD+ major isoform to another DBD+ one, yet 22% were between a DBD+ and a DBD- major isoform, and only 3% of switches involved two DBD- isoforms (Fig. 5.6A, left). The low percentage of the latter is expected, as major DBD- isoforms are much more rare than major DBD+ isoforms (Fig. 4.6A-B and section 4.3.2). Interestingly, only 39% of DBD+ to DBD+ (+/+) switches affected any defined protein domains, while another 40% of the +/+ switches affected the protein sequence outside domains and the remaining 21% did not affect the protein sequence at all (Fig. 5.6A, right) and hence must have affected UTRs.

Based on my functional annotation of non-DBDs in TF isoforms (Supplementary Table 2) and on the published experimental data (Fig. 3.10-3.11), I predicted possible functional consequences of the +/+, +/- and -/- switches (Fig. 5.6B; Supplementary Table 3). As expected from the above, the majority of the +/+ switches were predicted to affect UTRs and the protein sequence outside domains. Also, a change in DNA binding was among the top predicted consequences, which warranted an investigation into possible differences between switching DBD sequences. Next, the scarce -/- switches grouped into four different categories among which the “Change outside of domains” was the biggest one. However, an additional interesting category of the -/- switches was the “Activator / Domainless protein.” It comprised switches

produced by two C2H2 ZF factors, ZNF268 and ZNF655, which differentially included the KRAB domain in their otherwise domainless DBD- isoforms. Finally, the +/- switches were mainly predicted to reverse the transcription regulation (turn transcription activation into repression and vice versa) or turn TFs into domainless proteins (Fig. 5.6B).

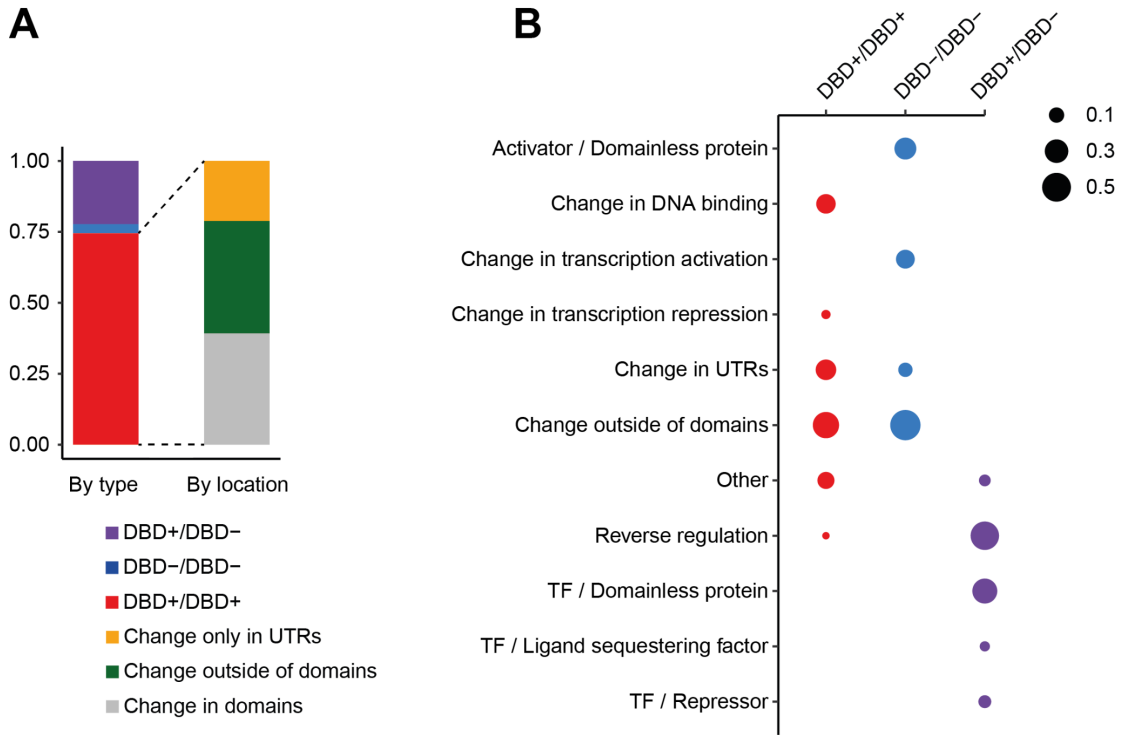


Fig. 5.6. The types and functional categories of major isoform switches. (A) Proportions of the DBD+/DBD-, DBD-/DBD- and DBD+/DBD+ switches. **(B)** Proportions of predicted functional categories of the three types of switches. See section 5.2.1 for methodological details.

Next, I considered the functional switch categories in more detail. Specifically, as "Change in DNA binding" was one of the top three most frequent +/+ switch categories, I studied DBD sequence changes in these switches and predicted possible functional consequences of these changes. To this end, from the whole set of 298 +/+ switches, I selected those occurring between isoforms with only one DBD (N = 130). This simplification made the interpretation of the results more straightforward, as it would be unclear how to detect changes between several DBDs. Next, from all the switches between single-DBD isoforms, I selected those which affected the DBD sequence (N = 26). They were produced by 14 TFs: CREM, ERF, FOSL2, MBD2, MEIS3, NFATC1, PGR, PITX2, RARB, TBPL1, TBX3, TCF3, THAP5, and THAP6. To predict the consequences of changes in the DBDs, I used the Nucleic acid-

Protein Interaction DataBase, NPIDB (Kirsanov et al., 2013; Zanegina et al., 2016), to annotate DBD residues that confer binding specificity (through the binding to DNA bases) or binding stability (via the binding to the sugar-phosphate backbone of the DNA). From the selected 14 TFs, 7 had protein-DNA structures in PDB that were annotated in the NPIDB: ERF (PDB IDs 7JSA and 7JSL), MBD2 (PDB ID 6C1A), NFATC1 (PDB ID 1A66), PGR (PDB ID 2C7A), PITX2 (PDB ID 2LKX), RARB (PDB ID 5UAN), and TCF3 (PDB ID 2YPB). I aligned isoforms of these TFs to the corresponding protein chains of their structures annotated in the NPIDB and propagated the annotation of the base- and backbone-binding residues to the isoforms (Fig. 5.8-5.9).

For MBD2 and TCF3, I found that the switching isoforms of these TFs actually have the same DBD sequence, but for an unclear reason, the boundaries of the DBD match differ by one amino acid between isoforms in each pair (Fig. 5.7A-B). In addition, in the switching isoforms of PITX2, DBD matches were very "sensitive" to the changes in the C-terminal part of the protein. Indeed, in isoform PITX2-207 the DBD was matched at the very end of the sequence shared between the two isoforms, while in PITX2-202 it was matched with a shift into the isoform-specific sequence. This did not allow me to definitively call a change in the DBD in the switching isoforms of PITX2 (Fig. 5.7C). Consequently, I excluded MBD2, TCF3 and PITX2 from this part of my analysis and proceeded with the other four TFs: ERF, RARB, PGR and NFATC1.

I found that in its +/+ switch, ERF loses the N terminus that contains more than a half of the DBD and a third of the backbone-contacting residues; however, residues that confer binding specificity of both the ERF dimer (Fig. 5.8A) and the ERF tetramer (Fig. 5.8B) are retained. Consequently, if the folding of the DBD trunk in the isoform ERF-202 is similar to the folding that this region obtains in the full DBD, then the DBD trunk could still bind DNA with the same specificity as the full DBD, although, presumably, the binding would be weaker. Of note, the ERF tetramer makes three more contacts with the DNA, in comparison to the dimer, but the conclusion for the dimer also holds for the tetramer. In contrast, RARB loses all residues that facilitate binding specificity and three out of five of its residues that confer binding stability; consequently, the remaining DBD trunk may not be able to bind DNA or it may bind DNA with a different specificity, contacting DNA with different residues (Fig. 5.8C). Finally, PGR and NFATC1 both lose the most part of their specificity-conferring residues, and hence, their DBD trunks may bind DNA non-specifically or with a different specificity (Fig. 5.9A-B).

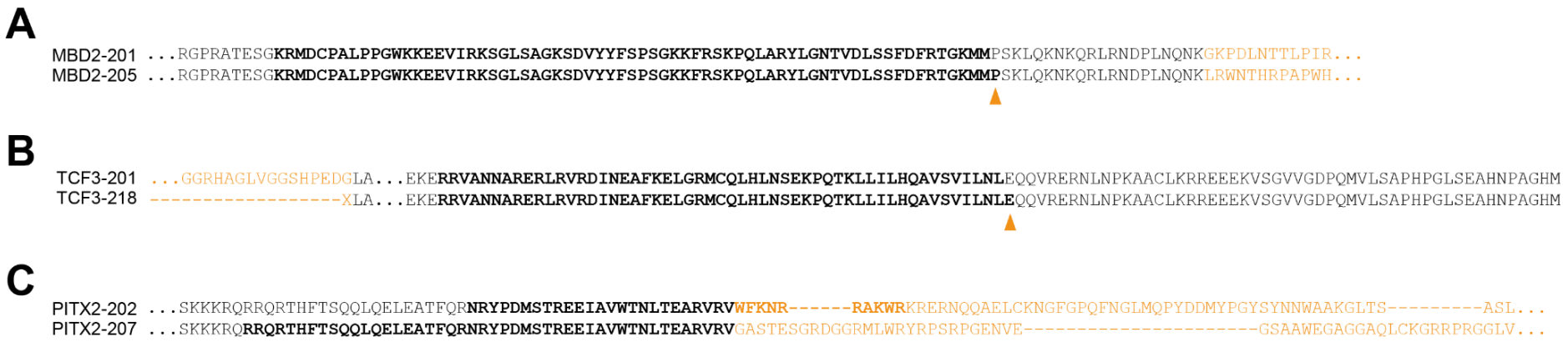


Fig. 5.7. Artificial "changes" in the DBD sequences of switching major isoforms. Aligned fragments of proteins encoded by switching major DBD+ isoforms of MBD2 **(A)**, TCF3 **(B)** and PITX2 **(C)**. Full DBDs or DBD fragments in the isoforms are shown in bold; residues that differ between isoforms in each alignment are shown in orange; full stops denote omitted residues, while dashes denote gaps in the alignment. Orange triangles in (A) and (B) point to single residues matched by a DBD in one isoform but not the other. See section 5.2.1 for methodological details.

Fig. 5.8. Changes in the DBD sequences of switching major isoforms. Part 1. (A) DNA-binding residues of the ERF dimer. **(B)** DNA-binding residues of the ERF tetramer. Coloured triangles mark additional DNA-binding residues predicted in the ERF tetramer but not in the ERF dimer. **(C)** DNA binding of RARB. In each panel, the top line shows a fragment of a DBD-containing chain from the corresponding protein structure (PDB IDs 7JSA for the ERF dimer, 7JSL for the ERF tetramer, and 5UAN for RARB), while the two lower lines show aligned fragments of proteins encoded by switching major DBD+ isoforms. Full DBDs or DBD fragments in the isoforms are shown in bold; red residues are predicted to confer binding specificity, while green residues are predicted to stabilise binding; full stops denote omitted residues, while dashes denote gaps in the alignment. See section 5.2.1 for methodological details.

Fig. 5.9. Changes in the DBD sequences of switching major isoforms. Part 2. DNA-binding residues of PGR **(A)** and NFATC1 **(B)**. In each panel, the top line shows a fragment of a DBD-containing chain from the corresponding protein structure (PDB IDs 2C7A for PGR and 1A66 for NFATC1), while the two lower lines show fragments of proteins encoded by switching

major DBD+ isoforms. Full DBDs or DBD fragments in the isoforms are shown in bold; red residues are predicted to confer binding specificity, while green residues are predicted to stabilise binding; full stops denote omitted residues, while dashes denote gaps in the alignment. See section 5.2.1 for methodological details.

After annotating switches with their possible functional consequences, I asked if any TF families or GO terms were overrepresented in TFs that switch major isoforms. To answer this question, I used a hypergeometric test of overrepresentation of TF families or GO terms (gene sets) among switching TFs (a signature), in comparison to all TFs that express multiple major isoforms (the background). With this, I did not find the overrepresentation of any TF families or GO terms in the switching TFs under the false discovery rate (FDR) of 5%. Additionally, the majority of TF families expressed in adult human tissues (namely, 47 out of 65, or 72%) had at least one TF switching its major isoforms.

Next, I asked if any TF families or GO terms were overrepresented among TFs that produce switches of particular functional categories. Using the same hypergeometric test and defining signatures as switch categories and the background as the whole set of switching TFs, I found that certain TF families were indeed overrepresented in certain switch categories under the FDR = 5% (Fig. 5.10A). Specifically, nuclear receptors were overrepresented among TFs whose switches were predicted to change ligand binding (adjusted p-value = 0.0005). This could be explained by the fact that 67% of all ligand-binding domains in my dataset are annotated in nuclear receptors (Fig. 5.10B). Additionally, MADS box TFs were overrepresented in switches that change protein binding (adjusted p-value = 0.04). This overrepresentation cannot be interpreted further, as “protein binding” is a very general term and at present cannot be made more specific in my functional annotation of non-DBDs. Finally, C2H2 ZF factors were overrepresented in the “Reverse regulation” category of switches (adjusted p-value = 0.002), where the reversal of the regulation happened predominantly via a loss/gain of a DBD or a transcription repression domain (Fig. 5.10D).

Overall, I found that 85% (277) of human TFs that express several major isoforms (hence have the potential to switch them), indeed switch their major isoforms across adult human tissues. Using my functional annotation of matched non-DBDs, I was able to predict and categorise general functional consequences of these switches, find TF families overrepresented in certain functional categories of switches and interpret switch-related changes in the DBD sequences of some TFs. Regarding the predicted functional consequences of switches, the majority of +/+ switches likely affected the non-domain protein sequence,

UTRs or the DNA binding of a protein; a change of a protein sequence outside domains was also characteristic of -/- switches, while about a half of +/- switches likely reversed the transcription regulation. Also, both +/- and -/- switches frequently replaced TFs with domainless proteins.

In the next section, I study the whole set of switches that I found in the context of the tissue specificity of the corresponding major isoforms.

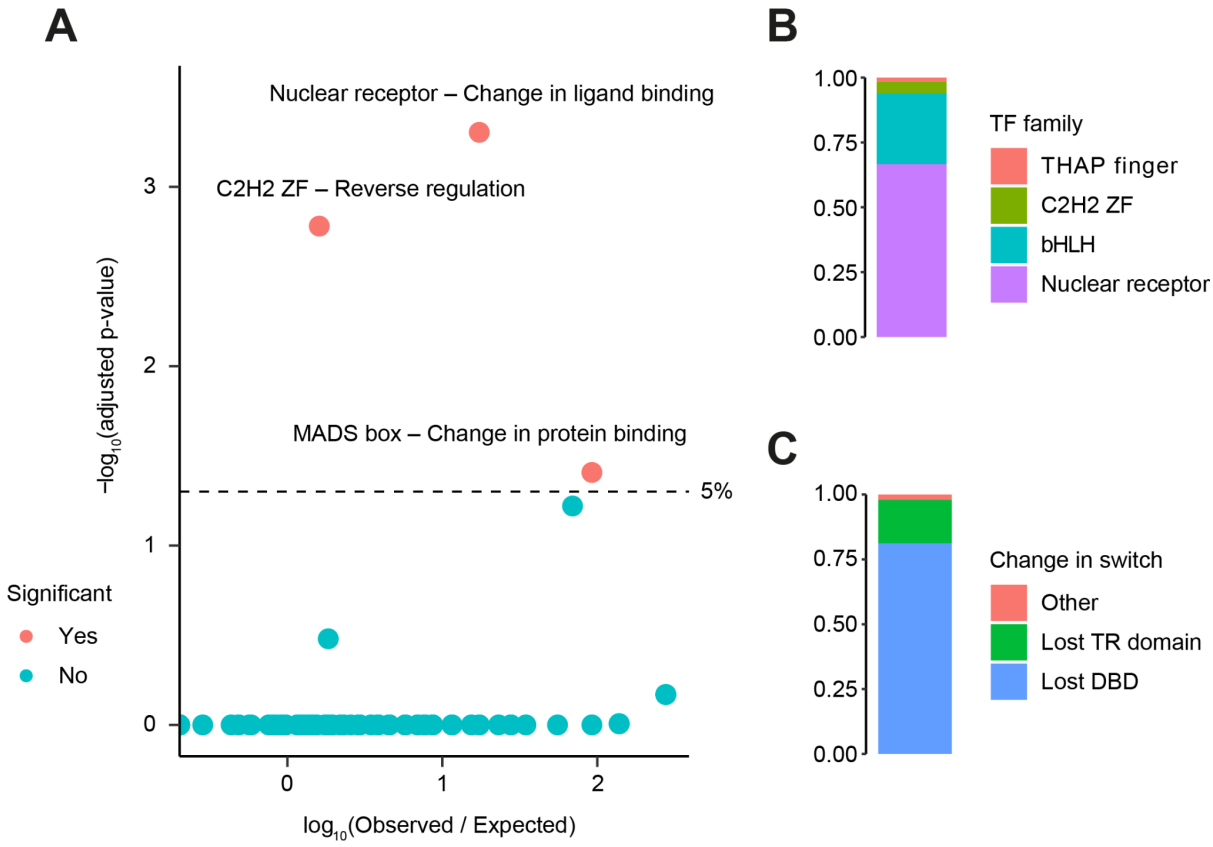


Fig. 5.10. Overrepresentation of TF families in the functional categories of switches. (A) Volcano plot of TF families significantly overrepresented in functional categories of switches (FDR = 5%). **(B)** Proportions of matches of ligand-binding domains across TF families. **(C)** Proportions of different changes that lead to a transcription regulation reversal in C2H2 ZF factors. TR stands for "transcription repression." See section 5.2.1 for methodological details.

5.3.2. Switches of major isoforms demonstrate biologically relevant patterns

After the general study of the switches of major isoforms, I investigated if these switches had characteristic patterns of tissue specificity. To this end, I hierarchically clustered switching TFs (N = 277) and tissues (N = 46) by the types of switching major isoforms (Fig. 5.11; see section 3.2 for methodological details). Among tissue clusters (Fig. 5.11A, rows), the

testis and brain clearly separated, which agrees with the known uniqueness of their transcriptomes (Barbosa-Morais et al., 2012; Soumillon et al., 2013; Yeo et al., 2004). However, I also found that the non-brain tissues split into two unequal groups, and one of them, "Non-brain 1," separated earlier in the hierarchy of clusters than the brain and all other tissues. Also, the cerebellum, pituitary gland and prostate separated into the "Other" cluster (Fig. 5.11A, rows). As these two separations were hard to explain, I checked if they were caused by possible isoform quantification artefacts in GTEx leading to artificially similar transcriptional profiles of tissues within the "Non-brain 1" and "Other" clusters. For this, I did the principal component analysis (PCA) on the tissue samples that I used in my switch analysis, based on 2,000 top-expressed protein-coding isoforms of TFs and non-TFs (Fig. 5.12). I found that the pancreas and liver were inseparable by the first two principal components (PC1 and PC2), and also that the heart and skeletal muscle tissues were highly overlapping, which was expected. However, these two groups of tissues were separated by PC2 and hence their clustering together in the switch heatmap was not due to artificially similar transcriptional profiles. Likewise, the prostate, pituitary gland and cerebellum were largely separated by PC1, as expected. Of note, the brain subtissues, including the cerebellum, were separated from all non-brain tissues by PC1, and the cerebellum projected to the same area as all other brain subtissues, which is exactly as expected (Fig. 5.12). Hence, the grouping of the "Non-brain 1" and "Other" tissues by their switch patterns needs a biological interpretation, and I suggest an interpretation of the "Non-brain 1" cluster in the Discussion (see section 5.4.1).

The 7 TF clusters that I obtained group TFs by their switch patterns and tissue specificity (Fig. 5.11A, columns; Supplementary Table 4). Namely, cluster 6 ($N = 44$) contains extremely tissue-specific TFs, while cluster 3 ($N = 71$) contains TFs expressed mostly in tissue clusters "Non-brain 2" and "Other." TFs from both clusters switch two DBD+ major isoforms. The rest of the TF clusters consist of broadly expressed TFs and show clear switch patterns. In cluster 1 ($N = 6$), the broadest expressed major isoform is DBD-; it is expressed mainly in the brain and switched for a DBD+ major isoform in particular non-brain tissues. In cluster 2 ($N = 14$), the broadest expressed major isoform is also DBD-; it is expressed across all tissues and occasionally switched for a DBD+ major isoform. In cluster 4 ($N = 36$), the broadest expressed major isoform is DBD+; it is expressed across all tissues and switched for another DBD+ major isoform predominantly in the "Non-brain 2" tissues. In cluster 5 ($N = 92$), the switch pattern is similar but the broadest expressed DBD+ major isoform is switched for another DBD+ or DBD- major isoform mainly in the brain. Finally, in cluster 7 ($N = 14$), the broadest expressed

major isoform is DBD+; it is mainly expressed in the brain and switched for another DBD+ major isoform in particular non-brain tissues.

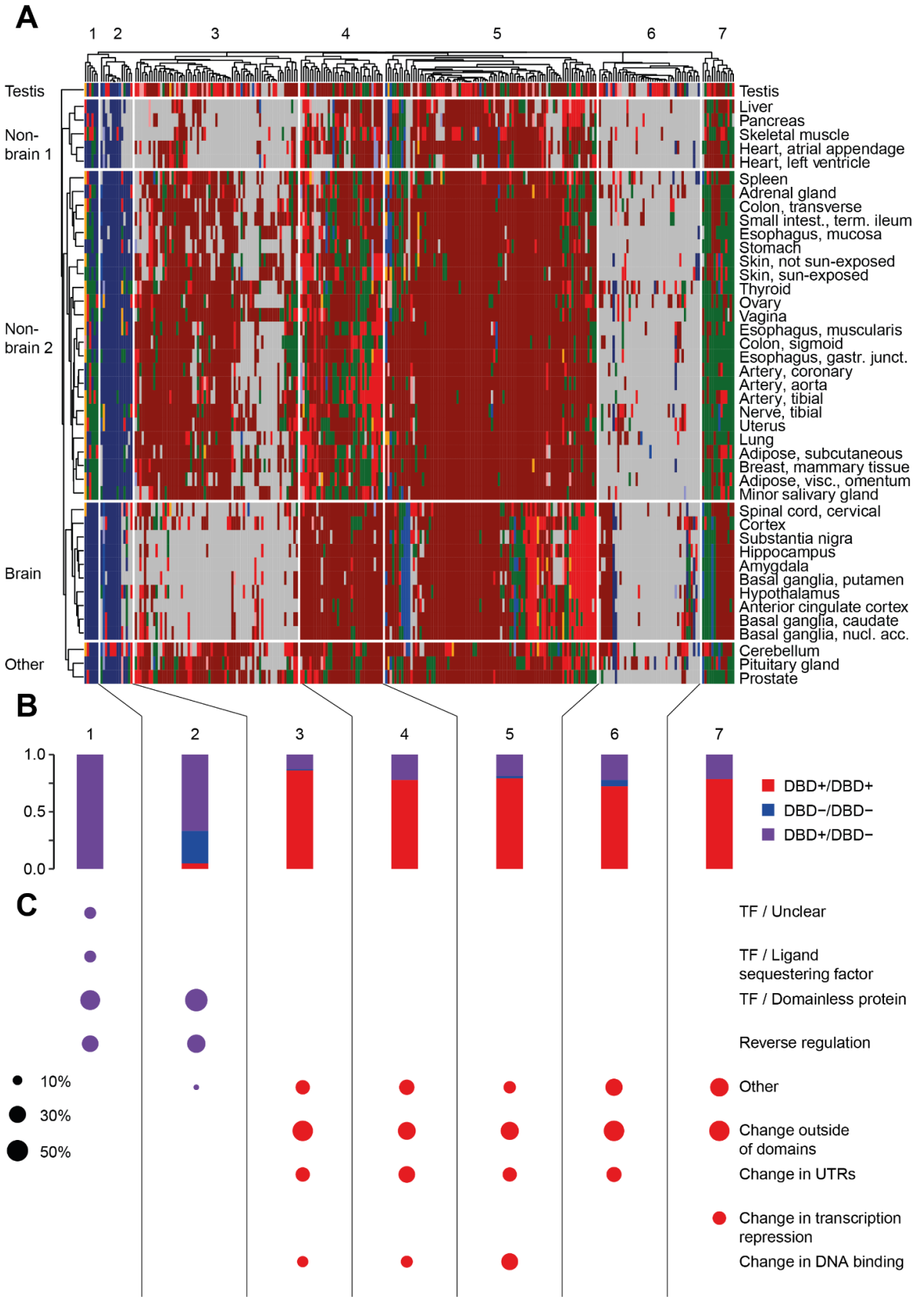


Fig. 5.11. Expression patterns of switching major isoforms across adult human tissues. **(A)** Switches of major TF isoforms. TFs (columns) and tissues (rows) are clustered hierarchically as vectors of TF scores to visualise switch patterns (see section 5.2.2 for methodological details and Fig. 5.4 for the colour scheme). **(B)** Proportions of switch types in each TF cluster. **(C)** Proportions of functional categories within the most frequent type of switches in each TF cluster.

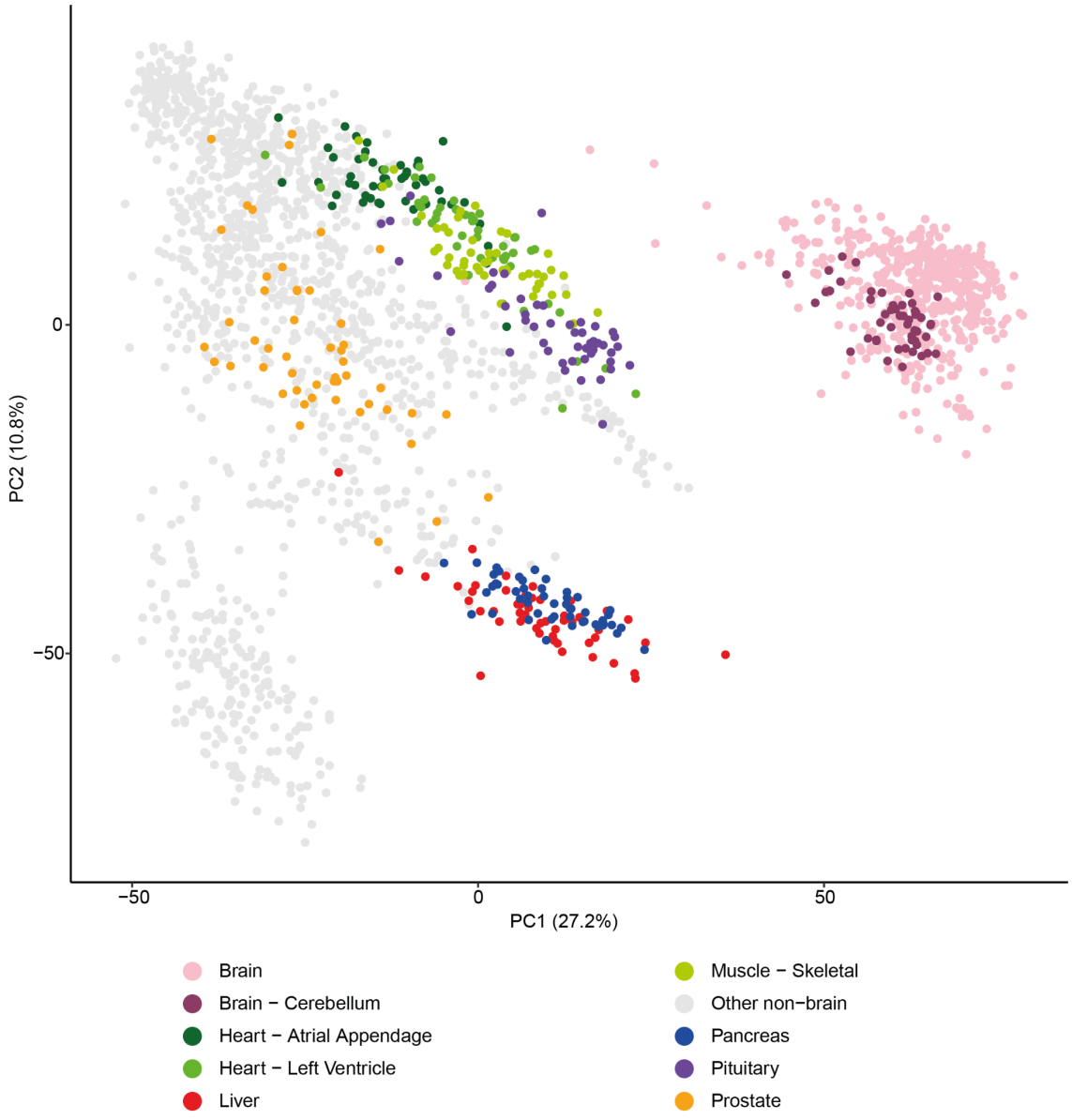


Fig. 5.12. PCA of human adult tissue samples. The PCA is based on the 2,000 top-expressed protein-coding isoforms of TFs and non-TFs. Percentages next to the principal component labels denote the proportion of the variation explained by the components. Samples are coloured by tissue. See section 5.2.2 for methodological details.

Because TF clusters delineate different switch patterns, I studied each cluster further to understand putative functions of switches in different clusters. Specifically, I calculated the proportions of the +/+, +/- and -/- switches (Fig. 5.11B), and for the most frequent switch type in each cluster I also calculated the proportions of functional switch categories (Fig. 5.11C). I found that the +/- switches from clusters 1 and 2 have the same most frequent categories, "TF / Domainless protein" and "Reverse regulation," similarly to the +/- switches in general (Fig. 5.6B). Additionally, clusters 3 to 7, where the +/+ switches prevail, demonstrate very similar distributions of switches across functional categories. However, interestingly, clusters 6 and 7 do not have switches that change a DBD, while cluster 7 is the only one comprising the switch category "Change in transcription repression."

After the overall characterisation of the TF isoform switches, I asked if among them I could observe any switches between experimentally characterised isoforms of the 24 TFs shown in Fig. 3.10-3.11. For 22 out of the 24 TFs, I cannot expect to see switches of their major isoforms in my data, because they switch their experimentally characterised isoforms in development (rather than between adult tissues that I study here), or between cancer cells or tissues that are not in my dataset, or between different conditions (for example, upon activation in the case of immune cells or before and after meiosis in the case of germ cells); also, some of these TFs were not known to switch major isoforms at all. However, for 8 of these TFs (CREM, PAX8, ESR1, NFATC1, NR3C1, TCF3, PGR, HNF4A) I did observe switches in my data, although, those switches involved isoforms other than those described in Fig. 3.10-3.11. On the other hand, for the remaining 2 TFs (HNF1A, TCF4) I expected to see switches in my data but did not observe them. Possible reasons for this include the absence of switching isoforms from the Ensembl annotation, assignment of low TSLs to these isoforms (so I could not select them for my analysis) or misquantification of these isoforms in GTEx.

Taking into account the fact that alternative isoforms are described to some extent in only about a hundred of human TFs (Tables 1-3) and that I found 277 TFs that switch their major isoforms, my results may include a considerable number of unstudied switches. Consequently, to guide future experimental investigations of TF isoform switches, I considered several groups of TFs with pronounced brain-non-brain switch patterns, as well as some individual TFs, in greater detail (Fig. 5.13-5.14; Table 5).

Table 5. Groups of TFs with pronounced switches of major isoforms between the brain and non-brain tissues.

Group	TF list	Switch pattern
I	FOXJ3, MAZ, NFIB, NR2F2, ZNF197, ZNF226	DBD- in the brain and some other tissues / DBD+ in particular non-brain tissues
II	ZNF12, ZNF195, ZNF274, ZNF548, ZNF721, ZNF746, ZNF76	One DBD+ mainly in the brain / Another DBD+ in particular non-brain tissues
III	IRF3, RARA, THAP4, ZNF470, ZNF571	DBD- in the brain / DBD+ broadly expressed outside the brain
IV	ADNP, CAMTA2, CEBPG, JAZF1, MEF2A, MEIS1, SREBF1, TFEB, ZBTB45, ZC3H8, ZNF202, ZNF384, ZNF415, ZNF529, ZNF592, ZNF605	One DBD+ in the brain / Another DBD+ broadly expressed outside the brain.

First of all, TFs from group I (FOXJ3, MAZ, NFIB, NR2F2, ZNF197, ZNF226) which equals TF cluster 1 (Fig. 5.13) switch a DBD- major isoform expressed in the brain and some other tissues for a DBD+ major isoform expressed in particular non-brain tissues. To date, none of these switches has been characterised; however, the pronounced brain-non-brain pattern of these switches suggests their potential functional relevance. For example, FOXJ3 is a TF in arteries, the adipose tissue, skeletal muscles, parts of the digestive system, testis and the adrenal gland, but becomes a domainless protein, hence losing its TF function, in the brain and some other non-brain tissues (Fig. 5.14A).

Secondly, TFs from group II (ZNF12, ZNF195, ZNF274, ZNF548, ZNF721, ZNF746, ZNF76), which consists of C2H2 zinc finger factors and is a part of cluster 7 (Fig. 5.13), switch a DBD+ major isoform expressed predominantly in the brain for another DBD+ major isoform expressed in particular non-brain tissues. Again, to the best of my knowledge, none of these switches has been characterised, although they might be functionally important. For instance, ZNF12 changes the sequence of its transcription repression domain (KRAB) between its two major isoforms and hence, potentially, changes the mode of transcription repression between the brain and some of the non-brain tissues (Fig. 5.14B).

Next, TFs in group III (IRF3, RARA, THAP4, ZNF470, ZNF571) which is a part of cluster 5 switch a DBD- major isoform expressed in the brain for a DBD+ major isoform broadly expressed outside the brain (Fig. 5.13). Among these switches, only the one by IRF3 has been characterised to some extent (Karpova et al., 2001). Specifically, this switch could differentially control IFN-beta activation between the brain and non-brain tissues (Fig. 5.14C). However, several other TFs in this group switch their major isoforms in exactly the same way and their switches have still not been characterised, while my analysis suggests their functional importance. For example, ZNF571 (Fig. 5.14D) switches a DBD+ major isoform (a likely repressor) for a DBD- major isoform (predicted in my analysis to be an activator).

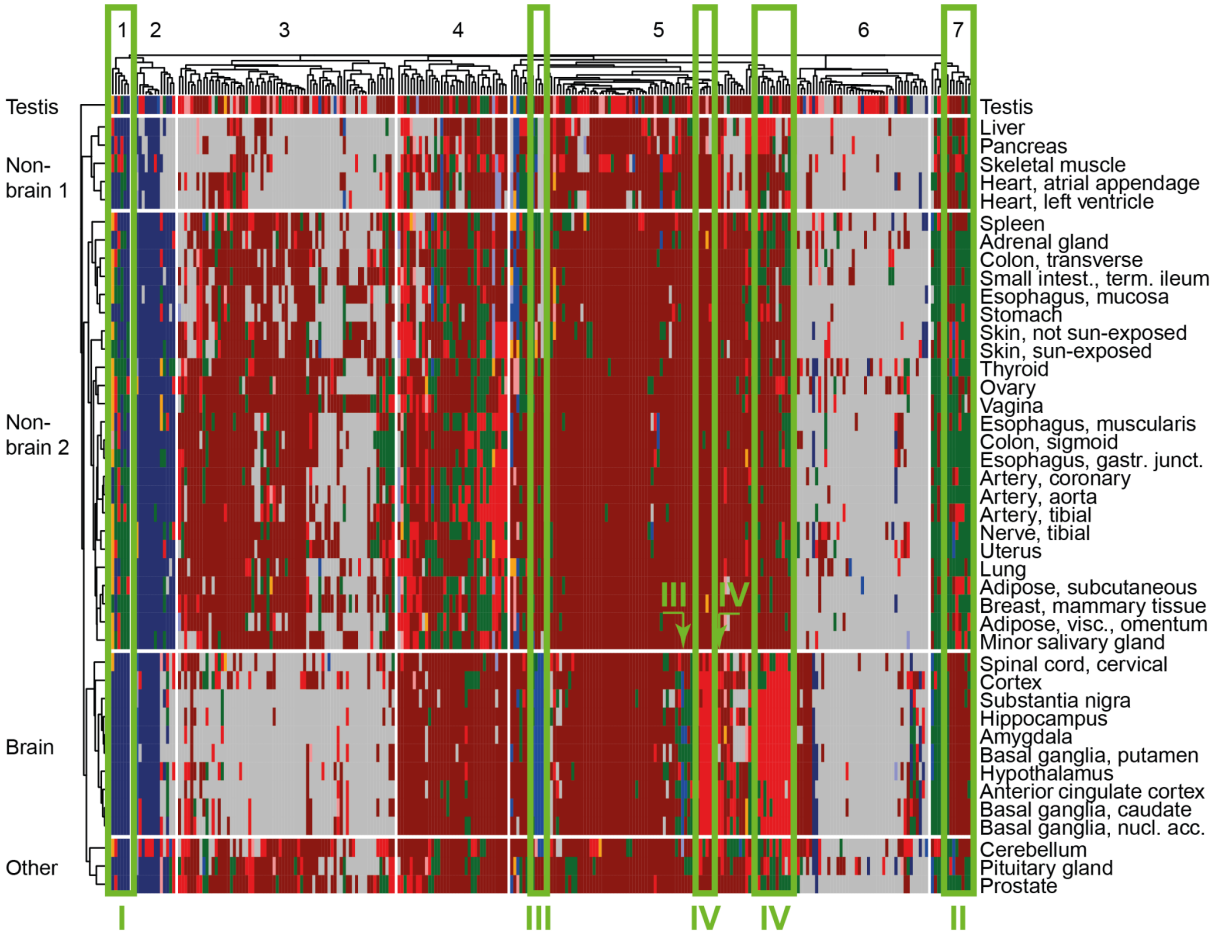


Fig. 5.13. Groups of TFs with pronounced brain-non-brain switch patterns. Group I equals cluster 1, group II comprises the major part of cluster 7, groups III and IV are parts of cluster 5. Group III consists of a continuous block and a single TF with the same pattern, while group IV consists of two blocks and a single TF. Single TFs are marked with labelled arrows.

Finally, TFs from group IV (ADNP, CAMTA2, CEBPG, JAZF1, MEF2A, MEIS1, SREBF1, TFEB, ZBTB45, ZC3H8, ZNF202, ZNF384, ZNF415, ZNF529, ZNF592, ZNF605), which is also a part of cluster 5, switch a brain-specific DBD+ major isoform for a broadly expressed non-brain DBD+ major isoform (Fig. 5.13). Similarly to groups I and II, none of these switches have been functionally characterised. However, my analysis and the pronounced brain-non-brain pattern of these switches suggest their potential functional relevance. For instance, a switch in TFEB could change the binding of a partner protein (Fig. 5.14E). Overall, alternative isoforms of only 15 out of the 34 TFs that I considered in the examples above have been studied to some extent (CEBPG, FOXJ3, IRF3, MEF2A, MEIS1, NFIB, NR2F2, RARA, SREBF1, TFEB, ZNF197, ZNF202, ZNF415, ZNF746, ZNF76; see Table 3), and for only 1 out of these 34 TFs (IRF3) a functional role has been suggested for its switch that I observe in my analysis.

Apart from these highly visible groups of switching TFs, I also considered single TFs with particularly interesting switches. First of all, PRDM2 from cluster 7 switches a DBD- major isoform expressed in the brain for a DBD+ major isoform expressed outside the brain, and I predicted this unstudied switch to differentially affect H3K9 trimethylation between the brain and non-brain tissues (Fig. 5.14F). Other unstudied, but functionally promising, switches include:

1. A +/+ switch between the *adult* vagina and testis of a regulator of sex determination WT1 (Hammes et al., 2001; Shimamura et al., 1997). This switch could regulate the binding of a partner protein (Fig. 5.14G).
2. A +/+ switch of MEF2C, a well-known regulator of muscle differentiation (Potthoff et al., 2007; Ridgeway et al., 2000; Zetser et al., 1999), that occurs between the *adult* skeletal muscle and other adult tissues and may entail differential binding of partner proteins (Fig. 5.14H).
3. A +/+ switch by TBX3 between (a) spleen and liver and (b) other adult tissues. It may facilitate differential DNA binding affinity of TBX3 (Fig. 5.14I).
4. A -/- switch by an understudied TF ZNF655 that occurs, in part, between two groups of the brain subtissues and involves two domainless major isoforms (Fig. 5.14J).

Overall, I found that switching TFs cluster by their switch patterns: by tissues in which the switches occur and by the types of switching major isoforms (DBD+ or DBD-). Importantly, the clusters showed that a number of major isoform switches occur between

particular types of tissues (especially, the brain and non-brain tissues), while my further analysis demonstrated that many major isoform switches could be functional. Therefore, my *in silico* finding of the cross-tissue switches of the major TF isoforms warrants experimental characterisation.

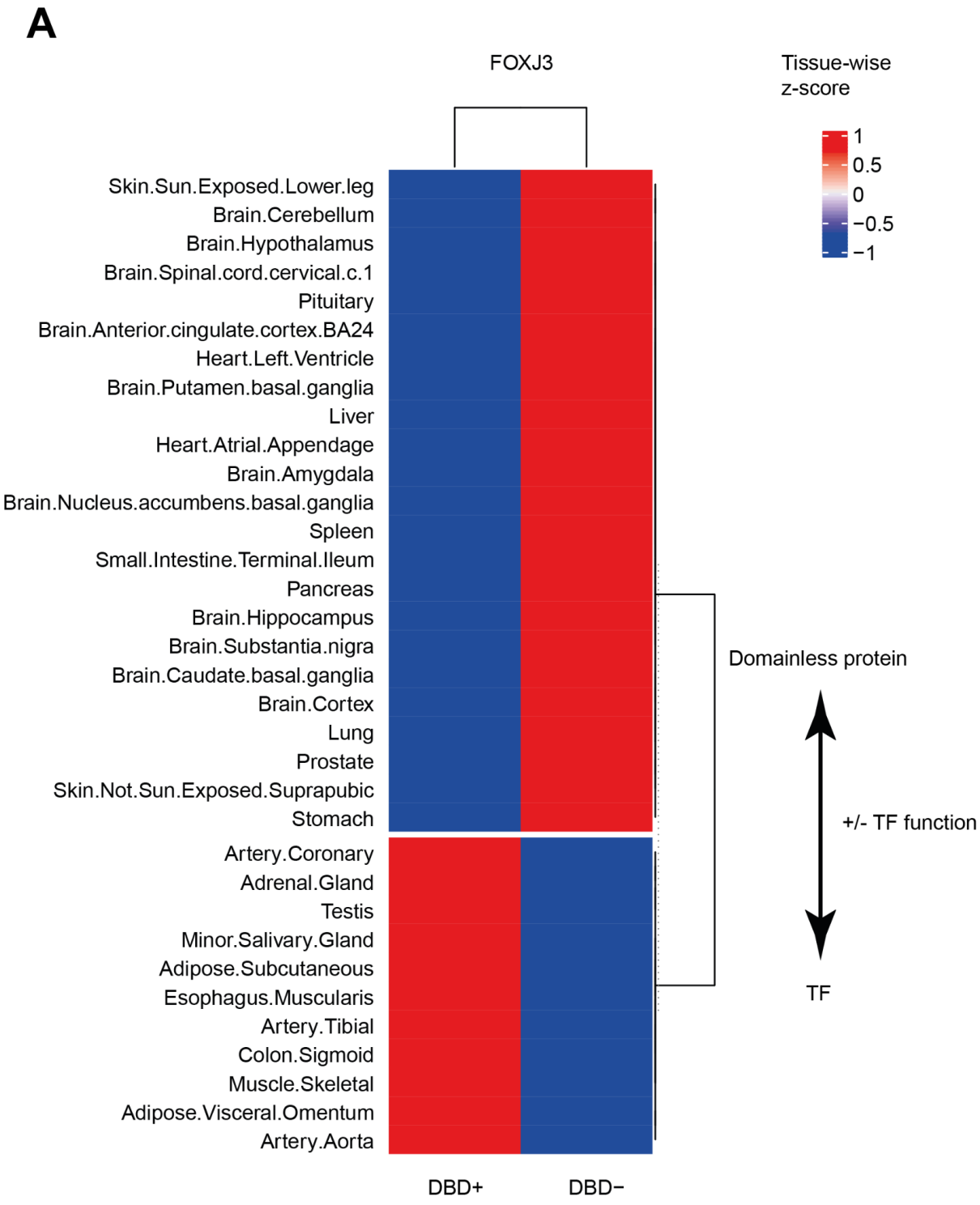


Fig. 5.14. Major isoform switches in particular TFs. (A) Major isoforms of FOXJ3, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

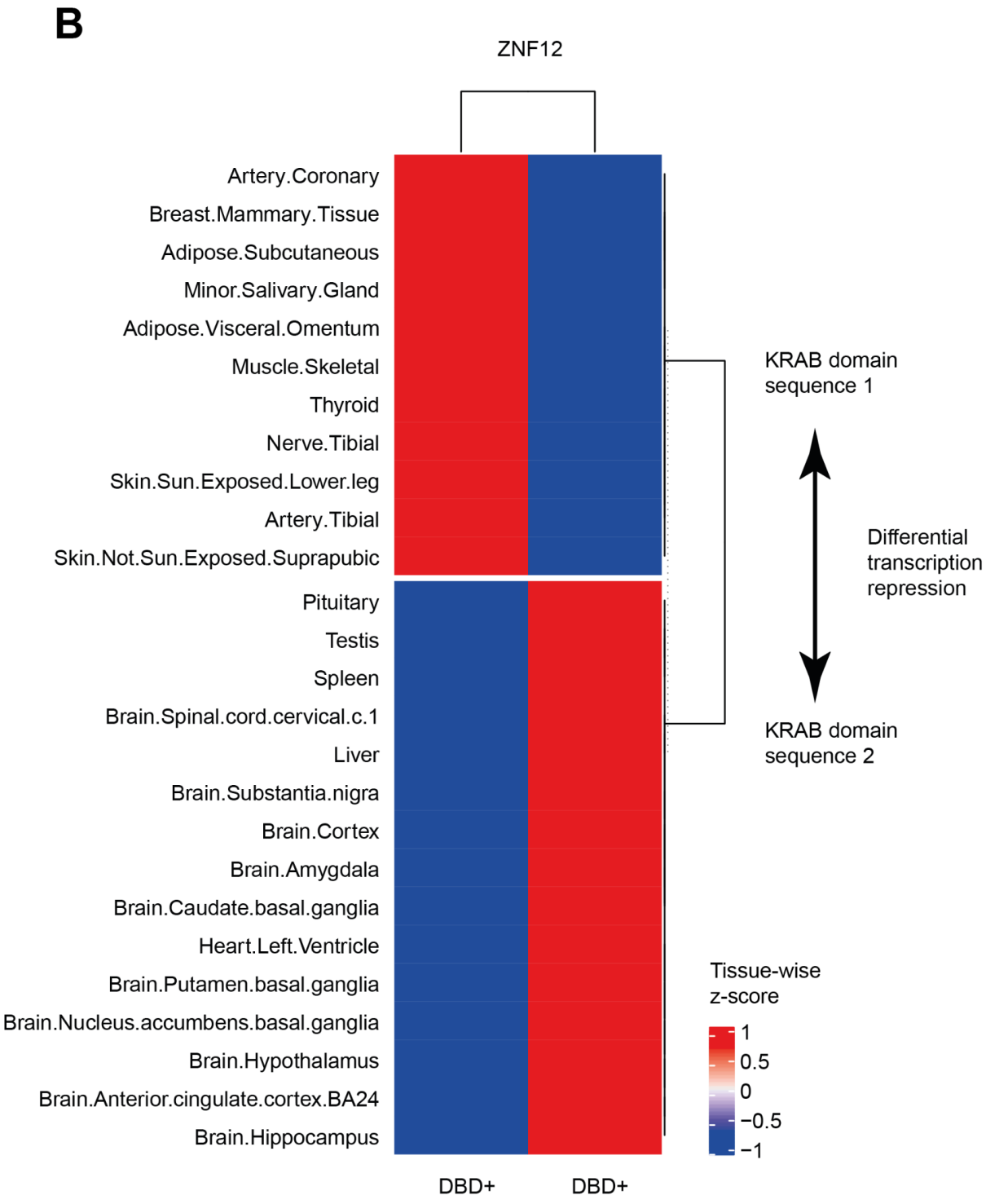


Fig. 5.14. Major isoform switches in particular TFs. (B) Major isoforms of ZNF12, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

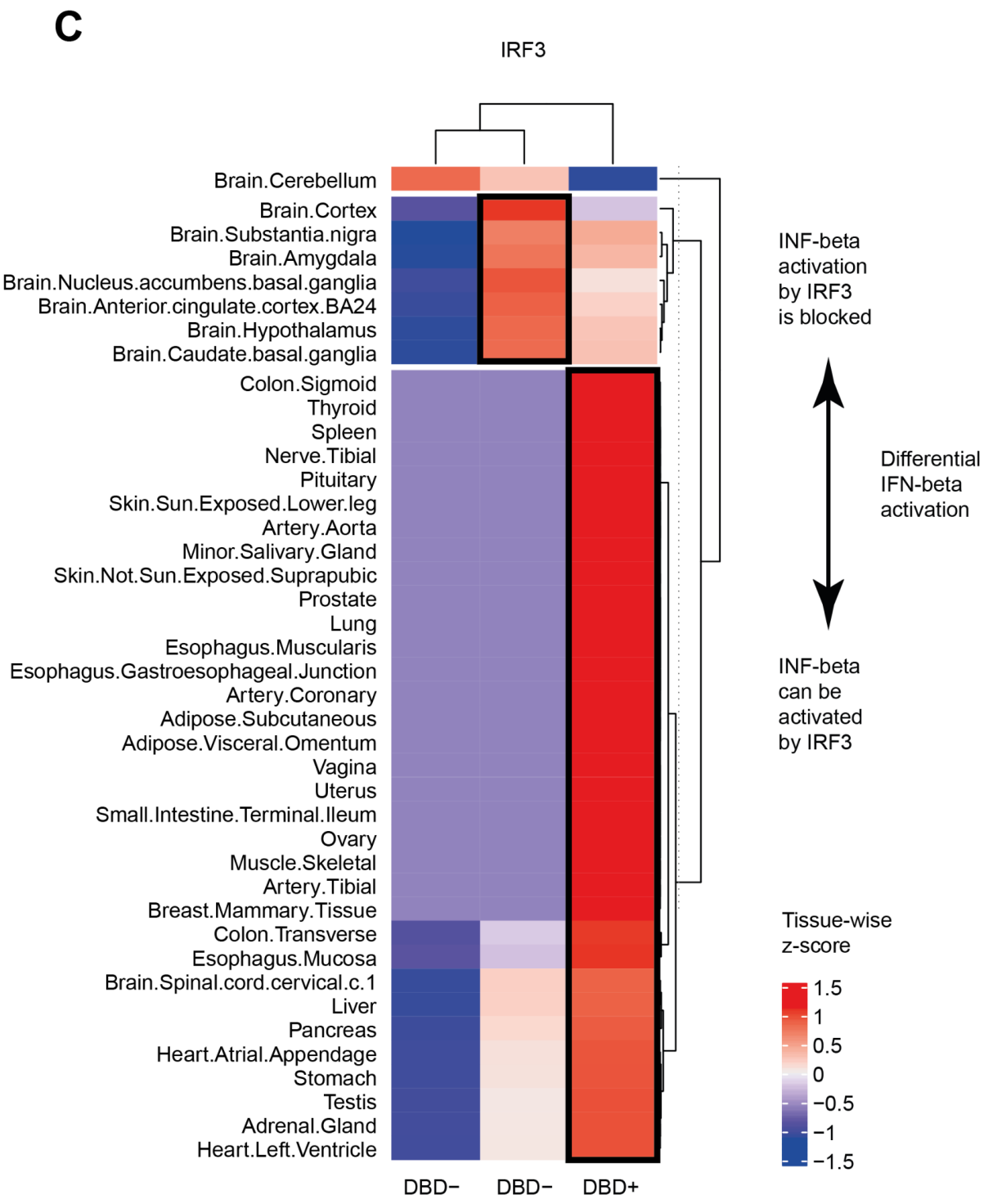


Fig. 5.14. Major isoform switches in particular TFs. (C) Major isoforms of IRF3, a switch of two of them (framed) and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

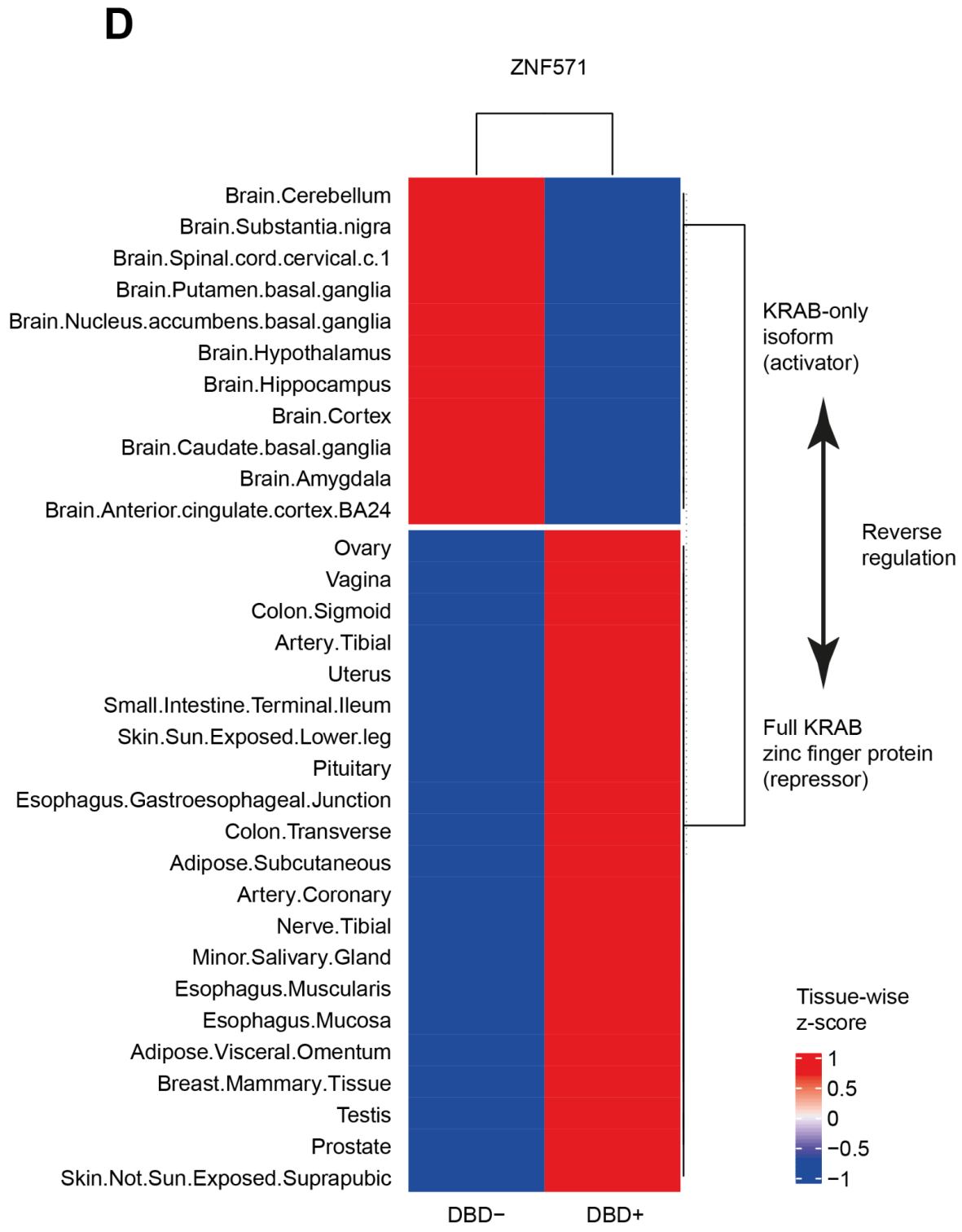


Fig. 5.14. Major isoform switches in particular TFs. (D) Major isoforms of ZNF571, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

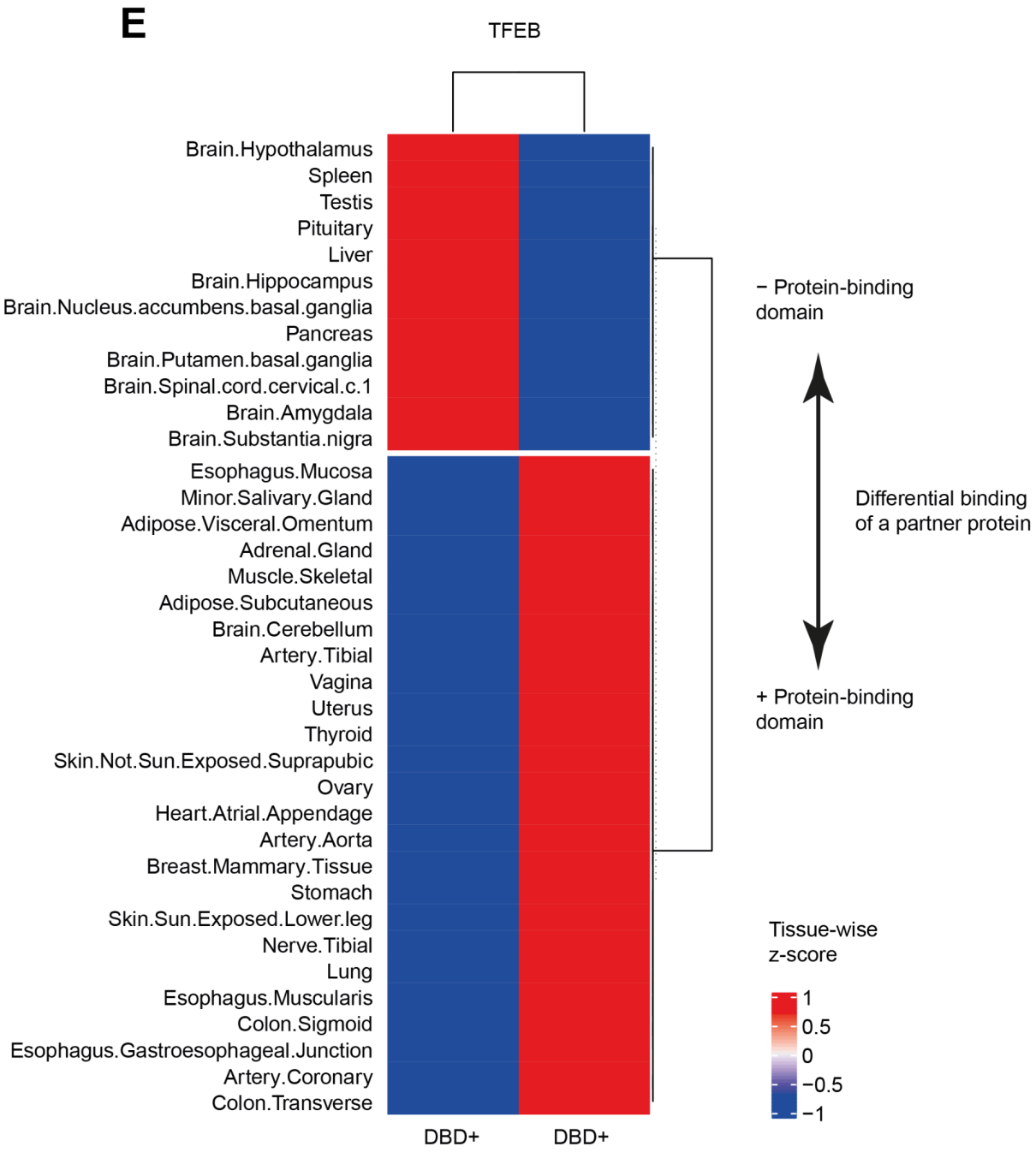


Fig. 5.14. Major isoform switches in particular TFs. (E) Major isoforms of TFEB, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

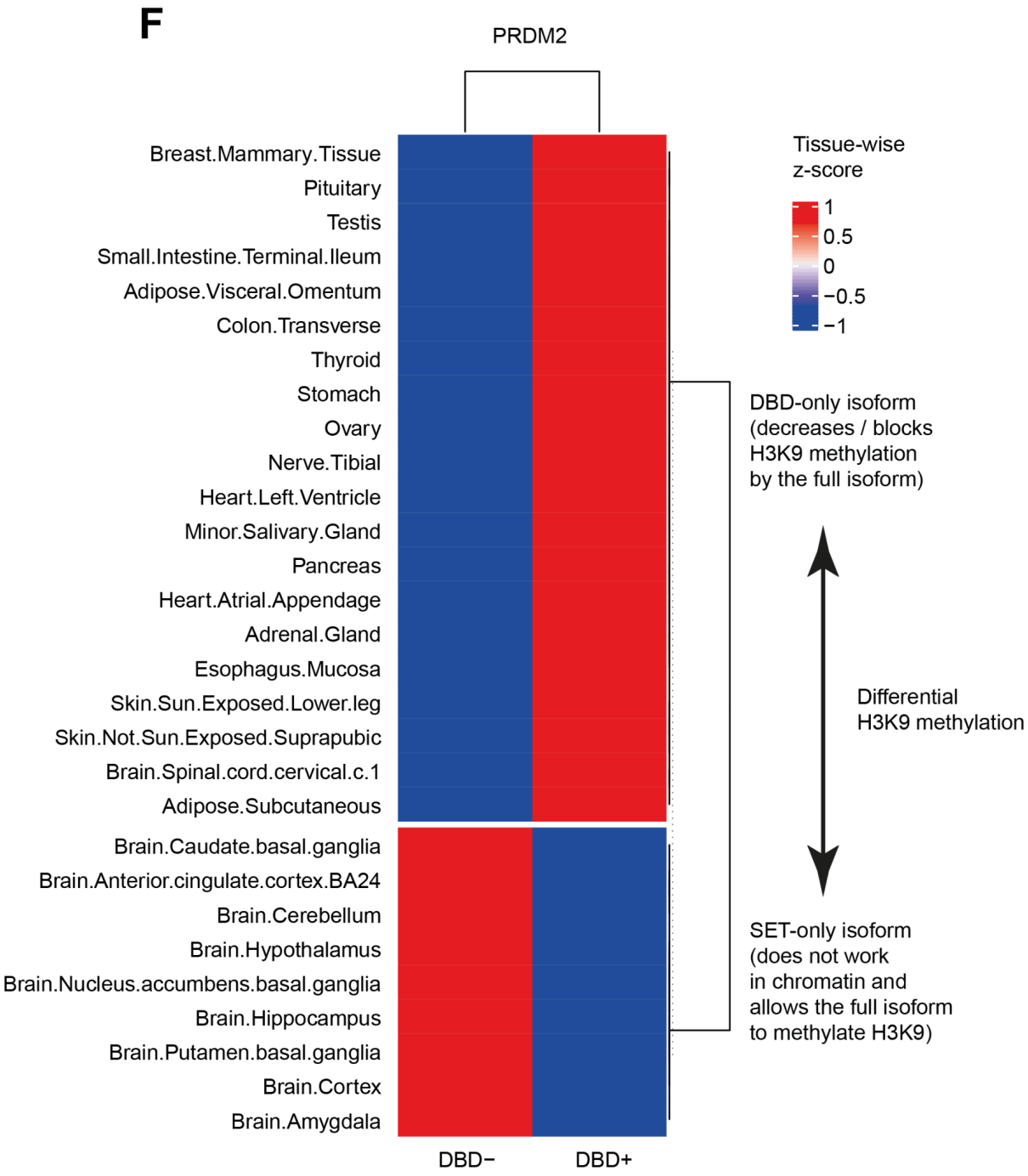


Fig. 5.14. Major isoform switches in particular TFs. (F) Major isoforms of PRDM2, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

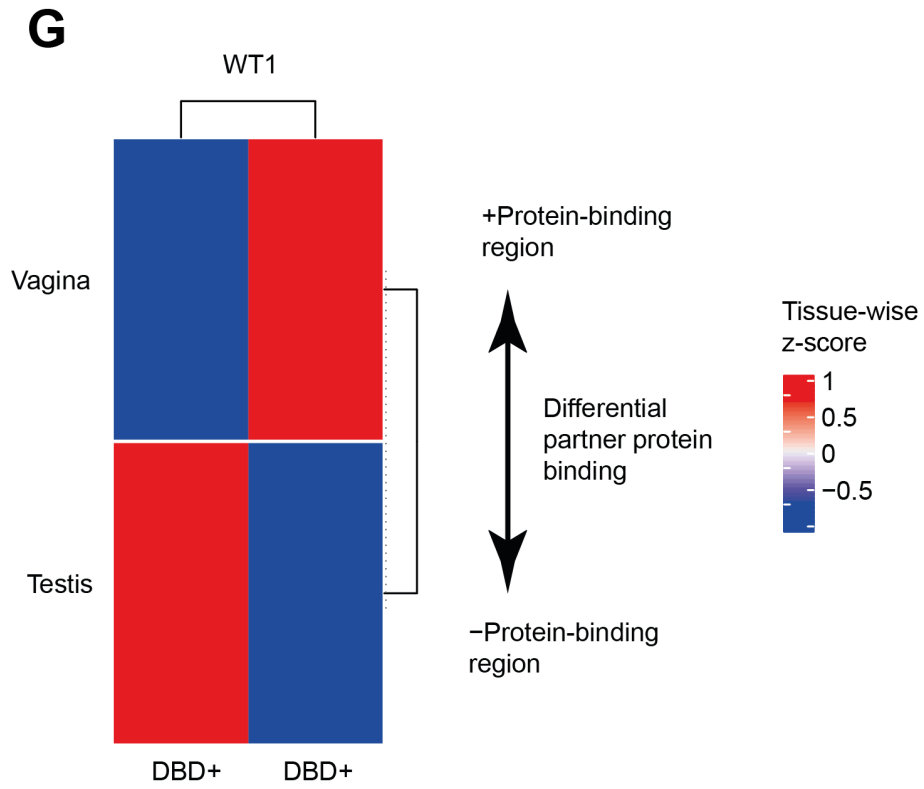


Fig. 5.14. Major isoform switches in particular TFs. (G) Major isoforms of WT1, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

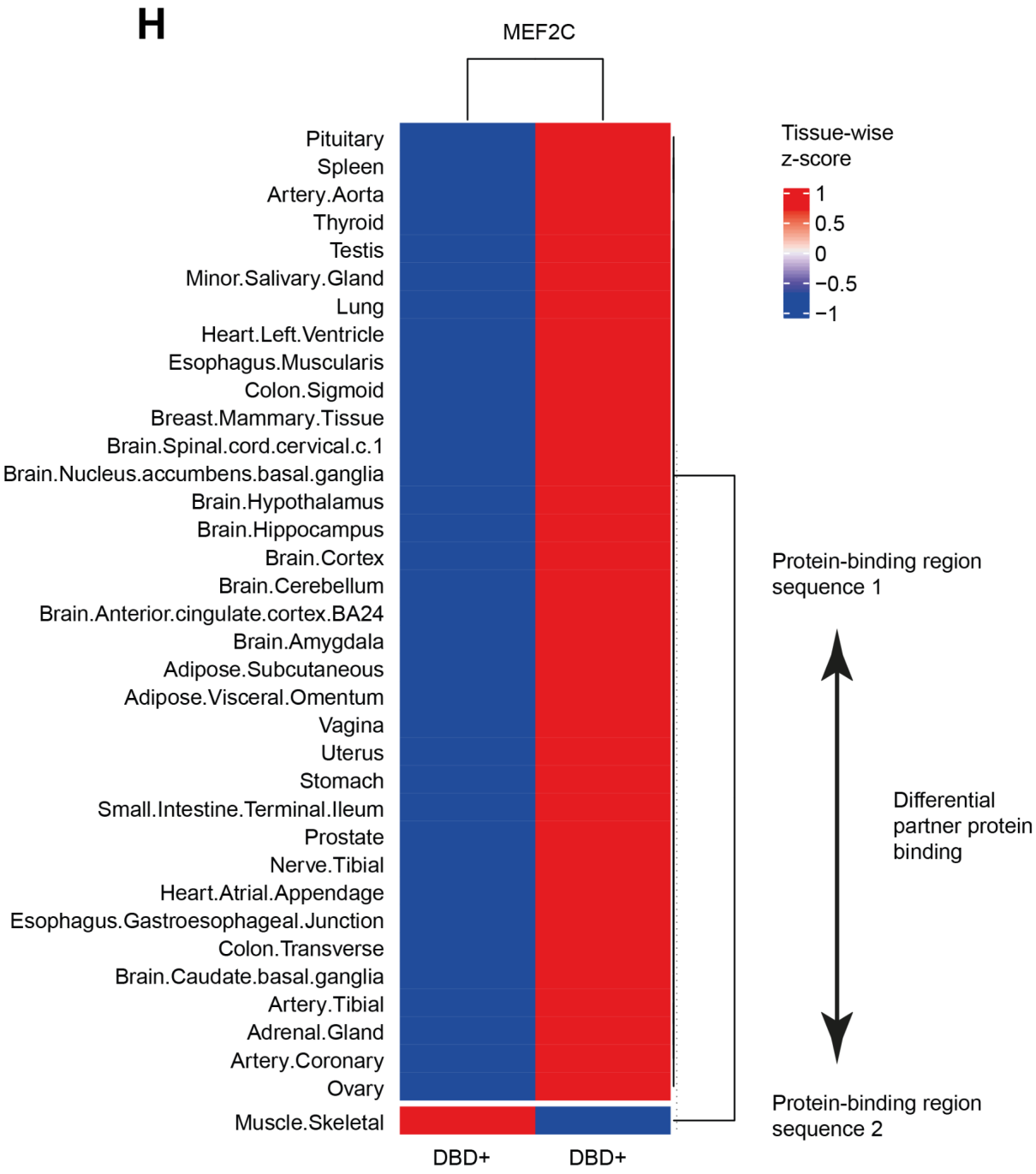


Fig. 5.14. Major isoform switches in particular TFs. (H) Major isoforms of MEF2C, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

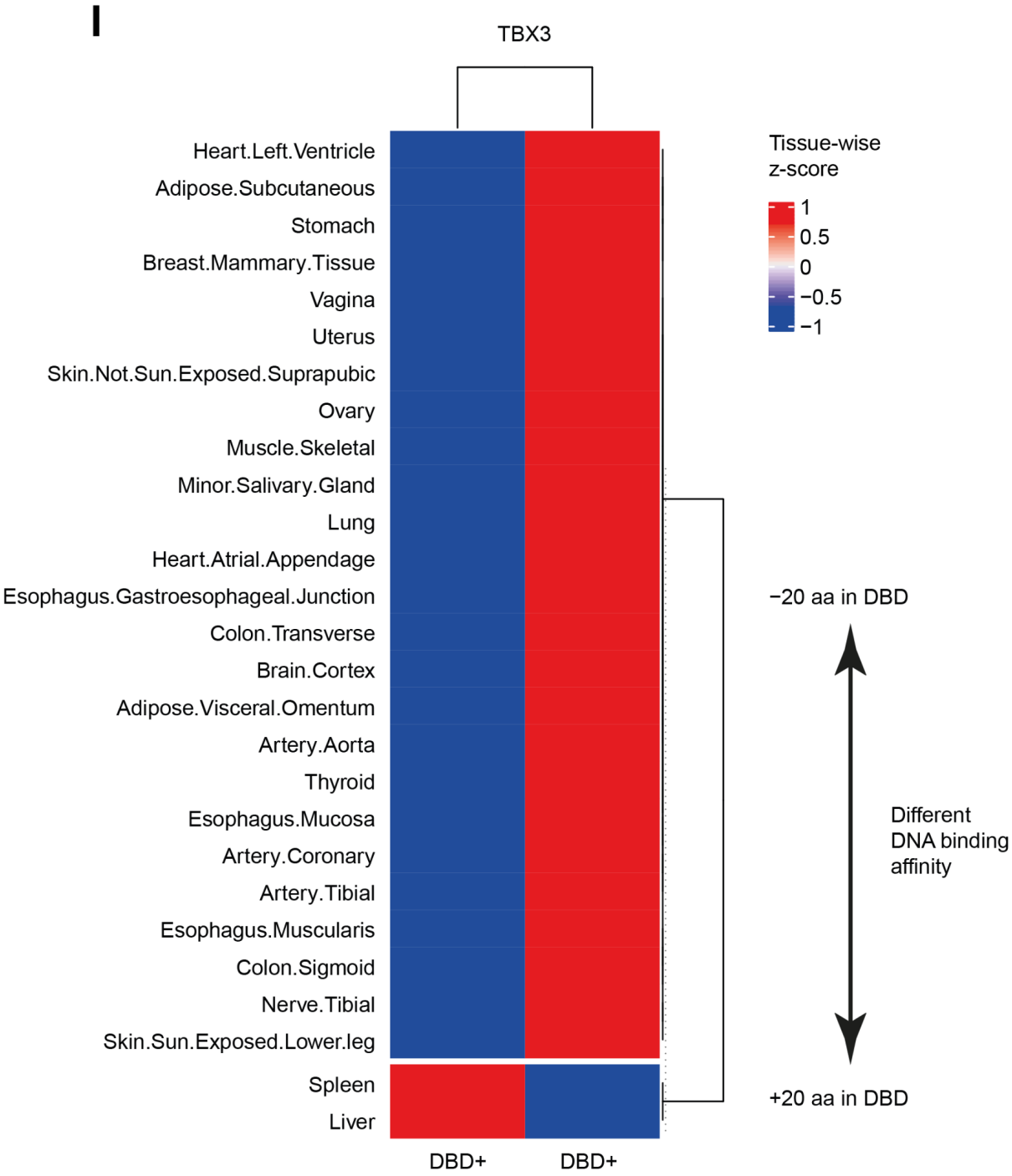


Fig. 5.14. Major isoform switches in particular TFs. (I) Major isoforms of TBX3, their switch and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

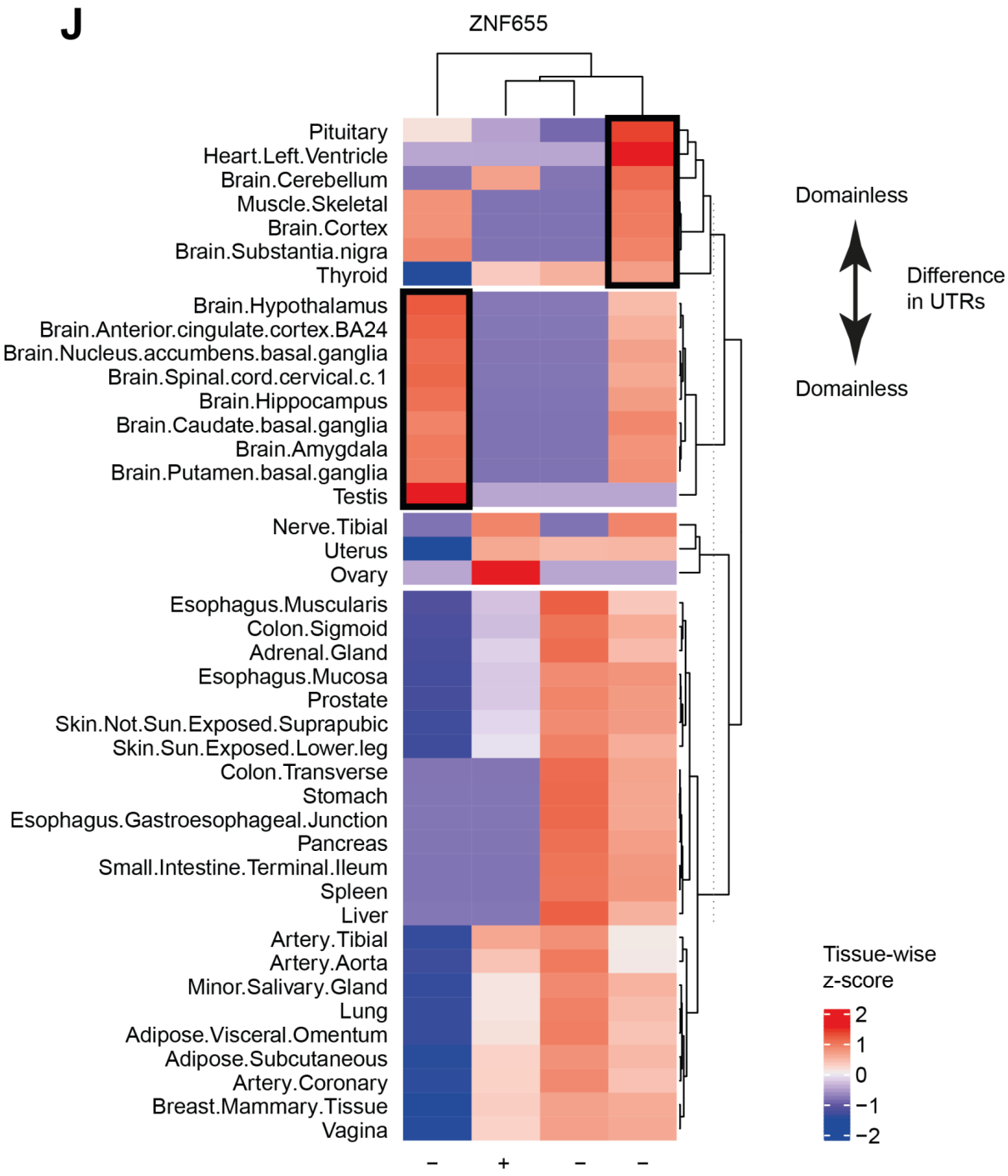


Fig. 5.14. Major isoform switches in particular TFs. (J) Major isoforms of ZNF655, a switch of two of them (framed) and its predicted functional consequence. The heatmap shows tissue-wise z-scores of isoform expression levels.

5.4. Discussion

5.4.1. Summary and interpretation of the results

In this chapter, I defined a switch between two major TF isoforms and analysed switches between adult human tissues. I found that 22% of human TFs produce several major isoforms depending on the tissue, and the majority of such TFs (85%) switch at least one pair of major isoforms between tissues. The fact that the most part of TFs that produce more than one major isoform switch these isoforms could be partially explained in terms of probability. Indeed, each major isoform is transcribed in at least some of more than 40 adult human tissues that I used in my analysis, and the more tissues I study, the higher the chance is that the difference between the transcription levels of major isoforms expressed in any two of these tissues becomes larger than 10%, which is the criterion for calling a switch. However, if the difference in the transcription levels is substantial, then any differences in the putative protein sequences of the switching isoforms become interesting as they may be functionally important. Hence, I predicted possible functional consequences of the 400 switches that I found.

Interestingly, among the +/+ switches that comprise the majority of all switches, I found that 61% do not affect domains. Hence, these switches may affect UTRs or protein sequences between domains. In the former case, a switch may influence the mRNA stability, localisation or translational efficiency (Byeon et al., 2021; Jia et al., 2020; Mayr, 2017), and this fact suggests an interesting project to study possible differences in UTRs between switching major isoforms or alternative TF isoforms in general. In the latter case, when the protein sequence is affected, a switch can alter intrinsically disordered regions (IDRs) between domains. Hence, as IDRs are also functional units of the protein sequence, the alternative splicing of IDRs in human TFs needs to be studied alongside the alternative splicing of domains. For instance, previous studies showed that alternative splicing of IDRs likely creates tissue-specific networks of protein-protein interactions (Buljan et al., 2012).

One of the top three most frequent predicted functional effects of the +/+ switches was a change in the TF-DNA binding induced by a change in the sequence or the number of DBDs. I studied differences between the DBD sequences of switching major isoforms of four TFs – ERF, RARB, PGR and NFATC1 – and found that ERF1 retained the residues that confer DNA-binding specificity, while the other three TFs completely (or nearly completely) lost these residues in the switches. Consequently, depending on the folding of the truncated protein,

ERF1 could bind DNA with a specificity similar to the full isoform, while the truncated forms of RARB, PGR and NFATC1, if still able to bind DNA, would do so with specificities different from the ones of their full isoforms.

After predicting the functional consequences of switches of the major TF isoforms, I tested the overrepresentation of TF families and GO terms in switching TFs. I found that there was no such overrepresentation and that the vast majority (72%) of TF families expressed in adult human tissues had at least one switching TF. This result suggests that the switching of major isoforms is not specific to any structural or functional group of TFs. In contrast, I found an overrepresentation of particular TF families among TFs producing switches of particular functional consequences, which suggests that TFs from different families "employ" switches of their major isoforms to "achieve" certain characteristic effects.

Finally, I clustered TFs and adult human tissues by their characteristic switch patterns. The majority of the brain tissues clustered together and separately from the other tissues suggesting that the brain expresses a unique set of major TF isoforms. Interestingly, several non-brain tissues (the liver, pancreas, skeletal muscles and two heart subtissues) also clustered together, while an additional analysis suggested that this clustering is not an artefact. Although I do not have a definite biological explanation of this clustering, I can suggest a physiological phenomenon that is common to all of these tissues: a key role of the glucose metabolism. Indeed, the liver controls the level of glucose in the bloodstream through the production and utilisation of glycogen, as well as through the glycolysis and gluconeogenesis (Dashty, 2013; Han et al., 2016), while the pancreas secretes insulin and glucagon to control these processes (Titchenell et al., 2017). In turn, the contraction of skeletal muscles primarily depends on the internally stored glycogen and also on the glucose supply from the bloodstream, especially in endurance exercise when the glucose uptake from the blood can account for up to 40% of the oxidative metabolism in this tissue (Richter and Hargreaves, 2013). Skeletal muscles store about two thirds of the whole amount of glycogen in the organism and during intense physical activity produce lactate that feeds back to the liver and regulates the *de novo* production of additional glucose (Dashty, 2013). Finally, the heart demonstrates one of the highest glucose uptakes among all organs in normal conditions (Zhao et al., 2021) and also depends on a tight control of its glucose metabolism to keep the normal contractile function (Dhar-Chowdhury et al., 2007).

However, this cluster lacks two other tissues that are key for the glucose metabolism and are present in my analysis: the brain and the adipose tissue. Indeed, glucose is the primary source of energy for the brain, and this organ has a share of 20% of the whole glucose

consumption in the body (Zhang et al., 2021). In turn, the adipose tissue regulates the amount of glucose in the body via the lipogenesis that converts the excess of glucose into triglycerides and the lipolysis that produces glycerol, one of the main non-carbohydrate precursors of glucose in the gluconeogenesis (Han et al., 2016). Overall, the physiological relevance of the non-brain tissue clusters requires further investigation.

TFs clustered by their tissue specificity and switch patterns. Five clusters out of seven consisted of broadly expressed TFs that demonstrated clear patterns of the major isoform switching between the tissues. I outlined several groups of TFs, with 34 factors in total, that demonstrated the particularly interesting switches of their major isoforms between the brain and non-brain tissues, as well as additional 5 TFs that demonstrated interesting switches between other tissues. For example, based on the disparities between the switching major isoforms, these switches could turn a TF into a non-TF protein (FOXJ3) or differentially affect the transcription regulation mode of a TF (ZNF12, ZNF571), change the binding of a partner protein (TFEB, WT1, MEF2C), H3K9 trimethylation (PRDM2), or the DNA binding affinity (TBX3). As 38 of these 39 switches have not been studied, I suggested their experimental characterisation. See section 5.4.3 for a brief discussion of possible experiments.

5.4.2. Limitations of the analysis

First of all, in my analysis of the major isoform switches, I did not consider non-coding TF isoforms, because I would not be able to predict their functions (as opposed to coding isoforms whose functions mainly follow from their domains). Nevertheless, at least some non-coding TF isoforms may be functional, as, for instance, non-coding RNAs produced by FOXO3 and CEBPA have been found to be functionally important (Di Ruscio et al., 2013; Du et al., 2016, 2017).

Secondly, I considered only one protein per mRNA isoform, as Ensembl v99 presents a one-to-one correspondence between mRNAs and proteins. However, a TF mRNA can be alternatively translated and give rise to functionally different protein isoforms (Bégay et al., 2018; Kennett et al., 1997; Lu and Cidlowski, 2005; Sapetschnig et al., 2004), which I did not account for in my analysis.

Also, any disparate functions of alternative protein isoforms may be regulated by the differential presence of residues that can be post-translationally modified. In my analysis, I did not account for post-translational modifications (PTMs) of alternative TF isoforms, primarily because of the lack of the genome-wide isoform-specific publicly available PTM data. One potential method to overcome this limitation would be to use the annotations and interpretations

of PTMs from the dbPTM database (https://awi.cuhk.edu.cn/dbPTM/) (Li et al., 2022) and propagate the PTM annotations from the UniProt amino-acid sequences used in dbPTM to individual TF isoforms annotated in Ensembl. However, in this approach I would need to assume that the presence or absence of non-modified residues does not affect the modification of the constantly present residues. This assumption will be false in the case of a change in the folding of a protein which hides the otherwise modified residues from the modifying enzymes or the gain/loss of a binding site for a modifying enzyme. For instance, Pim kinases and also a complex of CDK2 and cyclin A3 bind to residues located *near* those that they phosphorylate (Brown et al., 1999; Bullock et al., 2005). Consequently, the interpretation of PTMs *per protein isoform* may still be problematic.

In addition, the same protein isoform may be differentially post-translationally modified in different physiological conditions and hence exhibit different properties. Indeed, (Ellis et al., 2006) showed *in vitro* that the transactivation potential of short SP3 isoforms depends on their sumoylation status. However, I could not account for this phenomenon because in order to do so I would need data about PTMs per TF isoform per tissue, which is unrealistic.

Importantly, the limitations listed above are also relevant to my predictions of the functions of DBD- isoforms (section 4.3.3).

Furthermore, I defined major isoform switches based on a cutoff for the difference in isoform expression, rather than via testing the statistical significance of this difference. To refine this analysis in future studies, one could use the IsoformSwitchAnalyzeR R package (Vitting-Seerup and Sandelin, 2017, 2019) to find pairs of major isoforms with large and statistically significant opposite changes in expression. Also, the statistical assessment of isoform switches would help to expand the analysis from major isoforms to all expressed isoforms.

Moreover, for the analysis of differences between DBD sequences of switching major isoforms I selected only the switches between isoforms with a single DBD, so that I could easily interpret the changes occurring in the only DBD sequence. However, such switches comprise only 43% of all DBD+ to DBD+ (+/+) switches, and the major isoforms in the remaining 57% of +/+ switches (that occur between a single-DBD and a multi-DBD isoforms or between two multi-DBD isoforms) may differ in the number of their DBDs, as well as in the DBD sequences. These differences could be included into the analysis by taking into account the exonic structures of TF genes, as otherwise it would not be possible to detect which DBDs are gained/lost in a switch, or which DBDs, out of several present, changed their sequences.

However, even taking the exonic structures of TF genes into account, one would not be able to interpret the observed changes if they are complex; for example, a simultaneous gain/loss of DNA backbone-binding residues in one DBD and base-binding residues in another DBD.

Additionally, the small number of TFs that I was able to analyse in terms of the differences in the DBD sequences between switching major isoforms is explained by the fact that DBDs differed between switching major isoforms in only 20% of all single-DBD switches. This is in line with the observation I made above, namely that the majority of +/+ switches (61%) do not affect DBDs or any other domains. Another reason is the lack of structures for some TF-DNA complexes in PDB, and hence, – in the NPIDB which predicts the DNA base- and backbone-binding residues in the structures published in PDB. Indeed, only a half of the 14 TFs that I selected for my analysis had such annotated structures.

Finally, my predictions about the specificity and stability of the DNA binding of switching major isoforms are only qualitative and hinge on the assumption that the switching isoforms fold in a similar way. However, the actual DNA-binding specificities of TF isoforms that have different numbers of DBDs or different DBD sequences could be investigated *in vitro* using the multiplexed massively parallel SELEX (Jolma et al., 2010). For each protein isoform, this method would yield the DNA-binding sites of the highest affinity, hence allowing the generation of DNA-binding motifs and the comparison of these motifs between isoforms.

However, although *in vitro*-established DNA-binding motifs of TFs usually correspond well to the motifs found *in vivo* (Jolma et al., 2010, 2013; Liu et al., 2006b; Mukherjee et al., 2004; Wei et al., 2010), the DNA-binding specificity *in vitro* cannot, in general, predict TF-bound loci *in vivo*. The reason is that, apart from the binding motif, the *in vivo* DNA binding is regulated by the nucleosome occupancy (Ballaré et al., 2013; Barozzi et al., 2014; Biddie et al., 2011), the differential DNA methylation (Hainer et al., 2016; Yin et al., 2017) and the cooperative binding of TFs that are simultaneously present in living cells but may not be present *in vitro* (Amin et al., 2015; Biddie et al., 2011; Cernilogar et al., 2019; Kumar et al., 2017). However, the characterisation of the *in vivo* DNA binding of different isoforms of the same TF may be complicated by the need to use isoform-specific antibodies.

Noteworthy, my analysis of switches of the *major* TF isoforms ignores, for the sake of interpretation clarity, any separate roles of *minor* isoforms that may be co-expressed with the major ones. However, examples of functionally important minor TF isoforms have been known for a long time. For instance, (Mangal et al., 1997) showed that the expression level of the major isoform A of the progesterone receptor is constant during the menstrual cycle, while the expression level of its minor isoform B changes. The change in the ratio between the two co-

expressed isoforms, occurring solely due to the changes in the expression level of the minor one, may influence the overall effect of the progesterone, because isoforms A and B were previously shown to have different transcription regulation properties (Kastner et al., 1990; Meyer et al., 1992; Tora et al., 1988). Consequently, a more sophisticated analysis that takes into account functional interactions between at least two co-expressed TF isoforms may be very useful. This analysis, however, would depend on the availability of methods to quantify the isoform expression from single-cell RNA-seq data to detect the co-expression of two or more isoforms of the same TF within the same cell.

Overall, because my analysis described in Chapters 3, 4 and 5 was limited by the incomplete domain collection in InterPro and also did not take into account non-coding RNAs, the alternative translation of mRNAs and the localisation and PTMs of the resulting protein isoforms, it may have underestimated the functional significance of the alternative isoform production and of major isoform switches by human TFs.

5.4.3. Perspective

Here, I outline some research directions that follow from the results that I described in this chapter. First of all, as I suggested in section 5.3.2, my functional predictions for major isoform switches could be experimentally validated. A physiological characterisation of the outlined examples of the brain-non-brain switches would be especially interesting. Below, I briefly describe some experiments that could be done to test my predictions.

In a switch of the major isoforms of a TF, one major isoform is expressed in one set of tissues, while the other major isoform of the same TF is expressed in another set of tissues. Consequently, for verification experiments, one would need to use two groups of relevant human cell lines or primary cells from healthy clinical samples, if available. One would start with the validation of the expression of the switching major isoforms in the two groups of cells at the RNA level (using real-time PCR with isoform-specific primers) and at the protein level (using Western blot on proteins selected with the TF-specific antibody whose binding to both isoforms could be tested *in vitro*). Further, one of the two cell groups could be modified by silencing the endogenous major TF isoform using RNA interference with isoform-specific silencing RNAs and then by transfecting the cells with the other major isoform. In this way, one would control for all factors that are different between the two initial groups of cells, except for the difference between the two major TF isoforms. Finally, one would do a differential analysis of the biological characteristics of interest between the wild type and the modified cells.

For example, using ChIP-seq or CUT&Tag (Kaya-Okur et al., 2019, 2020), one could compare H3K9 methylation profiles between the relevant wild type and modified brain cells to test if the switch of the PRDM2 major isoforms affects histone methylation (Fig. 5.14F). Similarly, one could use ChIP-mass spectrometry to find proteins differentially interacting with TFEB, WT1 and MEF2C major isoforms (Fig. 5.14E, G-H). Also, to test the predicted differential DNA binding affinity of the two major TBX3 isoforms, one could find and compare isoform-specific DNA-binding motifs using protein binding microarrays or SELEX; then, one could verify these motifs by profiling TBX3 binding with ChIP-exo or CUT&RUN in relevant wild type and modified cells and by scanning the obtained peaks with the isoform-specific motifs (Fig. 5.14I). Furthermore, one could test the putative differential transcription regulation by the major isoforms of ZNF12 and ZNF571 (Fig. 5.14B, D) using luciferase reporters transfected into wild type and modified cells. In general, to measure the overall physiological effect of a switch, one could do a differential gene expression and a ChIP-mass spectrometry analyses followed by pathway enrichment analyses in wild type and modified cells to elucidate the molecular processes affected by the switch. Also, by transfecting cells with GFP-fused switching isoforms, one could measure their differential cellular localisation using confocal microscopy.

Secondly, my primary supervisor Professor Nicholas Luscombe and I formulated two more questions about the evolutionary conservation of alternative TF isoforms: (1) If two isoforms are switched between adult human tissues and are conserved between the human and another animal species, do these isoforms also switch in that other species? (2) If yes, are the tissues that switch the isoforms similar between the considered species and human? The conservation of the switches of alternative TF isoforms would suggest the functional importance of these switches and not only single isoforms. Additionally, as I mentioned in section 4.4.3, software tools for the analysis of isoform conservation, such as ThorAxe (Zea et al., 2021), have been recently emerging.

Next, in addition to studying switches of major TF isoforms between adult human tissues, one could study switches that occur in development. To this end, one could quantify the expression of TF isoforms across several time points in a differentiation time course of a human or murine cell lineage, or in murine embryogenesis, find possible novel TF isoform switches, predict their functional role and validate it experimentally. Switches of major TF isoforms during development are indeed known from previous experimental studies. In addition to the isoform switches by FOXP1 and REST that I described in sections 4.1 and 5.1, respectively, PAX6 and CREM also present the well-known examples of developmental

isoform switches. Specifically, PAX6 changes the major isoform during retina development in chicken (Azuma et al., 2005), while CREM switches its major isoform in the development of testis in mice (Foulkes et al., 1992). Consequently, it would be quite interesting to expand the set of known developmental switches of major TF isoforms. Bulk RNA-seq data gathered by some large developmental sequencing initiatives, such as dGTEx (https://dgtex.org/) and BrainSpan (https://www.brainspan.org/), could be useful, if re-quantified per mRNA isoform.

Next, as I mentioned in section 5.4.1, sequences of IDRs and UTRs in alternative TF isoforms deserve to be studied alongside domain sequences. In addition to these regions, alternative isoforms of the same TF could differ by their subcellular localisation signals. For instance, while the two normal DBD+ isoforms of IKZF2 are localised to the nucleus, the pathological DBD- isoforms of this TF are mostly located in the cytoplasm (Zhao et al., 2016). Hence, one could study the scale and predict the biological consequences of the alternative presence of various localisation signals in the human TF isoforms. To this end, one could use the NLSdb database (https://rostlab.org/services/nlsdb/) (Bernhofer et al., 2018; Nair et al., 2003) which provides a vast collection of nuclear localisation and export signals, as well as the LocTree3 software (https://rostlab.org/services/loctree3/) (Goldberg et al., 2012, 2014) which predicts the subcellular localisation of a protein based on its sequence. However, one would need to keep in mind that the interpretation of the results is limited by the possible presence of PTMs that may change the localisation of a protein (Rocks et al., 2005; Ross et al., 2002) and, in general, would be very hard, if at all possible, to account for.

Finally, the 400 switches of the major TF isoforms across adult human tissues that I summarised in this chapter could provide an initial basis for the computational study of the tissue-specific DNA binding of TF isoforms that I outlined in section 4.4.3.

Chapter 6. Discussion

In this thesis, I described my studies of the putative gyre-spanning nucleosomal binding of human TFs and my genome-wide analysis of the human TF isoforms.

6.1. Summary

In Chapter 2, I presented my computational search for evidence of an unusual nucleosomal binding of TBXT and other human TFs in which a TF homodimer binds across DNA gyres. This binding mode, termed the gyre-spanning binding, was observed *in vitro* for TBXT by (Zhu et al., 2018), and in their experiments, the gyre-spanning binding of TBXT stabilised nucleosomes. In the nucleosomal DNA, this binding was manifested by a gapped motif with a gap size approximately equal to the length of one turn of the DNA molecule around the histone core. Based on these observations, I searched for gapped TBXT motifs of different gap sizes in the strongly positioned nucleosomes in the human genome.

I found more than 11,000 matches of the motifs with the gap sizes of 77 and 79 bp (GTM77 and GTM79, respectively) that would be compatible with the TBXT binding across DNA gyres. However, the TBXT ChIP-seq data from hESC-derived mesoderm progenitors did not show the TBXT binding in these matches, and further analysis demonstrated that GTM77 and GTM79 are features of Alu repeats. I additionally searched for matches of gapped motifs of other human TFs, but the matched motifs were either incompatible with the gyre-spanning binding or lay in repeats. With this, I concluded that the gyre-spanning binding of TFs either does not occur in the human genome *in vivo* or is extremely rare.

As my results do not completely eliminate the possibility of this type of TF binding *in vivo*, I thought of additional experiments and analysis that would help to further search for this binding. Specifically, one could use the CUT&RUN data for TBXT that were generated by (Meers et al., 2019) to additionally search for the TBXT binding at the GTMs. Of note, this would require accounting for the mappability issues in Alu repeats. If the CUT&RUN data show the binding signal, one could also profile the nucleosome occupancy at the bound GTMs in two cell lines, only one of which expresses TBXT. A higher occupancy at the bound GTMs in the TBXT-expressing cells would give additional indirect evidence for the gyre-spanning binding of TBXT, although it would still not be possible to resolve the question completely. Finally, I speculate that if such evidence is indeed found, then microscopy could help to further clarify the mode of TBXT binding at GTMs.

In Chapter 3, I described the first part of another project, a computational study of the alternative isoforms of human TFs. First, I found that a higher proportion of TFs, in comparison to non-TFs, produce alternative isoforms (66.5%, in comparison to 62.5%, respectively). I also found that the average number of alternative isoforms is higher for a TF than for a non-TF (2.73, in comparison to 2.45). Next, I showed that this result does not depend on the gene length or the number of exons and holds for almost all deciles of these two characteristics. Moreover, I demonstrated that with the increase of the gene length or the number of exons, the average number of isoforms produced by a gene increases faster for TFs, than for non-TFs. These results demonstrate that TFs more broadly sample the set of all alternative isoforms combinatorially possible for a given number of exons, and hence, the production of alternative isoforms may be especially important for the TF function. Overall, my findings not only confirm previous results, published more than a decade ago, on an up-to-date human genome annotation, but also extend those results by taking into account the gene length and the number of exons and motivate further analysis of the alternative isoforms of human TFs.

To be able to interpret the differences between the alternative protein isoforms of TFs, I did an extensive literature search and compiled a collection of 24 human TFs whose alternative isoforms were studied functionally and characterised based on the presence of certain domains (Tables 1-2). The main characteristics that I list for each TF in these tables are the functional difference between a pair of isoforms, the domain-based explanation of this difference, a literature-based description of the expression patterns of the TF isoforms in various tissues and cell lines, and the corresponding literature references. Additionally, in Table 3 I collected references to the literature about alternative isoforms of another 68 human TFs or small groups of TFs with known alternative isoforms. I hope that the total compendium of 92 TFs that I compiled from the literature will provide a basis for future studies of alternative TF isoforms.

In Chapter 4, I started my computational analysis by considering the presence of a DBD in alternative TF isoforms. I found that almost a third of all human TF genes express DBD- isoforms that by definition cannot function as TFs, and such isoforms comprised, in total, 17% of all human TF isoforms. Moreover, some TF families with a comparable number of factors had considerably different proportions of TFs that produced DBD- isoforms, which may have an evolutionary explanation. Aside from the presence or absence of a DBD, alternative TF isoforms can be distinguished by the DBD sequence. I found that only 15% of single-DBD TFs changed their DBD sequence across isoforms, and the majority of such TFs produced only two different forms of the DBD sequence. Therefore, I expect that the majority of single-DBD TFs

do not change their DNA-binding specificity across alternative isoforms. Additionally, I demonstrated that the single-DBD TFs never change the type of the DBD between their isoforms. Finally, I found that 30% of multi-DBD TFs changed a combination of their DBDs (specifically, the number or types of DBDs) across their isoforms, and TF families differed considerably by the proportion of TFs that change their DBD combinations. Consequently, I expect that about two thirds of all multi-DBD TFs do not change the DNA-binding specificity across their alternative isoforms. However, this does not take into account a possible role in the DNA binding of amino-acid differences outside DBDs.

As about a third of all human TFs lose a DBD in some of their isoforms, I next collaborated with Dr. Koustav Pal to study the expression levels and tissue specificity of the DBD- isoforms to elucidate their possible functional importance. We found that about a half of all DBD- isoforms were expressed in the adult human tissues, and of those, about a third became the top-expressed (major) in at least one tissue. Moreover, when we stratified human TFs into quantiles by the breadth of their expression across the tissues and calculated the median tissue specificity of the DBD+ and DBD- isoforms produced by TFs in each quantile, we found that many DBD- isoforms of broadly expressed TFs are tissue-specific. These findings suggest that at least a subset of DBD- isoforms of human TFs are functional. Furthermore, we found that the median expression levels and the tissue specificities of DBD- isoforms can vary across TF families, which suggests different roles of DBD- isoforms in different families.

Therefore, I computationally predicted the functions of the DBD- isoforms. In order to do so, I first manually annotated the general functions of 262 non-DBD InterPro entries (Supplementary Table 2), based on the InterPro entry descriptions and literature. In addition to aiding my analysis, I hope that this annotation serves as a useful resource for future studies. I also used the well-characterised examples of DBD- TF isoforms that I found in literature and described in Chapter 3. With this, I showed that the most frequent predicted general function of a DBD- isoform was the reversion of the transcription regulation exerted by the corresponding DBD+ isoform, so that the DBD- isoform of a transcription activator was a predicted repressor and vice versa. Importantly, I also found that more than a third of the DBD- isoforms of C2H2 zinc finger TFs and about two-thirds of the DBD- isoforms of other TFs did not have any domains, and hence did not have a predicted function. Such isoforms could either be non-functional or exert a function based on their disordered regions that could bind DNA, RNA, or proteins.

In Chapter 5, I described the last part of my analysis of alternative TF isoforms. Specifically, I studied the switches of the major alternative TF isoforms across human tissues. I found that 85% of TFs with multiple major isoforms switch these isoforms across tissues and that there are, in total, 400 switches (namely, pairs of major isoforms that switch across tissues). The most common are DBD+ to DBD+ (+/+) switches (they comprise 75% of all switches), but only 39% of them affect domains, while another 40% affect the protein sequence outside domains, and the remaining 21% do not affect the protein sequence. The switches that affect the sequence outside domains could change intrinsically disordered regions that are known to be enriched in TFs and, in general, be able to bind DNA, RNA, and proteins. Hence, these changes may affect the affinity or specificity of the TF-DNA binding or the TF binding to other TFs and cofactors. The switches that do not affect the protein sequence of isoforms must affect UTRs and hence – the localisation, stability, or translational efficiency of the isoforms. The analysis and interpretation of the differences in UTRs between the alternative isoforms of human TFs could comprise a separate project.

Next, based on my functional annotation of non-DBDs and the well-studied examples of functional alternative TF isoforms that I found in literature and described in Chapter 3, I predicted the functional consequences of the major isoform switches. As expected, among the +/+ switches, the majority affected the UTRs or the protein sequence outside domains. Also, among the top predicted consequences of the +/+ switches was a change in DNA binding. For the DBD+ to DBD- (+/-) switches, the most frequent predicted consequence was the reversion of the mode of transcription regulation.

As one of the most frequent categories of +/+ switches was a change in the DNA binding, I studied the possible effects of the differences between the DBD sequences in switching major isoforms with a single DBD. I found that these differences may affect DNA binding in different ways. Namely, ERF lost a third of its residues that confer binding stability but retained all the residues that confer binding specificity; hence, if the DNA binding of the truncated isoform occurs, it could be weaker but have a specificity similar to that of the full isoform. In contrast, PGR and NFATC1 lost the majority of specificity-conferring but retained almost all stability-conferring residues; hence, their truncated isoforms could either bind DNA unspecifically or with a specificity different from that of the full isoforms. Finally, RARB lost all of its specificity-conferring and three out of five stability-conferring residues; therefore, its truncated isoform may not be able to bind DNA or may bind it with a different specificity, using different residues.

In my further analysis of the major isoform switches, I demonstrated that, overall, they were not specific to any particular TF family, but certain families were overrepresented among TFs that produced switches of certain functional consequences. Namely, nuclear receptors were overrepresented among TFs whose switches changed the ligand binding; MADS box TFs were overrepresented among factors whose switches changed the protein binding; and C2H2 zinc finger TFs were overrepresented among factors whose switches reverse the mode of transcription regulation.

Finally, I studied the characteristic tissue specificity patterns of the major isoform switches. To this end, I clustered TFs and tissues by the types of the expressed major isoforms (DBD+ or DBD-) and found that the brain and testis made separate clusters, suggesting a unique set of major isoforms that they express. Interestingly, some non-brain tissues also formed tight clusters, which is not readily explained. Based on the evidence from the literature, I speculate that one of these clusters may be related to glucose metabolism. Next, I analysed the seven TF clusters that I obtained and found that many TFs demonstrated clear switch patterns, especially, between the brain and the non-brain tissues. Hence, after describing the predicted functions of switches in each cluster, I considered 34 TFs in several groups with particularly interesting brain-non-brain switches that have not been characterised to date. In my analysis, these switches were predicted to, for instance, turn a TF outside the brain into a non-TF protein in the brain (FOXJ3), to affect transcription repression via changes in a transcription repression domain (ZNF12), to turn a repressor outside the brain into an activator in the brain (ZNF571), or to affect the binding of a partner protein (TFEB). I also considered five more TFs with intriguing switches between other tissues. In conclusion, I suggest that all these switch examples, and probably many more from the switch clusters that I obtained, are worth further experimental investigation.

In summary, I integrated publicly available genomic, transcriptomic, and proteomic data to study the nucleosomal binding and alternative isoforms of the human transcription factors. I also suggested possible directions for future analyses and experiments.

6.2. Perspective

A big unresolved question in molecular biology is how TFs interact with nucleosomes. The gyre-spanning binding of TFs to nucleosomes that I studied in Chapter 2 is only one of several TF-nucleosome binding modes observed *in vitro* by (Zhu et al., 2018). Other modes include end binding, oriented binding, periodic binding and dyad binding (Zhu et al., 2018). However, testing the binding of TFs to nucleosomes *in vivo* is challenging because TFs may bind nucleosomal DNA at its ends, in short periods of time when the DNA is partially unwrapped from the histone core, rather than bind tightly wrapped DNA (North et al., 2012). Additionally, testing about 1,500 known human TFs *in vivo* would be quite expensive.

Hence, *in vivo* predictions generated using deep learning models, although not proving or disproving particular binding modes, could help to guide future experimental efforts. Namely, one may develop a deep learning model that would be trained for each TF separately on *in vitro* NCAP-SELEX data and then predict nucleosome binding for the same TF in the genome. The accuracy of such predictions could be increased using a transfer learning technique where a model learns some general properties of TF-nucleosome binding from several NCAP-SELEX datasets simultaneously and then, based on this prior knowledge, refines the predictions for individual TFs (Novakovsky et al., 2021; Zheng et al., 2021).

However, the decisive evidence for a certain TF-nucleosome binding mode *in vivo* can be obtained only experimentally, and I believe that the required experimental techniques will emerge in the near future.

* * *

In Chapters 3, 4 and 5, I computationally explored alternative isoforms of human TFs. A question about their possible functions is a part of a bigger unresolved question about the functional importance of alternative mRNA isoforms observed *in vivo* (Agosto et al., 2019; Pickrell et al., 2010; Reixachs-Solé and Eyras, 2022; Tress et al., 2017; Weatheritt et al., 2016). Like for any other type of genes, for TF-encoding genes the resolution of this question can be guided, at least in part, by the analysis of conservation of alternative mRNA isoforms across species (see section 5.4.3 for the discussion of particular questions related to the conservation of TF isoforms). Although a lack of conservation would not mean a lack of function, conserved isoforms could be prioritised in experimental studies. The search for conserved isoforms and

the assessment of the evolutionary span of their conservation will be aided by the ongoing progress in the assembly and annotation of the genomes of various non-model species by such consortia as The Genome 10K Project (Genome 10K Community of Scientists, 2009; Koepfli et al., 2015; Rhie et al., 2021) and The 10,000 Fish Genomes Project (Fan et al., 2020). Of note, a high-quality genome annotation is usually, at least in part, based on RNA sequencing data (Aken et al., 2016) which may also serve for isoform quantification once the annotation is produced.

Another major outstanding problem is the accurate annotation and quantification of RNA isoforms, as short sequencing reads often do not capture the full length of an isoform. Further advancement of the long-read RNA sequencing (Bayega et al., 2021; Leung et al., 2021; Sharon et al., 2013; Zhu et al., 2021) will improve the accuracy of isoform annotation, as well as quantification, and make the analysis of TF isoform expression across tissues or developmental time points more reliable.

Finally, a number of TFs are known to co-express their isoforms in one cell with important functional implications (Fernandez and Gudas, 2009; Figtree et al., 2003; Oakley et al., 1999; Qi et al., 2016; Sun et al., 1996; Yu et al., 2010). However, with bulk RNA sequencing it is not possible to differentiate between the co-expression of TF isoforms in one cell and their separate expression in different cell subtypes within a tissue. Consequently, attaining the ability to distinguish between the two scenarios, as well as to investigate and interpret the co-expression of TF isoforms, is another major challenge in TF biology. The emerging techniques of the full-length transcript sequencing in single cells (Al'Khafaji et al., 2021; Joglekar et al., 2021; Tian et al., 2021; Volden and Vollmers, 2022), as well as single-cell proteomics (Schoof et al., 2021), will open an exciting new avenue of research into TF isoform co-expression.

References

Agosto, L.M., Gazzara, M.R., Radens, C.M., Sidoli, S., Baeza, J., Garcia, B.A., and Lynch, K.W. (2019). Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing. Genome Res. *29*, 2046–2055. https://doi.org/10.1101/gr.248435.119.

Akalin, A., Franke, V., Vlahovicek, K., Mason, C., and Schubeler, D. (2014). genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinforma. Oxf. Engl. https://doi.org/10.1093/bioinformatics/btu775.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., et al. (2016). The Ensembl gene annotation system. Database *2016*, baw093. https://doi.org/10.1093/database/baw093.

Al Chiblak, M., Steinbeck, F., Thiesen, H.-J., and Lorenz, P. (2019). DUF3669, a "domain of unknown function" within ZNF746 and ZNF777, oligomerizes and contributes to transcriptional repression. BMC Mol. Cell Biol. *20*, 60. https://doi.org/10.1186/s12860-019-0243-y.

Alders, M., Ryan, A., Hodges, M., Bliek, J., Feinberg, A.P., Privitera, O., Westerveld, A., Little, P.F.R., and Mannens, M. (2000). Disruption of a Novel Imprinted Zinc-Finger Gene, ZNF215, in Beckwith-Wiedemann Syndrome. Am. J. Hum. Genet. *66*, 1473–1484. https://doi.org/10.1086/302892.

Al'Khafaji, A.M., Smith, J.T., Garimella, K.V., Babadi, M., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M.A., Popic, V., Blaum, E.M., et al. (2021). High-throughput RNA isoform sequencing using programmable cDNA concatenation. bioRxiv *2021.10.01.462818*. https://doi.org/10.1101/2021.10.01.462818.

Amae, S., Fuse, N., Yasumoto, K., Sato, S., Yajima, I., Yamamoto, H., Udono, T., Durlu, Y.K., Tamai, M., Takahashi, K., et al. (1998). Identification of a Novel Isoform of Microphthalmia-Associated Transcription Factor That Is Enriched in Retinal Pigment Epithelium. Biochem. Biophys. Res. Commun. *247*, 710–715. https://doi.org/10.1006/bbrc.1998.8838.

Amemiya-Kudo, M., Shimano, H., Hasty, A.H., Yahagi, N., Yoshikawa, T., Matsuzaka, T., Okazaki, H., Tamura, Y., Iizuka, Y., Ohashi, K., et al. (2002). Transcriptional activities of nuclear SREBP-1a, -1c, and -2 to different target promoters of lipogenic and cholesterogenic genes. J. Lipid Res. *43*, 1220–1235.

https://doi.org/10.1194/jlr.M100417-JLR200.

Amin, S., Donaldson, I.J., Zannino, D.A., Hensman, J., Rattray, M., Losa, M., Spitz, F., Ladam, F., Sagerström, C., and Bobola, N. (2015). Hoxa2 Selectively Enhances Meis Binding to Change a Branchial Arch Ground State. Dev. Cell *32*, 265–277. https://doi.org/10.1016/j.devcel.2014.12.024.

Amoutzias, G.D., Robertson, D.L., Van de Peer, Y., and Oliver, S.G. (2008). Choose your partners: dimerization in eukaryotic transcription factors. Trends Biochem. Sci. *33*, 220–229. https://doi.org/10.1016/j.tibs.2008.02.002.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461. https://doi.org/10.1038/nature12787.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Asahara, H., Dutta, S., Kao, H.-Y., Evans, R.M., and Montminy, M. (1999). Pbx-Hox Heterodimers Recruit Coactivator-Corepressor Complexes in an Isoform-Specific Manner. Mol. Cell. Biol. *19*, 8219–8225. https://doi.org/10.1128/MCB.19.12.8219.

Aurora, R., and Herr, W. (1992). Segments of the POU domain influence one another's DNA-binding specificity. Mol. Cell. Biol. *12*, 455–467. https://doi.org/10.1128%2Fmcb.12.2.455.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366. https://doi.org/10.1038/s41588-021-00782-6.

Azuma, N., Tadokoro, K., Asaka, A., Yamada, M., Yamaguchi, Y., Handa, H., Matsushima, S., Watanabe, T., Kohsaka, S., Kida, Y., et al. (2005). The Pax6 isoform bearing an alternative spliced exon promotes the development of the neural retinal structure. Hum. Mol. Genet. *14*, 735–745. https://doi.org/10.1093/hmg/ddi069.

Bach, I., and Yaniv, M. (1993). More potent transcriptional activators or a transdominant inhibitor of the HNF1 homeoprotein family are generated by alternative RNA processing. EMBO J. *12*, 4229–4242. https://doi.org/10.1002/j.1460-2075.1993.tb06107.x.

Bader, A.G., and Vogt, P.K. (2006). Leucine Zipper Transcription Factors: bZIP Proteins. In Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine, (Berlin,

Heidelberg: Springer), pp. 964–967.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res. *43*, W39–W49. https://doi.org/10.1093/nar/gkv416.

Ballaré, C., Castellano, G., Gaveglia, L., Althammer, S., González-Vallinas, J., Eyras, E., Le Dily, F., Zaurin, R., Soronellas, D., Vicent, G.P., et al. (2013). Nucleosome-Driven Transcription Factor Binding and Gene Regulation. Mol. Cell *49*, 67–79. https://doi.org/10.1016/j.molcel.2012.10.019.

Balwierz, P.J., Carninci, P., Daub, C.O., Kawai, J., Hayashizaki, Y., Van Belle, W., Beisel, C., and van Nimwegen, E. (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. *10*, R79. https://doi.org/10.1186/gb-2009-10-7-r79.

Bamberger, C.M., Bamberger, A.M., Castro, M. de, and Chrousos, G.P. (1995). Glucocorticoid receptor beta, a potential endogenous inhibitor of glucocorticoid action in humans. J. Clin. Invest. *95*, 2435–2441. https://doi.org/10.1172/JCI117943.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science *338*, 1587–1593. https://doi.org/10.1126/science.1230612.

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., and Natoli, G. (2014). Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. Mol. Cell *54*, 844–857. https://doi.org/10.1016/j.molcel.2014.04.006.

Bayega, A., Oikonomopoulos, S., Gregoriou, M.-E., Tsoumani, K.T., Giakountis, A., Wang, Y.C., Mathiopoulos, K.D., and Ragoussis, J. (2021). Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci. Rep. *11*, 7878. https://doi.org/10.1038/s41598-021-86753-7.

Bégay, V., Baumeier, C., Zimmermann, K., Heuser, A., and Leutz, A. (2018). The C/EBPβ LIP isoform rescues loss of C/EBPβ function in the mouse. Sci. Rep. *8*, 8417. https://doi.org/10.1038/s41598-018-26579-y.

Belluti, S., Rigillo, G., and Imbriano, C. (2020). Transcription Factors in Cancer: When Alternative Splicing Determines Opposite Cell Fates. Cells *9*, 760. https://doi.org/10.3390/cells9030760.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–

300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berlato, C., Chan, K.V., Price, A.M., Canosa, M., Scibetta, A.G., and Hurst, H.C. (2011). Alternative TFAP2A isoforms have distinct activities in breast cancer. Breast Cancer Res. *13*, R23. https://doi.org/10.1186/bcr2838.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.

Bernhofer, M., Goldberg, T., Wolf, S., Ahmed, M., Zaugg, J., Boden, M., and Rost, B. (2018). NLSdb—major update for database of nuclear localization signals and nuclear export signals. Nucleic Acids Res. *46*, D503–D508. https://doi.org/10.1093/nar/gkx1021.

Beverly, L.J., and Capobianco, A.J. (2003). Perturbation of Ikaros isoform selection by MLV integration is a cooperative event in NotchIC-induced T cell leukemogenesis. Cancer Cell *3*, 551–564. https://doi.org/10.1016/S1535-6108(03)00137-5.

Bharti, B., and Mishra, R. (2015). Spleen-specific isoforms of Pax5 and Ataxin-7 as potential proteomic markers of lymphoma-affected spleen. Mol. Cell. Biochem. *402*, 181–191. https://doi.org/10.1007/s11010-014-2325-7.

Bharti, K., Liu, W., Csermely, T., Bertuzzi, S., and Arnheiter, H. (2008). Alternative promoter use in eye development: the complex role and regulation of the transcription factor MITF. Development *135*, 1169–1178. https://doi.org/10.1242/dev.014142.

Bickmore, W.A., Oghene, K., Little, M.H., Seawright, A., van Heyningen, V., and Hastie, N.D. (1992). Modulation of DNA Binding Specificity by Alternative Splicing of the Wilms Tumor wt1 Gene Transcript. Science *257*, 235–237. https://doi.org/10.1126/science.1321494.

Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A., Schiltz, R.L., Miranda, T.B., Sung, M.-H., Trump, S., Lightman, S.L., et al. (2011). Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. Mol. Cell *43*, 145–155. https://doi.org/10.1016/j.molcel.2011.06.016.

Blake, J.A., Baldarelli, R., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J., and the Mouse Genome Database Group (2021). Mouse Genome Database (MGD): Knowledgebase for mouse–human comparative biology. Nucleic Acids Res. *49*, D981–D987. https://doi.org/10.1093/nar/gkaa1083.

Blokken, J., De Rijck, J., Christ, F., and Debyser, Z. (2017). Protein–protein and protein–chromatin interactions of LEDGF/p75 as novel drug targets. Drug Discov. Today Technol. *24*, 25–31. https://doi.org/10.1016/j.ddtec.2017.11.002.

Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C., and Hochreiter, S. (2015). msa: an R package for multiple sequence alignment. Bioinformatics *31*, 3997–3999. https://doi.org/10.1093/bioinformatics/btv494.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842–1855.e16. https://doi.org/10.1016/j.cell.2018.10.042.

Breitbart, R.E., Liang, C.S., Smoot, L.B., Laheru, D.A., Mahdavi, V., and Nadal-Ginard, B. (1993). A fourth human MEF2 transcription factor, hMEF2D, is an early marker of the myogenic lineage. Development *118*, 1095–1106. https://doi.org/10.1242/dev.118.4.1095.

Brix, D.M., Bundgaard Clemmensen, K.K., and Kallunki, T. (2020). Zinc Finger Transcription Factor MZF1—A Specific Regulator of Cancer Invasion. Cells *9*, 223. https://doi.org/10.3390/cells9010223.

Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M., and Barkai, N. (2020). Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. Mol. Cell *79*, 459–471.e4. https://doi.org/10.1016/j.molcel.2020.05.032.

Brown, N.R., Noble, M.E.M., Endicott, J.A., and Johnson, L.N. (1999). The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Nat. Cell Biol. *1*, 438–443. https://doi.org/10.1038/15674.

Brown-Bryan, T.A., Leoh, L.S., Ganapathy, V., Pacheco, F.J., Mediavilla-Varela, M., Filippova, M., Linkhart, T.A., Gijsbers, R., Debyser, Z., and Casiano, C.A. (2008). Alternative Splicing and Caspase-Mediated Cleavage Generate Antagonistic Variants of the Stress Oncoprotein LEDGF/p75. Mol. Cancer Res. *6*, 1293–1307. https://doi.org/10.1158/1541-7786.MCR-08-0125.

Bruce, A.E.E., and Winklbauer, R. (2020). Brachyury in the gastrula of basal vertebrates. Mech. Dev. *163*, 103625. https://doi.org/10.1016/j.mod.2020.103625.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. Mol. Cell *46*, 871–883. https://doi.org/10.1016/j.molcel.2012.05.039.

Bullock, A.N., Debreczeni, J., Amos, A.L., Knapp, S., and Turk, B.E. (2005). Structure and Substrate Specificity of the Pim-1 Kinase. J. Biol. Chem. *280*, 41675–41682. https://doi.org/10.1074/jbc.M510711200.

Bürglin, T.R., and Affolter, M. (2016). Homeodomain proteins: an update. Chromosoma *125*, 497–521. https://doi.org/10.1007/s00412-015-0543-8.

Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M., et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. *49*, D437–D451. https://doi.org/10.1093/nar/gkaa1038.

Byeon, G.W., Cenik, E.S., Jiang, L., Tang, H., Das, R., and Barna, M. (2021). Functional and structural basis of extreme conservation in vertebrate 5′ untranslated regions. Nat. Genet. *53*, 729–741. https://doi.org/10.1038/s41588-021-00830-1.

Caballero, R., Setien, F., Lopez-Serra, L., Boix-Chornet, M., Fraga, M.F., Ropero, S., Megias, D., Alaminos, M., Sanchez-Tapia, E.M., Montoya, M.C., et al. (2007). Combinatorial effects of splice variants modulate function of Aiolos. J. Cell Sci. *120*, 2619–2630. https://doi.org/10.1242/jcs.007344.

Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D.B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., von Paleske, L., Renders, S., et al. (2014). Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis. Cell Stem Cell *15*, 507–522. https://doi.org/10.1016/j.stem.2014.07.005.

Cakiroglu, S. A., Steinhauser, S., Smith, J., Xing, W., and Luscombe, N. M. (2021). ChromWave: Deciphering the DNA-encoded competition between transcription factors and nucleosomes with deep neural networks. BioRxiv 2021.03.19.436198. https://doi.org/10.1101/2021.03.19.436198.

Calabretta, S., and Richard, S. (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. Trends Biochem. Sci. *40*, 662–672. https://doi.org/10.1016/j.tibs.2015.08.012.

Camacho-Vanegas, O., Till, J., Miranda-Lorenzo, I., Ozturk, B., Camacho, S.C., and Martignetti, J.A. (2013). Shaking the family tree: identification of novel and biologically active alternatively spliced isoforms across the KLF family of transcription factors. FASEB J. *27*, 432–436. https://doi.org/10.1096/fj.12-220319.

Candi, E., Dinsdale, D., Rufini, A., Salomoni, P., Knight, R.A., Mueller, M., Krammer, P.H., and Melino, G. (2007). TAp63 and ΔNp63 in Cancer and Epidermal Development. Cell Cycle *6*, 274–284. https://doi.org/10.4161/cc.6.3.3797.

de Castro, M., Elliot, S., Kino, T., Bamberger, C., Karl, M., Webster, E., and Chrousos, G.P. (1996). The Non-Ligand Binding β-Isoform of the Human Glucocorticoid Receptor (hGRβ): Tissue Levels, Mechanism of Action, and Potential Physiologic Role. Mol. Med. *2*, 597–607. https://doi.org/10.1007/BF03401643.

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173. https://doi.org/10.1093/nar/gkab1113.

Cernilogar, F.M., Hasenöder, S., Wang, Z., Scheibner, K., Burtscher, I., Sterr, M., Smialowski, P., Groh, S., Evenroed, I.M., Gilfillan, G.D., et al. (2019). Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. Nucleic Acids Res. *47*, 9069–9086. https://doi.org/10.1093/nar/gkz627.

Cervera, A.M., Apostolova, N., Luna-Crespo, F., Sanjuan-Pla, A., Garcia-Bou, R., and McCreath, K.J. (2006). An alternatively spliced transcript of the PHD3 gene retains prolyl hydroxylase activity. Cancer Lett. *233*, 131–138. https://doi.org/10.1016/j.canlet.2005.03.004.

Chang, J.-W., Zhang, W., Yeh, H.-S., Park, M., Yao, C., Shi, Y., Kuang, R., and Yong, J. (2018). An integrative model for alternative polyadenylation, IntMAP, delineates mTOR-modulated endoplasmic reticulum stress response. Nucleic Acids Res. *46*, 5996–6008. https://doi.org/10.1093/nar/gky340.

Chantalat, E., Boudou, F., Laurell, H., Palierne, G., Houtman, R., Melchers, D., Rochaix, P., Filleron, T., Stella, A., Burlet-Schiltz, O., et al. (2016). The AF-1-deficient estrogen receptor ERα46 isoform is frequently expressed in human breast tumors. Breast Cancer Res. *18*, 123. https://doi.org/10.1186/s13058-016-0780-7.

Cheedipudi, S., Puri, D., Saleh, A., Gala, H.P., Rumman, M., Pillai, M.S., Sreenivas, P., Arora, R., Sellathurai, J., Schrøder, H.D., et al. (2015). A fine balance: epigenetic control of cellular quiescence by the tumor suppressor PRDM2/RIZ at a bivalent domain in the cyclin a gene. Nucleic Acids Res. *43*, 6236–6256. https://doi.org/10.1093/nar/gkv567.

Chellappa, K., Deol, P., Evans, J.R., Vuong, L.M., Chen, G., Briançon, N., Bolotin, E., Lytle, C., Nair, M.G., and Sladek, F.M. (2016). Opposing roles of nuclear receptor HNF4α isoforms in colitis and colitis-associated colon cancer. eLife *5*, e10903. https://doi.org/10.7554/eLife.10903.

Chen, D., Liu, Y., Yin, S., Qiu, J., Jin, Q., King, G.J., Wang, J., Ge, X., and Li, Z. (2020).

Alternatively Spliced BnaPAP2.A7 Isoforms Play Opposing Roles in Anthocyanin Biosynthesis of Brassica napus L. Front. Plant Sci. *11*. https://doi.org/10.3389/fpls.2020.00983.

Chen, L., Kostadima, M., Martens, J.H.A., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033. https://doi.org/10.1126/science.1251033.

Cheng, Y., Wang, Y., Li, Y., Deng, Y., Hu, J., Mo, X., Li, N., Li, Y., Luo, N., Yuan, W., et al. (2006). A novel human gene ZNF415 with five isoforms inhibits AP-1- and p53-mediated transcriptional activity. Biochem. Biophys. Res. Commun. *351*, 33–39. https://doi.org/10.1016/j.bbrc.2006.09.161.

Choi, H.-S., Chung, M., Tzameli, I., Simha, D., Lee, Y.-K., Seol, W., and Moore, D.D. (1997). Differential Transactivation by Two Isoforms of the Orphan Nuclear Hormone Receptor CAR. J. Biol. Chem. *272*, 23565–23571. https://doi.org/10.1074/jbc.272.38.23565.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*. https://doi.org/10.1126/science.aar2555.

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., et al. (2011). Modernizing Reference Genome Assemblies. PLOS Biol. *9*, e1001091. https://doi.org/10.1371/journal.pbio.1001091.

Crist, R.C., Roth, J.J., Waldman, S.A., and Buchberg, A.M. (2011). A Conserved Tissue-Specific Homeodomain-Less Isoform of MEIS1 Is Downregulated in Colorectal Cancer. PLOS ONE *6*, e23665. https://doi.org/10.1371/journal.pone.0023665.

Cui, H., Schlesinger, J., Schoenhals, S., Tönjes, M., Dunkel, I., Meierhofer, D., Cano, E., Schulz, K., Berger, M.F., Haack, T., et al. (2016). Phosphorylation of the chromatin remodeling factor DPF3a induces cardiac hypertrophy through releasing HEY repressors from DNA. Nucleic Acids Res. *44*, 2538–2553. https://doi.org/10.1093/nar/gkv1244.

Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. Nucleic Acids Res. *50*, D988–D995. https://doi.org/10.1093/nar/gkab1049.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

Dashty, M. (2013). A quick look at biochemistry: Carbohydrate metabolism. Clin. Biochem. *46*, 1339–1352. https://doi.org/10.1016/j.clinbiochem.2013.04.027.

Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution. J. Mol. Biol. *319*, 1097–1113. https://doi.org/10.1016/S0022-2836(02)00386-8.

Dhar-Chowdhury, P., Malester, B., Rajacic, P., and Coetzee, W.A. (2007). The regulation of ion channels and transporters by glycolytically derived ATP. Cell. Mol. Life Sci. *64*, 3069–3083. https://doi.org/10.1007/s00018-007-7332-3.

Di Ruscio, A., Ebralidze, A.K., Benoukraf, T., Amabile, G., Goff, L.A., Terragni, J., Figueroa, M.E., De Figueiredo Pontes, L.L., Alberich-Jorda, M., Zhang, P., et al. (2013). DNMT1-interacting RNAs block gene-specific DNA methylation. Nature *503*, 371–376. https://doi.org/10.1038/nature12598.

Dolfini, D., Minuzzo, M., Pavesi, G., and Mantovani, R. (2012). The Short Isoform of NF-YA Belongs to the Embryonic Stem Cell Transcription Factor Circuitry. STEM CELLS *30*, 2450–2459. https://doi.org/10.1002/stem.1232.

Du, W.W., Yang, W., Liu, E., Yang, Z., Dhaliwal, P., and Yang, B.B. (2016). Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. Nucleic Acids Res. *44*, 2846–2858. https://doi.org/10.1093/nar/gkw027.

Du, W.W., Yang, W., Chen, Y., Wu, Z.-K., Foster, F.S., Yang, Z., Li, X., and Yang, B.B. (2017). Foxo3 circular RNA promotes cardiac senescence by modulating multiple factors associated with stress and senescence responses. Eur. Heart J. *38*, 1402–1412. https://doi.org/10.1093/eurheartj/ehw001.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchinson, G. (2002). Biological sequence analysis. Probabilistic models of proteins and nucleic acids. Cambridge University Press.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208. https://doi.org/10.1038/nrm1589.

Echigoya, K., Koyama, M., Negishi, L., Takizawa, Y., Mizukami, Y., Shimabayashi, H., Kuroda, A., and Kurumizaka, H. (2020). Nucleosome binding by the pioneer transcription factor OCT4. Sci. Rep. *10*, 11832. https://doi.org/10.1038/s41598-020-68850-1.

Eeckhoute, J., Moerman, E., Bouckenooghe, T., Lukoviak, B., Pattou, F., Formstecher, P., Kerr-Conte, J., Vandewalle, B., and Laine, B. (2003). Hepatocyte Nuclear Factor 4α Isoforms Originated from the P1 Promoter Are Expressed in Human Pancreatic β-Cells and Exhibit Stronger Transcriptional Potentials than P2 Promoter-Driven Isoforms. Endocrinology *144*, 1686–1694. https://doi.org/10.1210/en.2002-0024.

Ellis, D.J.P., Dehm, S.M., and Bonham, K. (2006). The modification of Sp3 isoforms by SUMOylation has differential effects on the SRC1A promoter. Gene *379*, 68–78. https://doi.org/10.1016/j.gene.2006.04.015.

Englander, E.W., Wolffe, A.P., and Howard, B.H. (1993). Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. J. Biol. Chem. *268*, 19565–19573. https://doi.org/10.1016/S0021-9258(19)36553-6.

Epstein, J.A., Glaser, T., Cai, J., Jepeal, L., Walton, D.S., and Maas, R.L. (1994). Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing. Genes Dev. *8*, 2022–2034. https://doi.org/10.1101/gad.8.17.2022.

Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and Hahn, S. (2020). A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. Mol. Cell *78*, 890–902.e6. https://doi.org/10.1016/j.molcel.2020.04.020.

Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W.B., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development *142*, 2121–2135. https://doi.org/10.1242/dev.117838.

Fan, G., Song, Y., Yang, L., Huang, X., Zhang, S., Zhang, M., Yang, X., Chang, Y., Zhang, H., Li, Y., Liu, S., Yu, L., Chu, J., Seim, I., et al. (2020). Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). GigaScience *9*, giaa080. https://doi.org/10.1093/gigascience/giaa080.

Fan, W., Huang, X., Chen, C., Gray, J., and Huang, T. (2004). TBX3 and Its Isoform TBX3+2a Are Functionally Distinctive in Inhibition of Senescence and Are Overexpressed in a Subset of Breast Cancer Cell Lines. Cancer Res. *64*, 5132–5139. https://doi.org/10.1158/0008-5472.CAN-04-0615.

Federico, A., and Monti, S. (2020). hypeR: an R package for geneset enrichment workflows.

Bioinformatics *36*, 1307–1308. https://doi.org/10.1093/bioinformatics/btz700.

Fernandez, C.C., and Gudas, L.J. (2009). The truncated Hoxa1 protein interacts with Hoxa1 and Pbx1 in stem cells. J. Cell. Biochem. *106*, 427–443. https://doi.org/10.1002/jcb.22023.

Figtree, G.A., McDonald, D., Watkins, H., and Channon, K.M. (2003). Truncated Estrogen Receptor α 46-kDa Isoform in Human Endothelial Cells. Circulation *107*, 120–126. https://doi.org/10.1161/01.CIR.0000043805.11780.F5.

Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. eLife *5*, e10921. https://doi.org/10.7554/eLife.10921.

Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470. https://doi.org/10.1038/nature13182.

Foulkes, N.S., Mellström, B., Benusiglio, E., and Sassone-Corsi, P. (1992). Developmental switch of CREM function during spermatogenesis: from antagonist to activator. Nature *355*, 80–84. https://doi.org/10.1038/355080a0.

Fraenkel, E., and Pabo, C.O. (1998). Comparison of X-ray and NMR structures for the Antennapedia homeodomain–DNA complex. Nat. Struct. Biol. *5*, 692–697. https://doi.org/10.1038/1382.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.

Frietze, S., and Farnham, P.J. (2011). Transcription Factor Effector Domains. In A Handbook of Transcription Factors, T.R. Hughes, ed. (Dordrecht: Springer Netherlands), pp. 261–277.

Gabut, M., Samavarchi-Tehrani, P., Wang, X., Slobodeniuc, V., O'Hanlon, D., Sung, H.-K., Alvarez, M., Talukder, S., Pan, Q., Mazzoni, E.O., et al. (2011). An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. Cell *147*, 132–146. https://doi.org/10.1016/j.cell.2011.08.023.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of Nucleosome Positioning in the Human Genome. PLOS Genet. *8*, e1003036. https://doi.org/10.1371/journal.pgen.1003036.

Gao, Y., Wang, X., Han, J., Xiao, Z., Chen, B., Su, G., and Dai, J. (2010). The novel OCT4 spliced variant OCT4B1 can generate three protein isoforms by alternative splicing into OCT4B. J. Genet. Genomics *37*, 461–465. https://doi.org/10.1016/S1673-8527(09)60065-5.

Garbett, S.P., Stephens, J., Simonov, K., Xie, Y., Dong, Z., Wickham, H., Horner, J., reikoch, Beasley, W., O'Connor, B., et al. (2022). yaml: Methods to convert R data to YAML and back.

Garcia, D.A., Johnson, T.A., Presman, D.M., Fettweis, G., Wagh, K., Rinaldi, L., Stavreva, D.A., Paakinaho, V., Jensen, R.A.M., Mandrup, S., et al. (2021). An intrinsically disordered region-mediated confinement state contributes to the dynamics and function of transcription factors. Mol. Cell *81*, 1484–1498.e6. https://doi.org/10.1016/j.molcel.2021.01.013.

Garton, M., Najafabadi, H.S., Schmitges, F.W., Radovani, E., Hughes, T.R., and Kim, P.M. (2015). A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. Nucleic Acids Res. *43*, 9147–9157. https://doi.org/10.1093/nar/gkv919.

Gehring, W.J., Affolter, M., and Bürglin, T. (1994). Homeodomain proteins. Annu. Rev. Biochem. *63*, 487–526. https://doi.org/10.1146/annurev.bi.63.070194.002415.

Gellersen, B., Kempf, R., and Telgmann, R. (1997). Human Endometrial Stromal Cells Express Novel Isoforms of the Transcriptional Modulator CREM and Up-Regulate ICER in the Course of Decidualization. Mol. Endocrinol. *11*, 97–113. https://doi.org/10.1210/mend.11.1.9875.

Genome 10K Community of Scientists. (2009). Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. J. Hered. *100*, 659–674. https://doi.org/10.1093/jhered/esp086.

Ghosh, R.P., Horowitz-Scherer, R.A., Nikitina, T., Shlyakhtenko, L.S., and Woodcock, C.L. (2010). MeCP2 Binds Cooperatively to Its Substrate and Competes with Histone H1 for Chromatin Binding Sites. Mol. Cell. Biol. *30*, 4656–4670. https://doi.org/10.1128/MCB.00379-10.

Girardet, C., Walker, W.H., and Habener, J.F. (1996). An alternatively spliced polycistronic mRNA encoding cyclic adenosine 3',5'-monophosphate (cAMP)-responsive transcription factor CREB (cAMP response element-binding protein) in human testis extinguishes expression of an internally translated inhibitor CREB isoform. Mol. Endocrinol. *10*, 879–891. https://doi.org/10.1210/mend.10.7.8813728.

Glur, C. (2020). data.tree: General purpose hierarchical data structure.

Goeke, S., Greene, E.A., Grant, P.K., Gates, M.A., Crowner, D., Aigaki, T., and Giniger, E. (2003). Alternative splicing of lola generates 19 transcription factors controlling axon guidance in Drosophila. Nat. Neurosci. *6*, 917–924. https://doi.org/10.1038/nn1105.

Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. Bioinformatics *28*, i458–i465. https://doi.org/10.1093/bioinformatics/bts390.

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., et al. (2014). LocTree3 prediction of localization. Nucleic Acids Res. *42*, W350–W355. https://doi.org/10.1093/nar/gku396.

Goldman, S., and Shalev, E. (2006). Difference in Progesterone-Receptor Isoforms Ratio Between Early and Late First-Trimester Human Trophoblast Is Associated with Differential Cell Invasion and Matrix Metalloproteinase 2 Expression. Biol. Reprod. *74*, 13–22. https://doi.org/10.1095/biolreprod.105.044925.

Goodson, M.L., Jonas, B.A., and Privalsky, M.L. (2005). Alternative mRNA Splicing of SMRT Creates Functional Diversity by Generating Corepressor Isoforms with Different Affinities for Different Nuclear Receptors. J. Biol. Chem. *280*, 7493–7503. https://doi.org/10.1074/jbc.M411514200.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J. Mol. Biol. *313*, 903–919. https://doi.org/10.1006/jmbi.2001.5080.

Grajkowska, L.T., Ceribelli, M., Lau, C.M., Warren, M.E., Tiniakou, I., Nakandakari Higa, S., Bunin, A., Haecker, H., Mirny, L.A., Staudt, L.M., et al. (2017). Isoform-Specific Expression and Feedback Regulation of E Protein TCF4 Control Dendritic Cell Lineage Specification. Immunity *46*, 65–77. https://doi.org/10.1016/j.immuni.2016.11.006.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.

Grishkevich, V., and Yanai, I. (2014). Gene length and expression level shape genomic novelties. Genome Res. *24*, 1497–1503. https://doi.org/10.1101/gr.169722.113.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

Guo, X., Bulyk, M.L., and Hartemink, A.J. (2011). Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. In Biocomputing 2012,

(WORLD SCIENTIFIC), pp. 104–115. https://doi.org/10.1142/9789814366496_0011.

Haberle, V., Forrest, A.R.R., Hayashizaki, Y., Carninci, P., and Lenhard, B. (2015). CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. *43*, e51. https://doi.org/10.1093/nar/gkv054.

Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T., and White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. *29*, 41–43. https://doi.org/10.1093/nar/29.1.41.

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. *41*, D387–D395. https://doi.org/10.1093/nar/gks1234.

Hahm, K., Ernst, P., Lo, K., Kim, G.S., Turck, C., and Smale, S.T. (1994). The lymphoid transcription factor LyF-1 is encoded by specific, alternatively spliced mRNAs derived from the Ikaros gene. Mol. Cell. Biol. *14*, 7111–7123. https://doi.org/10.1128/mcb.14.11.7111-7123.1994.

Hai, T., and Curran, T. (1991). Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. Proc. Natl. Acad. Sci. *88*, 3720–3724. https://doi.org/10.1073/pnas.88.9.3720.

Hainer, S.J., McCannell, K.N., Yu, J., Ee, L.-S., Zhu, L.J., Rando, O.J., and Fazzio, T.G. (2016). DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells. eLife *5*, e21964. https://doi.org/10.7554/eLife.21964.

Hammes, A., Guo, J.-K., Lutsch, G., Leheste, J.-R., Landrock, D., Ziegler, U., Gubler, M.-C., and Schedl, A. (2001). Two Splice Variants of the Wilms' Tumor 1 Gene Have Distinct Functions during Sex Determination and Nephron Formation. Cell *106*, 319–329. https://doi.org/10.1016/S0092-8674(01)00453-6.

Han, H.-S., Kang, G., Kim, J.S., Choi, B.H., and Koo, S.-H. (2016). Regulation of glucose metabolism from a liver-centric perspective. Exp. Mol. Med. *48*, e218. https://doi.org/10.1038/emm.2015.122.

Harries, L.W., Ellard, S., Stride, A., Morgan, N.G., Hattersley, A.T., and The European MODY consortium (2006). Isomers of the TCF1 gene encoding hepatocyte nuclear factor-1 alpha show differential expression in the pancreas and define the relationship between mutation position and clinical phenotype in monogenic diabetes. Hum. Mol. Genet. *15*, 2216–2224. https://doi.org/10.1093/hmg/ddl147.

Harries, L.W., Locke, J.M., Shields, B., Hanley, N.A., Hanley, K.P., Steele, A., Njølstad, P.R., Ellard, S., and Hattersley, A.T. (2008). The Diabetic Phenotype in HNF4A Mutation

Carriers Is Moderated By the Expression of HNF4A Isoforms From the P1 Promoter During Fetal Development. Diabetes *57*, 1745–1752. https://doi.org/10.2337/db07-1742.

Hemesath, T.J., Steingrímsson, E., McGill, G., Hansen, M.J., Vaught, J., Hodgkinson, C.A., Arnheiter, H., Copeland, N.G., Jenkins, N.A., and Fisher, D.E. (1994). microphthalmia, a critical factor in melanocyte development, defines a discrete transcription factor family. Genes Dev. *8*, 2770–2780. https://doi.org/10.1101/gad.8.22.2770.

Hermosilla, V.E., Hepp, M.I., Escobar, D., Farkas, C., Riffo, E.N., Castro, A.F., and Pincheira, R. (2017). Developmental SALL2 transcription factor: a new player in cancer. Carcinogenesis *38*, 680–690. https://doi.org/10.1093/carcin/bgx036.

Hermosilla, V.E., Salgado, G., Riffo, E., Escobar, D., Hepp, M.I., Farkas, C., Galindo, M., Morín, V., García-Robles, M.A., Castro, A.F., et al. (2018). SALL2 represses cyclins D1 and E1 expression and restrains G1/S cell cycle transition and cancer-related phenotypes. Mol. Oncol. *12*, 1026–1046. https://doi.org/10.1002/1878-0261.12308.

Herr, W., Sturm, R.A., Clerc, R.G., Corcoran, L.M., Baltimore, D., Sharp, P.A., Ingraham, H.A., Rosenfeld, M.G., Finney, M., and Ruvkun, G. (1988). The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and Caenorhabditis elegans unc-86 gene products. Genes Dev. *2*, 1513–1516. https://doi.org/10.1101/gad.2.12a.1513.

Hershey, C.L., and Fisher, D.E. (2005). Genomic analysis of the Microphthalmia locus and identification of the MITF-J/Mitf-J isoform. Gene *347*, 73–82. https://doi.org/10.1016/j.gene.2004.12.002.

Herynk, M.H., and Fuqua, S.A.W. (2004). Estrogen Receptor Mutations in Human Disease. Endocr. Rev. *25*, 869–898. https://doi.org/10.1210/er.2003-0010.

Herz, H.-M., Garruss, A., and Shilatifard, A. (2013). SET for life: biochemical activities and biological functions of SET domain-containing proteins. Trends Biochem. Sci. *38*, 621–639. https://doi.org/10.1016/j.tibs.2013.09.004.

Hisaoka, M., Nagata, K., and Okuwaki, M. (2014). Intrinsically disordered regions of nucleophosmin/B23 regulate its RNA binding activity through their inter- and intra-molecular association. Nucleic Acids Res. *42*, 1180–1195. https://doi.org/10.1093/nar/gkt897.

Hollenberg, S.M., Weinberger, C., Ong, E.S., Cerelli, G., Oro, A., Lebo, R., Brad Thompson, E., Rosenfeld, M.G., and Evans, R.M. (1985). Primary structure and expression of a functional human glucocorticoid receptor cDNA. Nature *318*, 635–641. https://doi.org/10.1038/318635a0.

Hong, Y.S., Kim, S.Y., Bhattacharya, A., Pratt, D.R., Hong, W.K., and Tainsky, M.A. (1995). Structure and function of the HOX A1 human homeobox gene cDNA. Gene *159*, 209–214. https://doi.org/10.1016/0378-1119(95)92712-G.

Horton, J.D. (2002). Sterol regulatory element-binding proteins: transcriptional activators of lipid synthesis. Biochem. Soc. Trans. *30*, 1091–1095. https://doi.org/10.1042/bst0301091.

Houlard, M., Romero-Portillo, F., Germani, A., Depaux, A., Regnier-Ricard, F., Gisselbrecht, S., and Varin-Blank, N. (2005). Characterization of VIK-1: a new Vav-interacting Kruppel-like protein. Oncogene *24*, 28–38. https://doi.org/10.1038/sj.onc.1208043.

Hovanes, K., Li, T.W.H., and Waterman, M.L. (2000). The human LEF-1 gene contains a promoter preferentially active in lymphocytes and encodes multiple isoforms derived from alternative splicing. Nucleic Acids Res. *28*, 1994–2003. https://doi.org/10.1093/nar/28.9.1994.

Hua, X., Wu, J., Goldstein, J.L., Brown, M.S., and Hobbs, H.H. (1995). Structure of the human gene encoding sterol regulatory element binding protein-1 (SREBF1) and localization of SREBF1 and SREBF2 to chromosomes 17p11.2 and 22q13. Genomics *25*, 667–673. https://doi.org/10.1016/0888-7543(95)80009-B.

Huber, R.M., Murphy, K., Miao, B., Link, J.R., Cunningham, M.R., Rupar, M.J., Gunyuzlu, P.L., Haws, T.F., Kassam, A., Powell, F., et al. (2002). Generation of multiple farnesoid-X-receptor isoforms through the use of alternative promoters. Gene *290*, 35–43. https://doi.org/10.1016/S0378-1119(02)00557-7.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. Nucleic Acids Res. *43*, D117–D122. https://doi.org/10.1093/nar/gku1045.

Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Mol. Cell *62*, 79–91. https://doi.org/10.1016/j.molcel.2016.03.001.

Izeddin, I., Récamier, V., Bosanac, L., Cissé, I.I., Boudarene, L., Dugast-Darzacq, C., Proux, F., Bénichou, O., Voituriez, R., Bensaude, O., et al. (2014). Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the nucleus. eLife *3*, e02230. https://doi.org/10.7554/eLife.02230.

Jain, P., Ballare, C., Blanco, E., Vizan, P., and Di Croce, L. (2020). PHF19 mediated regulation

of proliferation and invasiveness in prostate cancer cells. eLife *9*, e51373. https://doi.org/10.7554/eLife.51373.

Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., and Parcy, F. (2002). bZIP transcription factors in Arabidopsis. Trends Plant Sci. *7*, 106–111. https://doi.org/10.1016/S1360-1385(01)02223-3.

Janin, J., and Wodak, S.J. (1983). Structural domains in proteins and their role in the dynamics of protein function. Prog. Biophys. Mol. Biol. *42*, 21–78. https://doi.org/10.1016/0079-6107(83)90003-2.

Jia, L., Mao, Y., Ji, Q., Dersh, D., Yewdell, J.W., and Qian, S.-B. (2020). Decoding mRNA translatability and stability from the 5′ UTR. Nat. Struct. Mol. Biol. *27*, 814–821. https://doi.org/10.1038/s41594-020-0465-x.

Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A.K., Marrocco, J., Williams, S.R., Haase, B., Hayes, A., et al. (2021). A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat. Commun. *12*, 463. https://doi.org/10.1038/s41467-020-20343-5.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502. https://doi.org/10.1126/science.1141319.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873. https://doi.org/10.1101/gr.100552.109.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. Cell *152*, 327–339. https://doi.org/10.1016/j.cell.2012.12.009.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388. https://doi.org/10.1038/nature15518.

Jones, S. (2004). An overview of the basic helix-loop-helix proteins. Genome Biol. *5*, 226. https://doi.org/10.1186/gb-2004-5-6-226.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

https://doi.org/10.1093/bioinformatics/btu031.

Joseph, S.R., Pálfy, M., Hilbert, L., Kumar, M., Karschau, J., Zaburdaev, V., Shevchenko, A., and Vastenhouw, N.L. (2017). Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. eLife *6*, e23326. https://doi.org/10.7554/eLife.23326.

Kanezaki, R., Toki, T., Yokoyama, M., Yomogida, K., Sugiyama, K., Yamamoto, M., Igarashi, K., and Ito, E. (2001). Transcription Factor BACH1 Is Recruited to the Nucleus by Its Novel Alternative Spliced Isoform. J. Biol. Chem. *276*, 7278–7284. https://doi.org/10.1074/jbc.M004227200.

Karpova, A.Y., Ronco, L.V., and Howley, P.M. (2001). Functional Characterization of Interferon Regulatory Factor 3a (IRF-3a), an Alternative Splice Isoform of IRF-3. Mol. Cell. Biol. *21*, 4169–4176. https://doi.org/10.1128/MCB.21.13.4169-4176.2001.

Kastner, P., Krust, A., Turcotte, B., Stropp, U., Tora, L., Gronemeyer, H., and Chambon, P. (1990). Two distinct estrogen-regulated promoters generate transcripts encoding the two functionally different human progesterone receptor forms A and B. EMBO J. *9*, 1603–1614. https://doi.org/10.1002/j.1460-2075.1990.tb08280.x.

Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K., and Henikoff, S. (2020). Efficient low-cost chromatin profiling with CUT&Tag. Nat. Protoc. *15*, 3264–3283. https://doi.org/10.1038/s41596-020-0373-x.

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. *10*, 1–10. https://doi.org/10.1038/s41467-019-09982-5.

Kennett, S.B., Udvadia, A.J., and Horowitz, J.M. (1997). Sp3 encodes multiple proteins that differ in their capacity to stimulate or repress transcription. Nucleic Acids Res. *25*, 3110–3117. https://doi.org/10.1093/nar/25.15.3110.

Kirsanov, D.D., Zanegina, O.N., Aksianov, E.A., Spirin, S.A., Karyagina, A.S., and Alexeevski, A.V. (2013). NPIDB: nucleic acid—protein interaction database. Nucleic Acids Res. *41*, D517–D523. https://doi.org/10.1093/nar/gks1199.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. *20*, 207–220. https://doi.org/10.1038/s41576-018-0089-8.

Klug, C.A., Morrison, S.J., Masek, M., Hahm, K., Smale, S.T., and Weissman, I.L. (1998). Hematopoietic stem cells and lymphoid progenitors express different Ikaros isoforms,

and Ikaros is localized to heterochromatin in immature lymphocytes. Proc. Natl. Acad. Sci. *95*, 657–662. https://doi.org/10.1073/pnas.95.2.657.

Koepfli, K.-P., Paten, B., and and O'Brien, S. J. (2015). The Genome 10K Project: A Way Forward. Annu. Rev. Anim. Biosci. *3*, 57–111. https://doi.org/10.1146/annurev-animal-090414-014900.

Komeno, Y., Yan, M., Matsuura, S., Lam, K., Lo, M.-C., Huang, Y.-J., Tenen, D.G., Downing, J.R., and Zhang, D.-E. (2014). Runx1 exon 6–related alternative splicing isoforms differentially regulate hematopoiesis in mice. Blood *123*, 3760–3769. https://doi.org/10.1182/blood-2013-08-521252.

Kozmik, Z., Czerny, T., and Busslinger, M. (1997). Alternatively spliced insertions in the paired domain restrict the DNA sequence specificity of Pax6 and Pax8. EMBO J. *16*, 6793–6803. https://doi.org/10.1093/emboj/16.22.6793.

Kristie, T.M., and Sharp, P.A. (1990). Interactions of the Oct-1 POU subdomains with specific DNA sequences and with the HSV alpha-trans-activator protein. Genes Dev. *4*, 2383–2396. https://doi.org/10.1101/gad.4.12b.2383.

Krueger, F., James, F., Ewels, P., Afyounian, E., and Schuster-Boeckler, B. (2021). TrimGalore. https://doi.org/10.5281/zenodo.5127899.

Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. Brief. Bioinform. *18*, 205–214. https://doi.org/10.1093/bib/bbw008.

Kuiper, R.P., Schepens, M., Thijssen, J., Schoenmakers, E.F.P.M., and van Kessel, A.G. (2004). Regulation of the MiTF/TFE bHLH-LZ transcription factors through restricted spatial expression and alternative splicing of functional domains. Nucleic Acids Res. *32*, 2315–2322. https://doi.org/10.1093/nar/gkh571.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259. https://doi.org/10.1093/nar/gkx1106.

Kumar, R., and Litwack, G. (2009). Structural and functional relationships of the steroid hormone receptors' N-terminal transactivation domain. Steroids *74*, 877–883. https://doi.org/10.1016/j.steroids.2009.07.012.

Kumar, B.D., Parker, H.J., Paulson, A., Parrish, M.E., Pushel, I., Singh, N.P., Zhang, Y., Slaughter, B.D., Unruh, J.R., Florens, L., et al. (2017). HOXA1 and TALE proteins

display cross-regulatory interactions and form a combinatorial binding code on HOXA1 targets. Genome Res. *27*, 1501–1512. https://doi.org/10.1101/gr.219386.116.

Lađinović, D., Pinkas, D., Šopin, T., Raška, O., Liška, F., Raška, I., and Vacík, T. (2020). Alternative isoforms of KDM2A and KDM2B lysine demethylases negatively regulate canonical Wnt signaling. PLOS ONE *15*, e0236612. https://doi.org/10.1371/journal.pone.0236612.

Lai, J.S., Cleary, M.A., and Herr, W. (1992). A single amino acid exchange transfers VP16-induced positive control from the Oct-1 to the Oct-2 homeo domain. Genes Dev. *6*, 2058–2065. https://doi.org/10.1101/gad.6.11.2058.

Laity, J.H., Lee, B.M., and Wright, P.E. (2001). Zinc finger proteins: new insights into structural and functional diversity. Curr. Opin. Struct. Biol. *11*, 39–46. https://doi.org/10.1016/S0959-440X(00)00167-6.

Lambert, É., Babeu, J.-P., Simoneau, J., Raisch, J., Lavergne, L., Lévesque, D., Jolibois, É., Avino, M., Scott, M.S., Boudreau, F., et al. (2020). Human Hepatocyte Nuclear Factor 4-α Encodes Isoforms with Distinct Transcriptional Functions. Mol. Cell. Proteomics *19*, 808–827. https://doi.org/10.1074/mcp.RA119.001909.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029.

Lange, M., Kaynak, B., Forster, U.B., Tönjes, M., Fischer, J.J., Grimm, C., Schlesinger, J., Just, S., Dunkel, I., Krueger, T., et al. (2008). Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. Genes Dev. *22*, 2370–2384. https://doi.org/10.1101/gad.471408.

Langmann, T., Schumacher, C., Morham, S.G., Honer, C., Heimerl, S., Moehle, C., and Schmitz, G. (2003). ZNF202 is inversely regulated with its target genes ABCA1 and apoE during macrophage differentiation and foam cell formation. J. Lipid Res. *44*, 968–977. https://doi.org/10.1194/jlr.M300016-JLR200.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Laoide, B.M., Foulkes, N.S., Schlotter, F., and Sassone-Corsi, P. (1993). The functional versatility of CREM is determined by its modular structure. EMBO J. *12*, 1179–1191. https://doi.org/10.1002/j.1460-2075.1993.tb05759.x.

Lavigne, P., Crump, M.P., Gagné, S.M., Hodges, R.S., Kay, C.M., and Sykes, B.D. (1998). Insights into the mechanism of heterodimerization from the 1H-NMR solution structure

of the c-Myc-Max heterodimeric leucine zipper. J. Mol. Biol. *281*, 165–181. https://doi.org/10.1006/jmbi.1998.1914.

Lazzari, E., Korczeniewska, J., Gabhann, J.N., Smith, S., Barnes, B.J., and Jefferies, C.A. (2014). TRIpartite Motif 21 (TRIM21) Differentially Regulates the Stability of Interferon Regulatory Factor 5 (IRF5) Isoforms. PLOS ONE *9*, e103609. https://doi.org/10.1371/journal.pone.0103609.

Lee, H., and Schatz, M.C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics *28*, 2097–2105. https://doi.org/10.1093/bioinformatics/bts330.

Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu. Rev. Biochem. *84*, 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316.

Lee, E.J., Kim, J.M., Lee, M.K., and Jameson, J.L. (2008). Splice Variants of the Forkhead Box Protein AFX Exhibit Dominant Negative Activity and Inhibit AFXα-Mediated Tumor Cell Apoptosis. PLOS ONE *3*, e2743. https://doi.org/10.1371/journal.pone.0002743.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. *114*, 6589–6631. https://doi.org/10.1021/cr400525m.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233–245. https://doi.org/10.1038/nrg3163.

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. *46*, D493–D496. https://doi.org/10.1093/nar/gkx922.

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. *49*, D458–D460. https://doi.org/10.1093/nar/gkaa937.

Leung, S.K., Jeffries, A.R., Castanho, I., Jordan, B.T., Moore, K., Davies, J.P., Dempster, E.L., Bray, N.J., O'Neill, P., Tseng, E., et al. (2021). Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. Cell Rep. *37*, 110022. https://doi.org/10.1016/j.celrep.2021.110022.

Lewis, T.E., Sillitoe, I., Dawson, N., Lam, S.D., Clarke, T., Lee, D., Orengo, C., and Lees, J.

(2018). Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Res. *46*, D435–D439. https://doi.org/10.1093/nar/gkx1069.

Li, C., and Luscombe, N.M. (2020). Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. Nat. Commun. *11*, 1363. https://doi.org/10.1038/s41467-020-15185-0.

Li, C., Ma, L., and Chen, X. (2011). Interferon regulatory factor 3-CL, an isoform of IRF3, antagonizes activity of IRF3. Cell. Mol. Immunol. *8*, 67–74. https://doi.org/10.1038/cmi.2010.55.

Li, J., Huang, K., Hu, G., Babarinde, I.A., Li, Y., Dong, X., Chen, Y.-S., Shang, L., Guo, W., Wang, J., et al. (2019). An alternative CTCF isoform antagonizes canonical CTCF occupancy and changes chromatin architecture to promote apoptosis. Nat. Commun. *10*, 1535. https://doi.org/10.1038/s41467-019-08949-w.

Li, R., Zhang, H., Yu, W., Chen, Y., Gui, B., Liang, J., Wang, Y., Sun, L., Yang, X., Zhang, Y., et al. (2009). ZIP: a novel transcription repressor, represses EGFR oncogene and suppresses breast carcinogenesis. EMBO J. *28*, 2763–2776. https://doi.org/10.1038/emboj.2009.211.

Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M.K., Durkin, A.S., et al. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res. *49*, D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.

Li, X.Y., Hooft van Huijsduijnen, R., Mantovani, R., Benoist, C., and Mathis, D. (1992). Intron-exon organization of the NF-Y genes. Tissue-specific splicing modifies an activation domain. J. Biol. Chem. *267*, 8984–8990. https://doi.org/10.1016/S0021-9258(19)50377-5.

Li, Z., Wang, D., Na, X., Schoen, S.R., Messing, E.M., and Wu, G. (2003). The VHL protein recruits a novel KRAB-A domain protein to repress HIF-1α transcriptional activity. EMBO J. *22*, 1857–1867. https://doi.org/10.1093/emboj/cdg173.

Li, Z., Li, S., Luo, M., Jhong, J.-H., Li, W., Yao, L., Pang, Y., Wang, Z., Wang, R., Ma, R., et al. (2022). dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res. *50*, D471–D479. https://doi.org/10.1093/nar/gkab1017.

Lillycrop, K.A., and Latchman, D.S. (1992). Alternative splicing of the Oct-2 transcription factor RNA is differentially regulated in neuronal cells and B cells and results in protein isoforms with opposite effects on the activity of octamer/TAATGARAT-containing

promoters. J. Biol. Chem. *267*, 24960–24965. https://doi.org/10.1016/S0021-9258(19)73991-X.

Lillycrop, K.A., Dawson, S.J., Estridge, J.K., Gerster, T., Matthias, P., and Latchman, D.S. (1994). Repression of a herpes simplex virus immediate-early promoter by the Oct-2 transcription factor is dependent on an inhibitory region at the N terminus of the protein. Mol. Cell. Biol. *14*, 7633–7642. https://doi.org/10.1128/mcb.14.11.7633-7642.1994.

Linares, A.J., Lin, C.-H., Damianov, A., Adams, K.L., Novitch, B.G., and Black, D.L. (2015). The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. eLife *4*, e09268. https://doi.org/10.7554/eLife.09268.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006a). Intrinsic Disorder in Transcription Factors. Biochemistry *45*, 6873–6888. https://doi.org/10.1021/bi0602718.

Liu, X., Lee, C.-K., Granek, J.A., Clarke, N.D., and Lieb, J.D. (2006b). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. Genome Res. *16*, 1517–1528. https://doi.org/10.1101/gr.5655606.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22. https://doi.org/10.1186/s13059-014-0560-6.

Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C.C., de Hoon, M., Severin, J., Oki, S., Hayashizaki, Y., et al. (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Res. *47*, D752–D758. https://doi.org/10.1093/nar/gky1099.

Lluís, F., Ballestar, E., Suelves, M., Esteller, M., and Muñoz-Cánoves, P. (2005). E47 phosphorylation by p38 MAPK promotes MyoD/E47 association and muscle-specific gene transcription. EMBO J. *24*, 974–984. https://doi.org/10.1038/sj.emboj.7600528.

Lone, I.N., Shukla, M.S., Richard, J.L.C., Peshev, Z.Y., Dimitrov, S., and Angelov, D. (2013). Binding of NF-κB to Nucleosomes: Effect of Translational Positioning, Nucleosome Remodeling and Linker Histone H1. PLOS Genet. *9*, e1003830. https://doi.org/10.1371/journal.pgen.1003830.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585. https://doi.org/10.1038/ng.2653.

López, A.J. (1995). Developmental Role of Transcription Factor Isoforms Generated by Alternative Splicing. Dev. Biol. *172*, 396–411. https://doi.org/10.1006/dbio.1995.8050.

Lu, N.Z., and Cidlowski, J.A. (2005). Translational Regulatory Mechanisms Generate N-Terminal Glucocorticoid Receptor Isoforms with Unique Transcriptional Target Genes. Mol. Cell *18*, 331–342. https://doi.org/10.1016/j.molcel.2005.03.025.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. *48*, D265–D268. https://doi.org/10.1093/nar/gkz991.

Lucena, P.I., Faget, D.V., Pachulec, E., Robaina, M.C., Klumb, C.E., Robbs, B.K., and Viola, J.P.B. (2016). NFAT2 Isoforms Differentially Regulate Gene Expression, Cell Death, and Transformation through Alternative N-Terminal Domains. Mol. Cell. Biol. *36*, 119–131. https://doi.org/10.1128/MCB.00501-15.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *389*, 251–260. https://doi.org/10.1038/38444.

Lüscher, B., and Vervoorts, J. (2012). Regulation of gene transcription by the oncoprotein MYC. Gene *494*, 145–160. https://doi.org/10.1016/j.gene.2011.12.027.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874. https://doi.org/10.1093/nar/29.13.2860.

Maiti, S., Acharya, B., Boorla, V.S., Manna, B., Ghosh, A., and De, S. (2019). Dynamic Studies on Intrinsically Disordered Regions of Two Paralogous Transcription Factors Reveal Rigid Segments with Important Biological Functions. J. Mol. Biol. *431*, 1353–1369. https://doi.org/10.1016/j.jmb.2019.02.021.

Mangal, R.K., Wiehle, R.D., Poindexter, A.N., and Weigel, N.L. (1997). Differential expression of uterine progesterone receptor forms A and B during the menstrual cycle. J. Steroid Biochem. Mol. Biol. *63*, 195–202. https://doi.org/10.1016/S0960-0760(97)00119-2.

Marasco, L. E., and Kornblihtt, A. R. (2022). The physiology of alternative splicing. Nat. Rev. Mol. Cell. Biol., 1–13. https://doi.org/10.1038/s41580-022-00545-z.

Maritano, D., Sugrue, M.L., Tininini, S., Dewilde, S., Strobl, B., Fu, X., Murray-Tait, V., Chiarle, R., and Poli, V. (2004). The STAT3 isoforms α and β have unique and specific functions. Nat. Immunol. *5*, 401–409. https://doi.org/10.1038/ni1052.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

Marti-Solano, M., Crilly, S.E., Malinverni, D., Munk, C., Harris, M., Pearce, A., Quon, T., Mackenzie, A.E., Wang, X., Peng, J., et al. (2020). Combinatorial expression of GPCR isoforms affects signalling and drug responses. Nature *587*, 650–656. https://doi.org/10.1038/s41586-020-2888-2.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108–D110. https://doi.org/10.1093/nar/gkj143.

Mayr, C. (2017). Regulation by 3′-Untranslated Regions. Annu. Rev. Genet. *51*, 171–194. https://doi.org/10.1146/annurev-genet-120116-024704.

McDermott, J.C., Cardoso, M.C., Yu, Y.T., Andres, V., Leifer, D., Krainc, D., Lipton, S.A., and Nadal-Ginard, B. (1993). hMEF2C gene encodes skeletal muscle- and brain-specific transcription factors. Mol. Cell. Biol. *13*, 2564–2577. https://doi.org/10.1128/mcb.13.4.2564-2577.1993.

Meers, M.P., Janssens, D.H., and Henikoff, S. (2019). Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. Mol. Cell *75*, P562-575.E5. https://doi.org/10.1016/j.molcel.2019.05.025.

Meyer, M.E., Quirin-Stricker, C., Lerouge, T., Bocquel, M.T., and Gronemeyer, H. (1992). A limiting factor mediates the differential activation of promoters by the human progesterone receptor isoforms. J. Biol. Chem. *267*, 10882–10887. https://doi.org/10.1016/S0021-9258(19)50100-4.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. *45*, D183–D189. https://doi.org/10.1093/nar/gkw1138.

Michael, A.K., Grand, R.S., Isbel, L., Cavadini, S., Kozicka, Z., Kempf, G., Bunker, R.D., Schenk, A.D., Graff-Meyer, A., Pathare, G.R., et al. (2020). Mechanisms of OCT4-SOX2 motif readout on nucleosomes. Science *368*, 1460–1465. https://doi.org/10.1126/science.abb0074.

Mignon, J., Mottet, D., Leyder, T., Uversky, V.N., Perpète, E.A., and Michaux, C. (2022). Structural characterisation of amyloidogenic intrinsically disordered zinc finger protein isoforms DPF3b and DPF3a. Int. J. Biol. Macromol. *218*, 57–71.

https://doi.org/10.1016/j.ijbiomac.2022.07.102.

Milech, N., Kees, U.R., and Watt, P.M. (2001). Novel alternative PBX3 isoforms in leukemia cells with distinct interaction specificities. Genes Chromosomes Cancer *32*, 275–280. https://doi.org/10.1002/gcc.1190.

Miller, J.A., and Widom, J. (2003). Collaborative Competition Mechanism for Gene Activation In Vivo. Mol. Cell. Biol. *23*, 1623–1632. https://doi.org/10.1128/MCB.23.5.1623-1632.2003.

Miller, M., Shuman, J.D., Sebastian, T., Dauter, Z., and Johnson, P.F. (2003). Structural Basis for DNA Recognition by the Basic Region Leucine Zipper Transcription Factor CCAAT/Enhancer-binding Protein α. J. Biol. Chem. *278*, 15178–15184. https://doi.org/10.1074/jbc.M300417200.

Milne, T.A., Briggs, S.D., Brock, H.W., Martin, M.E., Gibbs, D., Allis, C.D., and Hess, J.L. (2002). MLL Targets SET Domain Methyltransferase Activity to Hox Gene Promoters. Mol. Cell *10*, 1107–1117. https://doi.org/10.1016/S1097-2765(02)00741-4.

Minezaki, Y., Homma, K., Kinjo, A.R., and Nishikawa, K. (2006). Human Transcription Factors Contain a High Fraction of Intrinsically Disordered Regions Essential for Transcriptional Regulation. J. Mol. Biol. *359*, 1137–1149. https://doi.org/10.1016/j.jmb.2006.04.016.

Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc. Natl. Acad. Sci. *107*, 22534–22539. https://doi.org/10.1073/pnas.0913805107.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

Molina, C.A., Foulkes, N.S., Lalli, E., and Sassone-Corsi, P. (1993). Inducibility and negative autoregulation of CREM: An alternative promoter directs the expression of ICER, an early response repressor. Cell *75*, 875–886. https://doi.org/10.1016/0092-8674(93)90532-U.

Molnár, A., Wu, P., Largespada, D.A., Vortkamp, A., Scherer, S., Copeland, N.G., Jenkins, N.A., Bruns, G., and Georgopoulos, K. (1996). The Ikaros gene encodes a family of lymphocyte-restricted zinc finger DNA binding proteins, highly conserved in human and mouse. J. Immunol. *156*, 585–592.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription

factors with DNA microarrays. Nat. Genet. *36*, 1331–1339. https://doi.org/10.1038/ng1473.

Murphy, P., and Hill, R.E. (1991). Expression of the mouse labial-like homeobox-containing genes, Hox 2.9 and Hox 1.6, during segmentation of the hindbrain. Development *111*, 61–74. https://doi.org/10.1242/dev.111.1.61.

Nair, S.K., and Burley, S.K. (2003). X-Ray Structures of Myc-Max and Mad-Max Recognizing DNA: Molecular Bases of Regulation by Proto-Oncogenic Transcription Factors. Cell *112*, 193–205. https://doi.org/10.1016/S0092-8674(02)01284-9.

Nair, R., Carter, P., and Rost, B. (2003). NLSdb: database of nuclear localization signals. Nucleic Acids Res. *31*, 397–399. https://doi.org/10.1093/nar/gkg001.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562. https://doi.org/10.1038/nbt.3128.

Narla, G., DiFeo, A., Reeves, H.L., Schaid, D.J., Hirshfeld, J., Hod, E., Katz, A., Isaacs, W.B., Hebbring, S., Komiya, A., et al. (2005a). A Germline DNA Polymorphism Enhances Alternative Splicing of the KLF6 Tumor Suppressor Gene and Is Associated with Increased Prostate Cancer Risk. Cancer Res. *65*, 1213–1222. https://doi.org/10.1158/0008-5472.CAN-04-4249.

Narla, G., DiFeo, A., Yao, S., Banno, A., Hod, E., Reeves, H.L., Qiao, R.F., Camacho-Vanegas, O., Levine, A., Kirschenbaum, A., et al. (2005b). Targeted Inhibition of the KLF6 Splice Variant, KLF6 SV1, Suppresses Prostate Cancer Cell Growth and Spread. Cancer Res. *65*, 5761–5768. https://doi.org/10.1158/0008-5472.CAN-05-0217.

Neve, J., Patel, R., Wang, Z., Louey, A., and Furger, A. M. (2017). Cleavage and polyadenylation: Ending the message expands gene regulation. RNA Biol. *14*, 865–890. https://doi.org/10.1080/15476286.2017.1306171.

Ng, I.H.W., Ng, D.C.H., Jans, D.A., and Bogoyevitch, M.A. (2012). Selective STAT3-α or -β expression reveals spliceform-specific phosphorylation kinetics, nuclear retention and distinct gene expression outcomes. Biochem. J. *447*, 125–136. https://doi.org/10.1042/BJ20120941.

Nishiyama, C., Takahashi, K., Nishiyama, M., Okumura, K., Ra, C., Ohtake, Y., and Yokota, T. (2000). Splice Isoforms of Transcription Factor Elf-1 Affecting Its Regulatory Function in Transcription—Molecular Cloning of Rat Elf-1. Biosci. Biotechnol. Biochem. *64*, 2601–2607. https://doi.org/10.1271/bbb.64.2601.

Noman, A., Liu, Z., Aqeel, M., Zainab, M., Khan, M.I., Hussain, A., Ashraf, M.F., Li, X., Weng, Y., and He, S. (2017). Basic leucine zipper domain transcription factors: the vanguards in plant immunity. Biotechnol. Lett. *39*, 1779–1791. https://doi.org/10.1007/s10529-017-2431-1.

Noro, B., Culi, J., McKay, D.J., Zhang, W., and Mann, R.S. (2006). Distinct functions of homeodomain-containing and homeodomain-less isoforms encoded by homothorax. Genes Dev. *20*, 1636–1650. https://doi.org/10.1101/gad.1412606.

Norris, R.A., Scott, K.K., Moore, C.S., Stetten, G., Brown, C.R., Jabs, E.W., Wulfsberg, E.A., Yu, J., and Kern, M.J. (2000). Human PRRX1 and PRRX2 genes: cloning, expression, genomic localization, and exclusion as disease genes for Nager syndrome. Mamm. Genome *11*, 1000–1005. https://doi.org/10.1007/s003350010193.

North, J. A., Shimko, J. C., Javaid, S., Mooney, A. M., Shoffner, M. A., Rose, S. D., Bundschuh, R., Fishel, R., Ottesen, J. J., and Poirier, M. G. (2012). Regulation of the nucleosome unwrapping rate controls DNA accessibility. Nucleic Acids Res. *40*, 10215–10227. https://doi.org/10.1093/nar/gks747.

Novakovsky, G., Saraswat, M., Fornes, O., Mostafavi, S., and Wasserman, W. W. (2021). Biologically relevant transfer learning improves transcription factor binding prediction. Genome Biol. *22*, 280. https://doi.org/10.1186/s13059-021-02499-5.

Oakley, R.H., Sar, M., and Cidlowski, J.A. (1996). The Human Glucocorticoid Receptor β Isoform: EXPRESSION, BIOCHEMICAL PROPERTIES, AND PUTATIVE FUNCTION. J. Biol. Chem. *271*, 9550–9559. https://doi.org/10.1074/jbc.271.16.9550.

Oakley, R.H., Webster, J.C., Sar, M., Parker, C.R., Jr., and Cidlowski, J.A. (1997). Expression and Subcellular Distribution of the β-Isoform of the Human Glucocorticoid Receptor. Endocrinology *138*, 5028–5038. https://doi.org/10.1210/endo.138.11.5501.

Oakley, R.H., Jewell, C.M., Yudt, M.R., Bofetiado, D.M., and Cidlowski, J.A. (1999). The Dominant Negative Activity of the Human Glucocorticoid Receptor β Isoform: SPECIFICITY AND MECHANISMS OF ACTION. J. Biol. Chem. *274*, 27857–27866. https://doi.org/10.1074/jbc.274.39.27857.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH – a hierarchic classification of protein domain structures. Structure *5*,

1093–1109. https://doi.org/10.1016/S0969-2126(97)00260-8.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415. https://doi.org/10.1038/ng.259.

Papaioannou, V.E. (2014). The T-box gene family: emerging roles in development, stem cells and cancer. Development *141*, 3819–3833. https://doi.org/10.1242/dev.104471.

Pardee, K., Necakov, A.S., and Krause, H. (2011). Nuclear Receptors: Small Molecule Sensors that Coordinate Growth, Metabolism and Reproduction. In A Handbook of Transcription Factors, T.R. Hughes, ed. (Dordrecht: Springer Netherlands), pp. 123–153.

Park, O.K., Schaefer, L.K., Wang, W., and Schaefer, T.S. (2000). Dimer Stability as a Determinant of Differential DNA Binding Activity of Stat3 Isoforms. J. Biol. Chem. *275*, 32244–32249. https://doi.org/10.1074/jbc.M005082200.

Parrado, A., Despouy, G., Kraïba, R., Pogam, C.L., Dupas, S., Choquette, M., Robledo, M., Larghero, J., Bui, H., Gall, I.L., et al. (2001). Retinoic acid receptor α1 variants, RARα1ΔB and RARα1ΔBC, define a new class of nuclear receptor isoforms. Nucleic Acids Res. *29*, 4901–4908. https://doi.org/10.1093/nar/29.24.4901.

Payne, K.J., Nicolas, J.-H., Zhu, J.Y., Barsky, L.W., and Crooks, G.M. (2001). Cutting Edge: Predominant Expression of a Novel Ikaros Isoform in Normal Human Hemopoiesis. J. Immunol. *167*, 1867–1870. https://doi.org/10.4049/jimmunol.167.4.1867.

Pedruzzi, I., Rivoire, C., Auchincloss, A.H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cuche, B.A., Bougueleret, L., Poux, S., et al. (2013). HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. *41*, D584–D589. https://doi.org/10.1093/nar/gks1157.

Pedruzzi, I., Rivoire, C., Auchincloss, A.H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cuche, B.A., Bougueleret, L., Poux, S., et al. (2015). HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Res. *43*, D1064–D1070. https://doi.org/10.1093/nar/gku1002.

Petitjean, A., Cavard, C., Shi, H., Tribollet, V., Hainaut, P., and Caron de Fromentel, C. (2005). The expression of TA and ΔNp63 are regulated by different mechanisms in liver cells. Oncogene *24*, 512–519. https://doi.org/10.1038/sj.onc.1208215.

Petitjean, A., Ruptier, C., Tribollet, V., Hautefeuille, A., Chardon, F., Cavard, C., Puisieux, A., Hainaut, P., and Caron de Fromentel, C. (2008). Properties of the six isoforms of p63: p53-like regulation in response to genotoxic stress and cross talk with ΔNp73.

Carcinogenesis *29*, 273–281. https://doi.org/10.1093/carcin/bgm258.

Pfurr, S., Chu, Y.-H., Bohrer, C., Greulich, F., Beattie, R., Mammadzada, K., Hils, M., Arnold, S.J., Taylor, V., Schachtrup, K., et al. (2017). The E2A splice variant E47 regulates the differentiation of projection neurons via p57(KIP2) during cortical development. Development *144*, 3917–3931. https://doi.org/10.1242/dev.145698.

Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010). Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. PLOS Genet. *6*, e1001236. https://doi.org/10.1371/journal.pgen.1001236.

Pierreux, C.E., Nicolás, F.J., and Hill, C.S. (2000). Transforming Growth Factor β-Independent Shuttling of Smad4 between the Cytoplasm and Nucleus. Mol. Cell. Biol. *20*, 9041–9054. https://doi.org/10.1128/MCB.20.23.9041-9054.2000.

Piggin, C.L., Roden, D.L., Gallego-Ortega, D., Lee, H.J., Oakes, S.R., and Ormandy, C.J. (2016). ELF5 isoform expression is tissue-specific and significantly altered in cancer. Breast Cancer Res. *18*, 4. https://doi.org/10.1186/s13058-015-0666-0.

Pinson, J., Mason, J.O., Simpson, T.I., and Price, D.J. (2005). Regulation of the Pax6 : Pax6(5a) mRNA ratio in the developing mammalian brain. BMC Dev. Biol. *5*, 13. https://doi.org/10.1186/1471-213X-5-13.

Pirkkala, L., Nykänen, P., and Sistonen, L. (2001). Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. FASEB J. *15*, 1118–1131. https://doi.org/10.1096/fj00-0294rev.

Pockrandt, C., Alzamel, M., Iliopoulos, C.S., and Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. Bioinformatics *36*, 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222.

Porsch-Özcürümez, M., Langmann, T., Heimerl, S., Borsukova, H., Kaminski, W.E., Drobnik, W., Honer, C., Schumacher, C., and Schmitz, G. (2001). The Zinc Finger Protein 202 (ZNF202) Is a Transcriptional Repressor of ATP Binding Cassette Transporter A1 (ABCA1) and ABCG1 Gene Expression and a Modulator of Cellular Lipid Efflux. J. Biol. Chem. *276*, 12427–12433. https://doi.org/10.1074/jbc.M100218200.

Potthoff, M.J., Arnold, M.A., McAnally, J., Richardson, J.A., Bassel-Duby, R., and Olson, E.N. (2007). Regulation of Skeletal Muscle Sarcomere Integrity and Postnatal Muscle Function by Mef2c. Mol. Cell. Biol. *27*, 8143–8151. https://doi.org/10.1128/MCB.01187-07.

Pradeepa, M.M., Sutherland, H.G., Ule, J., Grimes, G.R., and Bickmore, W.A. (2012). Psip1/Ledgf p52 Binds Methylated Histone H3K36 and Splicing Factors and

Contributes to the Regulation of Alternative Splicing. PLOS Genet. *8*, e1002717. https://doi.org/10.1371/journal.pgen.1002717.

Pugacheva, E.M., Kubo, N., Loukinov, D., Tajmul, M., Kang, S., Kovalchuk, A.L., Strunnikov, A.V., Zentner, G.E., Ren, B., and Lobanenkov, V.V. (2020). CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. Proc. Natl. Acad. Sci. *117*, 2020–2031. https://doi.org/10.1073/pnas.1911708117.

Pujols, L., Mullol, J., Roca-Ferrer, J., Torrego, A., Xaubet, A., Cidlowski, J.A., and Picado, C. (2002). Expression of glucocorticoid receptor α- and β-isoforms in human cells and tissues. Am. J. Physiol.-Cell Physiol. *283*, C1324–C1331. https://doi.org/10.1152/ajpcell.00363.2001.

Qi, Y., Yu, J., Han, W., Fan, X., Qian, H., Wei, H., Tsai, Y.S., Zhao, J., Zhang, W., Liu, Q., et al. (2016). A splicing isoform of TEAD4 attenuates the Hippo–YAP signalling to inhibit tumour proliferation. Nat. Commun. *7*, 1–12. https://doi.org/10.1038/ncomms11840.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team (2020). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Radhakrishnan, I., Pérez-Alvarado, G.C., Parker, D., Dyson, H.J., Montminy, M.R., and Wright, P.E. (1997). Solution Structure of the KIX Domain of CBP Bound to the Transactivation Domain of CREB: A Model for Activator:Coactivator Interactions. Cell *91*, 741–752. https://doi.org/10.1016/S0092-8674(00)80463-8.

Raj, B., O'Hanlon, D., Vessey, J.P., Pan, Q., Ray, D., Buckley, N.J., Miller, F.D., and Blencowe, B.J. (2011). Cross-Regulation between an Alternative Splicing Activator and a Transcription Repressor Controls Neurogenesis. Mol. Cell *43*, 843–850. https://doi.org/10.1016/j.molcel.2011.08.014.

Ram, D., Ziv, E., Lantner, F., Lardans, V., and Schechter, I. (2004). Stage-specific alternative splicing of the heat-shock transcription factor during the life-cycle of Schistosoma mansoni. Parasitology *129*, 587–596. https://doi.org/10.1017/S003118200400602X.

Ramachandran, S., and Henikoff, S. (2016). Transcriptional Regulators Compete with Nucleosomes Post-replication. Cell *165*, 580–592. https://doi.org/10.1016/j.cell.2016.02.062.

Reixachs-Solé, M., and Eyras, E. (2022). Uncovering the impacts of alternative splicing on the

proteome with current omics techniques. WIREs RNA *13*, e1707. https://doi.org/10.1002/wrna.1707.

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., and Lee, C. (2004). Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome. J. Proteome Res. *3*, 76–83. https://doi.org/10.1021/pr034064v.

Reyes, A., and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res. *46*, 582–592. https://doi.org/10.1093/nar/gkx1165.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell *147*, 1408–1419. https://doi.org/10.1016/j.cell.2011.11.013.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746. https://doi.org/10.1038/s41586-021-03451-0.

Richter, E.A., and Hargreaves, M. (2013). Exercise, GLUT4, and Skeletal Muscle Glucose Uptake. Physiol. Rev. *93*, 993–1017. https://doi.org/10.1152/physrev.00038.2012.

Ridgeway, A.G., Wilton, S., and Skerjanc, I.S. (2000). Myocyte Enhancer Factor 2C and Myogenin Up-regulate Each Other's Expression and Induce the Development of Skeletal Muscle in P19 Cells. J. Biol. Chem. *275*, 41–46. https://doi.org/10.1074/jbc.275.1.41.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods *4*, 651–657. https://doi.org/10.1038/nmeth1068.

Robichaud, G.A., Nardini, M., Laflamme, M., Cuperlovic-Culf, M., and Ouellette, R.J. (2004). Human Pax-5 C-terminal Isoforms Possess Distinct Transactivation Properties and Are Differentially Modulated in Normal and Malignant B Cells. J. Biol. Chem. *279*, 49956–49963. https://doi.org/10.1074/jbc.M407171200.

Rocks, O., Peyker, A., Kahms, M., Verveer, P.J., Koerner, C., Lumbierres, M., Kuhlmann, J., Waldmann, H., Wittinghofer, A., and Bastiaens, P.I.H. (2005). An Acylation Cycle Regulates Localization and Activity of Palmitoylated Ras Isoforms. Science *307*, 1746–1752. https://doi.org/10.1126/science.1105654.

Rosa, A., and Brivanlou, A.H. (2011). A regulatory circuitry comprised of miR-302 and the

transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation. EMBO J. *30*, 237–248. https://doi.org/10.1038/emboj.2010.319.

Ross, S., Best, J.L., Zon, L.I., and Gill, G. (2002). SUMO-1 Modification Represses Sp3 Transcriptional Activation and Modulates Its Subnuclear Localization. Mol. Cell *10*, 831–842. https://doi.org/10.1016/S1097-2765(02)00682-2.

Sanborn, A.L., Yeh, B.T., Feigerle, J.T., Hao, C.V., Townshend, R.J., Lieberman Aiden, E., Dror, R.O., and Kornberg, R.D. (2021). Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. eLife *10*, e68068. https://doi.org/10.7554/eLife.68068.

Sapetschnig, A., Koch, F., Rischitor, G., Mennenga, T., and Suske, G. (2004). Complexity of Translationally Controlled Transcription Factor Sp3 Isoform Expression. J. Biol. Chem. *279*, 42095–42105. https://doi.org/10.1074/jbc.M404989200.

Scheepers, A., Schmidt, S., Manolescu, A., Cheeseman, C.I., Bell, A., Zahn, C., Joost, H.-G., and Schürmann, A. (2005). Characterization of the human SLC2A11 (GLUT11) gene: alternative promoter usage, function, expression, and subcellular distribution of three isoforms, and lack of mouse orthologue. Mol. Membr. Biol. *22*, 339–351. https://doi.org/10.1080/09687860500166143.

Schmitges, F.W., Radovani, E., Najafabadi, H.S., Barazandeh, M., Campitelli, L.F., Yin, Y., Jolma, A., Zhong, G., Guo, H., Kanagalingam, T., et al. (2016). Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res. *26*, 1742–1752. https://doi.org/10.1101/gr.209643.116.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Schneider-Gädicke, A., Beer-Romero, P., Brown, L.G., Mardon, G., Luoh, S.-W., and Page, D.C. (1989). Putative transcription activator with alternative isoforms encoded by human ZFX gene. Nature *342*, 708–711. https://doi.org/10.1038/342708a0.

Schoof, E.M., Furtwängler, B., Üresin, N., Rapin, N., Savickas, S., Gentil, C., Lechman, E., Keller, U. auf dem, Dick, J.E., and Porse, B.T. (2021). Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. Nat. Commun. *12*, 3341. https://doi.org/10.1038/s41467-021-23667-y.

Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koča, J., and Rose, A.S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. *49*, W431–W437. https://doi.org/10.1093/nar/gkab314.

Seo, P.J., Kim, M.J., Ryu, J.-Y., Jeong, E.-Y., and Park, C.-M. (2011). Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. Nat. Commun. *2*, 303. https://doi.org/10.1038/ncomms1303.

Seo, P.J., Park, M.-J., and Park, C.-M. (2013). Alternative splicing of transcription factors in plant responses to low temperature stress: mechanisms and functions. Planta *237*, 1415–1424. https://doi.org/10.1007/s00425-013-1882-4.

Sepp, M., Kannike, K., Eesmaa, A., Urb, M., and Timmusk, T. (2011). Functional Diversity of Human Basic Helix-Loop-Helix Transcription Factor TCF4 Isoforms Generated by Alternative 5′ Exon Usage and Splicing. PLOS ONE *6*, e22138. https://doi.org/10.1371/journal.pone.0022138.

Serfling, E., Avots, A., Klein-Hessling, S., Rudolf, R., Vaeth, M., and Berberich-Siebelt, F. (2012). NFATc1/αA: The other Face of NFAT Factors in Lymphocytes. Cell Commun. Signal. *10*, 16. https://doi.org/10.1186/1478-811X-10-16.

Shabalina, S.A., Ogurtsov, A.Y., Spiridonov, N.A., and Koonin, E.V. (2014). Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. Nucleic Acids Res. *42*, 7132–7144. https://doi.org/10.1093/nar/gku342.

Shao, H., Quintero, A.J., and Tweardy, D.J. (2001). Identification and characterization of cis elements in the STAT3 gene regulating STAT3α and STAT3β messenger RNA splicing. Blood *98*, 3853–3856. https://doi.org/10.1182/blood.V98.13.3853.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014. https://doi.org/10.1038/nbt.2705.

Shen, W., Detmer, K., Simonitch-Eason, T.A., Lawrence, H.J., and Largman, C. (1991). Alternative splicing of the HOX 2.2 homeobox gene in human hematopoietic cells and murine embryonic and adult tissues. Nucleic Acids Res. *19*, 539–545. https://doi.org/10.1093/nar/19.3.539.

Sherman, M.A., Powell, D.R., Weiss, D.L., and Brown, M.A. (1999). NF-ATc Isoforms Are Differentially Expressed and Regulated in Murine T and Mast Cells. J. Immunol. *162*, 2820–2828.

Shimamura, R., Fraizer, G.C., Trapman, J., Lau, Y.-f.C., and Saunders, G.F. (1997). The Wilms' tumor gene WT1 can regulate genes involved in sex determination and differentiation: SRY, Müllerian-inhibiting substance, and the androgen receptor. Clin.

Cancer Res. *3*, 2571–2580.

Shin, S.W., Kokoza, V., Ahmed, A., and Raikhel, A.S. (2002). Characterization of three alternatively spliced isoforms of the Rel/NF-κB transcription factor Relish from the mosquito Aedes aegypti. Proc. Natl. Acad. Sci. *99*, 9978–9983. https://doi.org/10.1073/pnas.162345999.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286. https://doi.org/10.1038/nrg3682.

Shoemaker, B.A., Portman, J.J., and Wolynes, P.G. (2000). Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. Proc. Natl. Acad. Sci. *97*, 8868–8873. https://doi.org/10.1073/pnas.160259697.

Shore, P., and Sharrocks, A.D. (1995). The MADS-Box Family of Transcription Factors. Eur. J. Biochem. *229*, 1–13. https://doi.org/10.1111/j.1432-1033.1995.0001l.x.

Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013). New and continuing developments at PROSITE. Nucleic Acids Res. *41*, D344–D347. https://doi.org/10.1093/nar/gks1067.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., et al. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. *49*, D266–D273. https://doi.org/10.1093/nar/gkaa1079.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife *6*, e21856. https://doi.org/10.7554/eLife.21856.

Skinner, M.K., Rawls, A., Wilson-Rawls, J., and Roalson, E.H. (2010). Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. Differentiation *80*, 1–8. https://doi.org/10.1016/j.diff.2010.02.003.

Sladek, F.M. (2011). What are nuclear receptor ligands? Mol. Cell. Endocrinol. *334*, 3–13. https://doi.org/10.1016/j.mce.2010.06.018.

Smith, A.F.A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.

Soto, L.F., Li, Z., Santoso, C.S., Berenson, A., Ho, I., Shen, V.X., Yuan, S., and Bass, J.I.F. (2022). Compendium of human transcription factor effector domains. Mol. Cell *82*, 514–526. https://doi.org/10.1016/j.molcel.2021.11.007.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Rep. *3*, 2179–2190.

https://doi.org/10.1016/j.celrep.2013.05.031.

Stehle, J.H., Foulkes, N.S., Molina, C.A., Simonneaux, V., Pévet, P., and Sassone-Corsi, P. (1993). Adrenergic signals direct rhythmic expression of transcriptional represser CREM in the pineal gland. Nature *365*, 314–320. https://doi.org/10.1038/365314a0.

Steingrímsson, E., Moore, K.J., Lamoreux, M.L., Ferré-D'Amaré, A.R., Burley, S.K., Sanders Zimring, D.C., Skow, L.C., Hodgkinson, C.A., Arnheiter, H., Copeland, N.G., et al. (1994). Molecular basis of mouse microphthalmia (mi) mutations helps explain their developmental and phenotypic consequences. Nat. Genet. *8*, 256–263. https://doi.org/10.1038/ng1194-256.

Stone, R.C., Du, P., Feng, D., Dhawan, K., Rönnblom, L., Eloranta, M.-L., Donnelly, R., and Barnes, B.J. (2013). RNA-Seq for Enrichment and Analysis of IRF5 Transcript Expression in SLE. PLOS ONE *8*, e54487. https://doi.org/10.1371/journal.pone.0054487.

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA *12*, 2. https://doi.org/10.1186/s13100-020-00230-y.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16–23. https://doi.org/10.1093/bioinformatics/16.1.16.

Stubbs, L., Sun, Y., and Caetano-Anolles, D. (2011). Function and Evolution of C2H2 Zinc Finger Arrays. In A Handbook of Transcription Factors, T.R. Hughes, ed. (Dordrecht: Springer Netherlands), pp. 75–94.

Sturm, R.A., and Herr, W. (1988). The POU domain is a bipartite DNA-binding structure. Nature *336*, 601–604. https://doi.org/10.1038/336601a0.

Sugiaman-Trapman, D., Vitezic, M., Jouhilahti, E.-M., Mathelier, A., Lauter, G., Misra, S., Daub, C.O., Kere, J., and Swoboda, P. (2018). Characterization of the human RFX transcription factor family by regulatory and target gene analysis. BMC Genomics *19*, 181. https://doi.org/10.1186/s12864-018-4564-6.

Sun, X.-H., and Baltimore, D. (1991). An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell *64*, 459–470. https://doi.org/10.1016/0092-8674(91)90653-G.

Sun, L., Liu, A., and Georgopoulos, K. (1996). Zinc finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. EMBO J. *15*, 5358–5369. https://doi.org/10.1002/j.1460-2075.1996.tb00920.x.

Taberlay, P.C., Kelly, T.K., Liu, C.-C., You, J.S., De Carvalho, D.D., Miranda, T.B., Zhou,

X.J., Liang, G., and Jones, P.A. (2011). Polycomb-Repressed Genes Have Permissive Enhancers that Initiate Reprogramming. Cell *147*, 1283–1294. https://doi.org/10.1016/j.cell.2011.10.040.

Takeda, K., and Akira, S. (2000). STAT family of transcription factors in cytokine-mediated biological responses. Cytokine Growth Factor Rev. *11*, 199–207. https://doi.org/10.1016/S1359-6101(00)00005-8.

Takeda, H., Matsuzaki, T., Oki, T., Miyagawa, T., and Amanuma, H. (1994). A novel POU domain gene, zebrafish pou2: expression and roles of two alternatively spliced twin products in early development. Genes Dev. *8*, 45–59. https://doi.org/10.1101/gad.8.1.45.

Takimoto, G.S., Tung, L., Abdel-Hafiz, H., Abel, M.G., Sartorius, C.A., Richer, J.K., Jacobsen, B.M., Bain, D.L., and Horwitz, K.B. (2003). Functional properties of the N-terminal region of progesterone receptors and their mechanistic relationship to structure. J. Steroid Biochem. Mol. Biol. *85*, 209–219. https://doi.org/10.1016/S0960-0760(03)00197-3.

Talavera, D., Orozco, M., and de la Cruz, X. (2009). Alternative Splicing of Transcription Factors' Genes: Beyond the Increase of Proteome Diversity. International Journal of Genomics *2009*, 905894. https://doi.org/10.1155/2009/905894.

Tanabe, M., Sasai, N., Nagata, K., Liu, X.-D., Liu, P.C.C., Thiele, D.J., and Nakai, A. (1999). The Mammalian HSF4 Gene Generates Both an Activator and a Repressor of Heat Shock Genes by Alternative Splicing. J. Biol. Chem. *274*, 27845–27856. https://doi.org/10.1074/jbc.274.39.27845.

Tanaka, Y., Yamashita, R., Suzuki, Y., and Nakai, K. (2010). Effects of Alu elements on global nucleosome positioning in the human genome. BMC Genomics *11*, 309. https://doi.org/10.1186/1471-2164-11-309.

Taneri, B., Snyder, B., Novoradovsky, A., and Gaasterland, T. (2004). Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. Genome Biol. *5*, R75. https://doi.org/10.1186/gb-2004-5-10-r75.

Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768. https://doi.org/10.1101/gr.220962.117.

The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021.

Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

Thomas, M.C., and Chiang, C.-M. (2006). The General Transcription Machinery and General Cofactors. Crit. Rev. Biochem. Mol. Biol. *41*, 105–178. https://doi.org/10.1080/10409230600648736.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Res. *13*, 2129–2141. https://doi.org/10.1101/gr.772403.

Thompson, B., Davidson, E.A., Liu, W., Nebert, D.W., Bruford, E.A., Zhao, H., Dermitzakis, E.T., Thompson, D.C., and Vasiliou, V. (2021). Overview of PAX gene family: analysis of human tissue-specific variant expression and involvement in human disease. Hum. Genet. *140*, 381–400. https://doi.org/10.1007/s00439-020-02212-9.

Tian, L., Jabbari, J.S., Thijssen, R., Gouil, Q., Amarasinghe, S.L., Voogd, O., Kariyawasam, H., Du, M.R.M., Schuster, J., Wang, C., et al. (2021). Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. *22*, 310. https://doi.org/10.1186/s13059-021-02525-6.

Titchenell, P.M., Lazar, M.A., and Birnbaum, M.J. (2017). Unraveling the Regulation of Hepatic Metabolism by Insulin. Trends Endocrinol. Metab. *28*, 497–505. https://doi.org/10.1016/j.tem.2017.03.003.

Todeschini, A.-L., Georges, A., and Veitia, R.A. (2014). Transcription factors: specific DNA binding and specific gene regulation. Trends Genet. *30*, 211–219. https://doi.org/10.1016/j.tig.2014.04.002.

Tora, L., Gronemeyer, H., Turcotte, B., Gaub, M.-P., and Chambon, P. (1988). The N-terminal region of the chicken progesterone receptor specifies target gene activation. Nature *333*, 185–188. https://doi.org/10.1038/333185a0.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem. Sci. *42*, 98–110. https://doi.org/10.1016/j.tibs.2016.08.008.

Tsukada, J., Yoshida, Y., Kominato, Y., and Auron, P.E. (2011). The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. Cytokine *54*, 6–19. https://doi.org/10.1016/j.cyto.2010.12.019.

Tsukamoto, K., Nakamura, Y., and Niikawa, N. (1994). Isolation of two isoforms of the PAX3 gene transcripts and their tissue-specific alternative expression in human adult tissues.

Hum. Genet. *93*, 270–274. https://doi.org/10.1007/BF00212021.

Turlure, F., Maertens, G., Rahman, S., Cherepanov, P., and Engelman, A. (2006). A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. Nucleic Acids Res. *34*, 1653–1665. https://doi.org/10.1093/nar/gkl052.

Ullmark, T., Järvstråt, L., Sandén, C., Montano, G., Jernmark-Nilsson, H., Lilljebjörn, H., Lennartsson, A., Fioretos, T., Drott, K., Vidovic, K., et al. (2017). Distinct global binding patterns of the Wilms tumor gene 1 (WT1) −KTS and +KTS isoforms in leukemic cells. Haematologica *102*, 336–345. https://doi.org/10.3324/haematol.2016.149815.

Verrijzer, C.P., Kal, A.J., and Vliet, P.C. van der (1990). The oct-1 homeo domain contacts only part of the octamer sequence and full oct-1 DNA-binding activity requires the POU-specific domain. Genes Dev. *4*, 1964–1974. https://doi.org/10.1101/gad.4.11.1964.

Vihma, H., Luhakooder, M., Pruunsild, P., and Timmusk, T. (2016). Regulation of different human NFAT isoforms by neuronal activity. J. Neurochem. *137*, 394–408. https://doi.org/10.1111/jnc.13568.

Vitting-Seerup, K., and Sandelin, A. (2017). The Landscape of Isoform Switches in Human Cancers. Mol. Cancer Res. *15*, 1206–1220. https://doi.org/10.1158/1541-7786.MCR-16-0459.

Vitting-Seerup, K., and Sandelin, A. (2019). IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics *35*, 4469–4471. https://doi.org/10.1093/bioinformatics/btz247.

Vo ngoc, L., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C., and Kadonaga, J. T. (2017). The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. Genes Dev. *31*, 6–11. https://doi.org/10.1101/gad.293837.116.

Volden, R., and Vollmers, C. (2022). Single-cell isoform analysis in human immune cells. Genome Biol. *23*, 47. https://doi.org/10.1186/s13059-022-02615-z.

Vuong, L. (2014). The Role of the P1 and P2 Promoter-Driven HNF4a Isoforms in Cellular Proliferation and Differentiation in Human Colon Cancer and Mouse Embryonic Stem Cells. A Dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Cell, Molecular, and Developmental Biology. UC Riverside. ProQuest ID: Vuong_ucr_0032D_12023. Merritt ID: ark:/13030/m5kp9gg1. Retrieved from https://escholarship.org/uc/item/0gb1j7p3.

Vuzman, D., Azia, A., and Levy, Y. (2010). Searching DNA via a "Monkey Bar" Mechanism: The Significance of Disordered Tails. J. Mol. Biol. *396*, 674–684. https://doi.org/10.1016/j.jmb.2009.11.056.

Wagner, S., Hess, M.A., Ormonde-Hanson, P., Malandro, J., Hu, H., Chen, M., Kehrer, R., Frodsham, M., Schumacher, C., Beluch, M., et al. (2000). A Broad Role for the Zinc Finger Protein ZNF202 in Human Lipid Metabolism. J. Biol. Chem. *275*, 15685–15690. https://doi.org/10.1074/jbc.M910152199.

Walhout, A. J. M. (2006). Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping. Genome Res. *16*, 1445–1454. https://doi.org/10.1101/gr.5321506.

Wang, D., Zou, X., and Fai Au, K. (2021). A network-based computational framework to predict and differentiate functions for gene isoforms using exon-level expression data. Methods *189*, 54–64. https://doi.org/10.1016/j.ymeth.2020.06.005.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476. https://doi.org/10.1038/nature07509.

Wang, H., Sartini, B.L., Millette, C.F., and Kilpatrick, D.L. (2006a). A Developmental Switch in Transcription Factor Isoforms During Spermatogenesis Controlled by Alternative Messenger RNA 3'-End Formation. Biol. Reprod. *75*, 318–323. https://doi.org/10.1095/biolreprod.106.052209.

Wang, Q., Kumar, S., Slevin, M., and Kumar, P. (2006b). Functional Analysis of Alternative Isoforms of the Transcription Factor PAX3 in Melanocytes In vitro. Cancer Res. *66*, 8574–8580. https://doi.org/10.1158/0008-5472.CAN-06-0947.

Weatheritt, R.J., Sterne-Weiler, T., and Blencowe, B.J. (2016). The ribosome-engaged landscape of alternative splicing. Nat. Struct. Mol. Biol. *23*, 1117–1123. https://doi.org/10.1038/nsmb.3317.

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160. https://doi.org/10.1038/emboj.2010.106.

Weirauch, M.T., and Hughes, T.R. (2011). A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution. In A Handbook of Transcription Factors, (Science+Business Media B. V.), pp. 25–73.

Wetlaufer, D.B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in

Proteins. Proc. Natl. Acad. Sci. *70*, 697–701. https://doi.org/10.1073/pnas.70.3.697.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA Recognition by Cys2His2 Zinc Finger Proteins. Annu. Rev. Biophys. Biomol. Struct. *29*, 183–212. https://doi.org/10.1146/annurev.biophys.29.1.183.

Workman, J.L., and Kingston, R.E. (1992). Nucleosome Core Displacement in Vitro via a Metastable Transcription Factor-Nucleosome Complex. Science *258*, 1780–1784. https://doi.org/10.1126/science.1465613.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-S.L., Natale, D.A., Vinayaka, C.R., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P., et al. (2004). PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. *32*, D112–D114. https://doi.org/10.1093/nar/gkh097.

Xhani, S., Lee, S., Kim, H.M., Wang, S., Esaki, S., Ha, V.L.T., Khanezarrin, M., Fernandez, G.L., Albrecht, A.V., Aramini, J.M., et al. (2020). Intrinsic disorder controls two functionally distinct dimers of the master transcription factor PU.1. Sci. Adv. *6*, eaay3178. https://doi.org/10.1126/sciadv.aay3178.

Xiong, L., Catoire, H., Dion, P., Gaspar, C., Lafrenière, R.G., Girard, S.L., Levchenko, A., Rivière, J.-B., Fiori, L., St-Onge, J., et al. (2009). MEIS1 intronic risk haplotype associated with restless legs syndrome affects its mRNA and protein expression levels. Hum. Mol. Genet. *18*, 1065–1074. https://doi.org/10.1093/hmg/ddn443.

Yamazaki, T., Liu, L., Lazarev, D., Al-Zain, A., Fomin, V., Yeung, P.L., Chambers, S.M., Lu, C.-W., Studer, L., and Manley, J.L. (2018). TCF3 alternative splicing controlled by hnRNP H/F regulates E-cadherin expression and hESC pluripotency. Genes Dev. *32*, 1161–1174. https://doi.org/10.1101/gad.316984.118.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659. https://doi.org/10.1093/bioinformatics/bti042.

Yang, A., Kaghad, M., Wang, Y., Gillett, E., Fleming, M.D., Dötsch, V., Andrews, N.C., Caput, D., and McKeon, F. (1998). p63, a p53 Homolog at 3q27–29, Encodes Multiple Products with Transactivating, Death-Inducing, and Dominant-Negative Activities. Mol. Cell *2*, 305–316. https://doi.org/10.1016/S1097-2765(00)80275-0.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., et al. (2016). Widespread Expansion of

Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805–817. https://doi.org/10.1016/j.cell.2016.01.029.

Yang, Y., Hwang, C.K., D'Souza, U.M., Lee, S.-H., Junn, E., and Mouradian, M.M. (2000). Three-amino acid Extension Loop Homeodomain Proteins Meis2 and TGIF Differentially Regulate Transcription. J. Biol. Chem. *275*, 20734–20741. https://doi.org/10.1074/jbc.M908382199.

Yasumoto, K.-I., Amae, S., Udono, T., Fuse, N., Takeda, K., and Shibahara, S. (1998). A Big Gene Linked to Small Eyes Encodes Multiple Mitf Isoforms: Many Promoters Make Light Work. Pigment Cell Res. *11*, 329–336. https://doi.org/10.1111/j.1600-0749.1998.tb00491.x.

Yeo, G., Holste, D., Kreiman, G., and Burge, C.B. (2004). Variation in alternative splicing across human tissues. Genome Biol. *5*, R74. https://doi.org/10.1186/gb-2004-5-10-r74.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239. https://doi.org/10.1126/science.aaj2239.

Yu, W., Li, R., Gui, B., and Shang, Y. (2010). sZIP, an Alternative Splice Variant of ZIP, Antagonizes Transcription Repression and Growth Inhibition by ZIP. J. Biol. Chem. *285*, 14301–14307. https://doi.org/10.1074/jbc.M110.107508.

Yu, Y.T., Breitbart, R.E., Smoot, L.B., Lee, Y., Mahdavi, V., and Nadal-Ginard, B. (1992). Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. Genes Dev. *6*, 1783–1798. https://doi.org/10.1101/gad.6.9.1783.

Yura, K., Shionyu, M., Hagino, K., Hijikata, A., Hirashima, Y., Nakahara, T., Eguchi, T., Shinoda, K., Yamaguchi, A., Takahashi, K., et al. (2006). Alternative splicing in human transcriptome: Functional and structural influence on proteins. Gene *380*, 63–71. https://doi.org/10.1016/j.gene.2006.05.015.

Zaika, A.I., Slade, N., Erster, S.H., Sansome, C., Joseph, T.W., Pearl, M., Chalas, E., and Moll, U.M. (2002). ΔNp73, A Dominant-Negative Inhibitor of Wild-type p53 and TAp73, Is Up-regulated in Human Tumors. J. Exp. Med. *196*, 765–780. https://doi.org/10.1084/jem.20020179.

Zanegina, O., Kirsanov, D., Baulin, E., Karyagina, A., Alexeevski, A., and Spirin, S. (2016). An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. Nucleic Acids Res. *44*, D144–D153.

https://doi.org/10.1093/nar/gkv1339.

Zea, D.J., Laskina, S., Baudin, A., Richard, H., and Laine, E. (2021). Assessing conservation of alternative splicing with evolutionary splicing graphs. Genome Res. *31*, 1462–1473. https://doi.org/10.1101/gr.274696.120.

Zetser, A., Gredinger, E., and Bengal, E. (1999). p38 Mitogen-activated Protein Kinase Pathway Promotes Skeletal Muscle Differentiation: PARTICIPATION OF THE MEF2C TRANSCRIPTION FACTOR. J. Biol. Chem. *274*, 5193–5200. https://doi.org/10.1074/jbc.274.8.5193.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620. https://doi.org/10.1093/bioinformatics/btt593.

Zhang, J., Ren, J., Wei, J., Chong, C.C.N., Yang, D., He, Y., Chen, G.G., and Lai, P.B.S. (2016). Alternative splicing of estrogen receptor alpha in hepatocellular carcinoma. BMC Cancer *16*, 926. https://doi.org/10.1186/s12885-016-2928-3.

Zhang, S., Lachance, B.B., Mattson, M.P., and Jia, X. (2021). Glucose metabolic crosstalk and regulation in brain function and diseases. Prog. Neurobiol. *204*, 102089. https://doi.org/10.1016/j.pneurobio.2021.102089.

Zhang, Y., Huypens, P., Adamson, A.W., Chang, J.S., Henagan, T.M., Boudreau, A., Lenard, N.R., Burk, D., Klein, J., Perwitz, N., et al. (2009). Alternative mRNA Splicing Produces a Novel Biologically Active Short Isoform of PGC-1α. J. Biol. Chem. *284*, 32813–32826. https://doi.org/10.1074/jbc.M109.037556.

Zhang, Y., Yamada, Y., Fan, M., Bangaru, S.D., Lin, B., and Yang, J. (2010). The β Subunit of Voltage-gated Ca2+ Channels Interacts with and Regulates the Activity of a Novel Isoform of Pax6. J. Biol. Chem. *285*, 2527–2536. https://doi.org/10.1074/jbc.M109.022236.

Zhao, J., Tan, C., Imai, R., Ukon, N., Shimoyama, S., Maejima, Y., Omiya, Y., Takahashi, K., Ito, H., Nan, G., et al. (2021). Evaluation of organ glucose metabolism by 18F-FDG accumulation with insulin loading in aged mice compared with young normal mice. Sci. Rep. *11*, 7421. https://doi.org/10.1038/s41598-021-86825-8.

Zhao, S., Liu, W., Li, Y., Liu, P., Li, S., Dou, D., Wang, Y., Yang, R., Xiang, R., and Liu, F. (2016). Alternative Splice Variants Modulates Dominant-Negative Function of Helios in T-Cell Leukemia. PLOS ONE *11*, e0163328. https://doi.org/10.1371/journal.pone.0163328.

Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., and Gymrek, M. (2021). Deep neural

networks identify sequence context features predictive of transcription factor binding. Nature Machine Intelligence *3*, 172–180. https://doi.org/10.1038/s42256-020-00282-y.

Zheng, G., and Yang, Y.-C. (2006). Acetylation and alternative splicing regulate ZNF76-mediated transcription. Biochem. Biophys. Res. Commun. *339*, 1069–1075. https://doi.org/10.1016/j.bbrc.2005.11.122.

Zhou, J., Ng, A.Y., Tymms, M.J., Jermiin, L.S., Seth, A.K., Thomas, R.S., and Kola, I. (1998). A novel transcription factor, ELF5, belongs to the ELF subfamily of ETS genes and maps to human chromosome 11p13–15, a region subject to LOH and rearrangement in human carcinoma cell lines. Oncogene *17*, 2719–2732. https://doi.org/10.1038/sj.onc.1202198.

Zhou, K., Gaullier, G., and Luger, K. (2019). Nucleosome structure and dynamics are coming of age. Nat. Struct. Mol. Biol. *26*, 3–13. https://doi.org/10.1038/s41594-018-0166-x.

Zhu, C., Wu, J., Sun, H., Briganti, F., Meder, B., Wei, W., and Steinmetz, L.M. (2021). Single-molecule, full-length transcript isoform sequencing reveals disease-associated RNA isoforms in cardiomyocytes. Nat. Commun. *12*, 4203. https://doi.org/10.1038/s41467-021-24484-z.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M., et al. (2018). The interaction landscape between transcription factors and the nucleosome. Nature *562*, 76. https://doi.org/10.1038/s41586-018-0549-5.