Structural Basis of Gene Regulation by the Transcription Factors Tfcp2l1 and Tfcp2

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

Submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

JIANHUI WANG

from Henan, China

June 2021

The thesis is based on work performed from April 2017 until June 2021 under the supervision of Prof. Dr. Udo Heinemann at the Max-Delbrück Center for Molecular Medicine, Berlin-Buch.

1st Reviewer: Prof. Dr. Udo Heinemann

2nd Reviewer: Prof. Dr. Oliver Daumke

Date of defense: 26.08.2021

DECLARATION

I hereby declare that the present thesis *Structural Basis of Gene Regulation by the Transcription Factors Tfcp2l1 and Tfcp2* is my own work and effort and has not been submitted to any other university for any award. Where other sources of information have been used, they have been acknowledged.

Berlin, June 2021

Jianhui Wang

ABSTRACT

Mammalian Tfcp2l1 and Tfcp2 are grouped into the CP2 subfamily of the grainyhead/CP2 (Grh/CP2) transcription factors involved in pluripotency maintenance and self-renewal of embryonic stem cells. In addition, Tfcp2l1 has been implicated in a variety of cancers such as breast cancer, thyroid cancer, and clear cell renal cell carcinoma. Recent studies also reveal that the mammalian transcription factors Grhl1 and Grhl2 display a similar tertiary structure as the tumor suppressor TP53, although the protein sequences share only 10% identical residues. Despite the conserved tertiary structure, Grhl1/2 and TP53 have a different mode of DNA binding. This thesis presents structures of the ligand-free Tfcp2l1 and Tfcp2 DNA-binding domains (DBDs) and the DNA-bound Tfcp2l1 DBD. These structures provide insight into protein DNA recognition.

The Tfcp2l1 DBD and Tfcp2 DBD structures are similar, and they belong to the immunoglobulin-(Ig-) like fold, which is shared by the Grhl1/2 DBD structures. The study confirmed that the DBD structure is highly conserved within the Grh/CP2 family. Tfcp2l1 DBD binds to a 12-mer target DNA fragment in a parsimonious binding mode, which is similar to the Grhl1-DBD:DNA complex. The specific contacts performed by residues Arg225 and Gly183 interacting with guanosine G8 supply the selectivity of protein DNA recognition. Unspecific contacts play an additional role in anchoring the protein to DNA via residue interaction with DNA phosphate groups.

Tfcp2l1 DBD prefers to bind to the $AAAAC_5CGG_8TTTT$ sequence rather than the C_5CAG_8 sequence. The conserved nucleotides cytosine C5 and guanosine G8 of the duplex DNA play a primary role in the readout of the DNA sequence by the protein, and the DNA shape supplies additional selectivity for Tfcp2l1 to readout the DNA sequence. The conformation of the target DNA may fine-tune the protein:DNA interaction.

The SAM domain of Tfcp2l1 is involved in protein tetramerization, and Tfcp2l1 binds to the DNA sequence of $AAACCAGN_6CCAGTTT$ in a mode of four DBDs binding to two consensus DNA motifs. The spacing of the CCAG core motifs recognized by Tfcp2l1 is not fixed at 6 bps, but may be reduced to 5 bps without generating spatial clashes.

The work described in this thesis reveals the mechanism of target DNA recognition by CP2 subfamily transcription factors. Crystal structure analyses and biophysical experiments provide insight into protein:DNA interaction involving CP2 factors and open up novel avenues for diagnosis and therapies of various epithelial cancers and kidney diseases.

ABSTRACT

ZUSAMMENFASSUNG

Tfcp2l1 und Tfcp2 aus Säugern werden der CP2-Unterfamilie der Grainyhead/CP2 (Grh/CP2) Transkriptionsfaktoren zugeordnet, die an der Pluripotenzerhaltung und Selbsterneuerung embryonaler Stammzellen beteiligt sind. Darüber hinaus wurde Tfcp2l1 mit einer Vielzahl von Krebsarten wie Brustkrebs, Schilddrüsenkrebs und klarzelligem Nierenzellkarzinom in Verbindung gebracht. Neuere Studien zeigen auch, dass die Säuger-Transkriptionsfaktoren Grhl1 und Grhl2 eine ähnliche Tertiärstruktur aufweisen wie der Tumorsuppressor TP53, obwohl die Proteinsequenzen nur 10% identische Reste aufweisen. Trotz der konservierten Tertiärstruktur weisen Grhl1/2 und TP53 einen unterschiedlichen DNA-Bindungsmodus auf. Diese Arbeit präsentiert Strukturen der ligandenfreien Tfcp2l1- und Tfcp2-DNA-Bindungsdomänen (DBDs) und der DNA-gebundenen Tfcp2l1-DBD. Diese Strukturen geben einen Einblick in die Protein-DNA-Erkennung.

Die Tfcp2l1-DBD- und Tfcp2-DBD-Strukturen sind ähnlich und weisen eine Immunglobulin- (Ig-) ähnliche Faltung auf, die von den Grhl1/2-DBD-Strukturen geteilt wird. Die Studie bestätigte, dass die DBD-Struktur innerhalb der Grh/CP2-Familie hoch konserviert ist. Tfcp211 DBD bindet an ein 12-mer DNA-Zielfragment in einem „sparsamen“ Bindungsmodus, der dem Grhl1-DBD:DNA-Komplex ähnelt. Die spezifischen Kontakte, die von den mit Guanosin-G8 wechselwirkenden Resten Arg225 und Gly183 ausgebildet werden, bestimmen die Selektivität der Protein-DNA-Erkennung. Unspezifische Kontakte spielen eine zusätzliche Rolle bei der Verankerung des Proteins an DNA über die Wechselwirkung von Resten mit DNA-Phosphatgruppen.

Tfcp211 DBD bindet bevorzugt an die AAAAC$_5$CGG$_8$TTTT-Sequenz anstatt an die C$_5$CAG$_8$-Sequenz. Die konservierten Nukleotide Cytidin C5 und Guanosin G8 der Duplex-DNA spielen eine Hauptrolle beim Auslesen der DNA-Sequenz durch das Protein, und Geometrie der DNA bietet zusätzliche Selektivität für Tfcp2l1 beim Auslesen der DNA-Sequenz. Die Konformation der Ziel-DNA kann die Protein:DNA-Interaktion feinsteuern.

Die SAM-Domäne von Tfcp2l1 trägt entscheidend zur Tetramerisierung des Proteins bei, und Tfcp2l1 bindet an die DNA-Sequenz von AAACCAGN$_6$CCAGTTT in einem Modus von vier DBDs, die an zwei Konsensus-DNA-Motive binden. Der Abstand der von Tfcp2l1 erkannten CCAG-Kernmotive ist nicht strikt auf 6 bps festgelegt, sondern kann auf 5 bps reduziert werden, ohne räumliche Kollisionen zu erzeugen.

ABSTRACT

Die in dieser Dissertation beschriebene Arbeit enthüllt den Mechanismus der Ziel-DNA-Erkennung durch Transkriptionsfaktoren der CP2-Unterfamilie. Kristallstrukturanalysen und biophysikalische Experimente geben Einblicke in die Protein-DNA-Interaktion mit CP2-Faktoren und eröffnen neue Wege für Diagnose und Therapie verschiedener epithelialer Krebsarten und Nierenerkrankungen.

CONTENTS

1. INTRODUCTION

Transcription factors (TFs) are essential for gene transcription. Through coordination with the general transcription machinery and chromatin regulators, TFs control gene expression, culminating in cell fate determination[1]. Overexpressed or dysfunctional TFs may cause various cancers and diseases[2,3]. The Grainyhead/CP2 (Grh/CP2) TF family is highly conserved from fly to human[4]. Its members play critical roles in regulating embryonic development, maintaining epithelial integrity and proper function of the epidermis[5,6]. The Grh/CP2 transcription factor family is divided into two main branches: the CP2 subfamily (Tfcp2, Tfcp2l, and Ubp1)[5] and the Grainyhead-like (Grhl) subfamily (Grhl1, Grhl2 and, Grhl3)[4]. The CP2 subfamily TFs are involved in the development of a variety of cancers. The biological functions of CP2 subfamily members have been well documented. However, it is still unclear how the CP2 factors interact with DNA recognition sites. Knowledge of the structural basis and molecular details of protein:DNA interactions, will contribute to understanding the impact of modifications on members of the CP2 TF subfamily. Therefore, in this thesis, I focus on analyzing the three-dimensional structures of CP2 family members through X-ray crystallography and applying biochemical and biophysical methods to elucidate the molecular basis of the Tfcp2l1-DNA interaction and target sequence recognition. I expect to clarify a new DNA binding pattern in the CP2 subfamily transcription factors, which may open up novel avenues for diagnosing and treating various epithelial cancers as well as kidney diseases.

1.1 Evolution of the Grh/CP2 transcription factor family

The late simian virus (SV) 40 transcription factor (LSF) was first described in HeLa cells extracts in 1987, where it served as a transcriptional activator binding specifically to the SV40 21-bp repeat promoter region[7,8]. LSF was independently identified by different laboratories and assigned various synonyms: TFCP2c (transcription factor CP2c)[9], LBP-1c (leader binding protein-1c), and LBP-1d[10,11]. The CCAAT binding protein 2 (CP2, targeting the murine α-globin promoter) was identified in 1988 and suggested to be identical to LSF. However, later studies reported that LSF does not bind to the CCAAT DNA sequence[12]. Therefore, the authentic CCAAT binding protein 2 and LSF are unrelated, which causes some confusion in the term “CP2”[13]. The CP2 subfamily TFs studied in this thesis are unrelated to CCAAT binding protein 2.

The term “Grh” was used to describe the grainyhead gene mutant in *Drosophila* embryos displaying a phenotype with flimsy cuticles, grainy and discontinuous head skeletons and patchy tracheal tubes[14]. The Grh protein was first identified in the *Drosophila melanogaster* central nervous system in 1988, where it binds to the dopa decarboxylase (*Ddc*) gene[15]. Grh was independently identified by different laboratories and assigned the synonyms neuronal transcription factor 1 (NTF-1)[16,17] and *cis*-acting element factor 1 (Elf-1)[14,18].

The Grh/CP2 TF family shares an immunoglobulin-like DNA binding domain and is highly conserved from *Drosophila* to humans in evolution[5,13]. In mammalian species, the Grh/CP2 TF family is divided into two main branches: the CP2 subfamily (Tfcp2, Tfcp2l1, and Ubp1)[5] and the Grainyhead-like (Grhl) subfamily (Grhl1, Grhl2, and Grhl3)[4], depending on whether they are more closely related to the *Drosophila* protein CP2 or Grh[6] (Fig. 1-1). The split of the Grh/CP2 family occurred around 700 million years ago[5]. Although the Grh/CP2 family proteins may have evolved from the common ancestor TP53[19], there are many differences in biological function and protein structure between the Grhl and the CP2 subfamily.

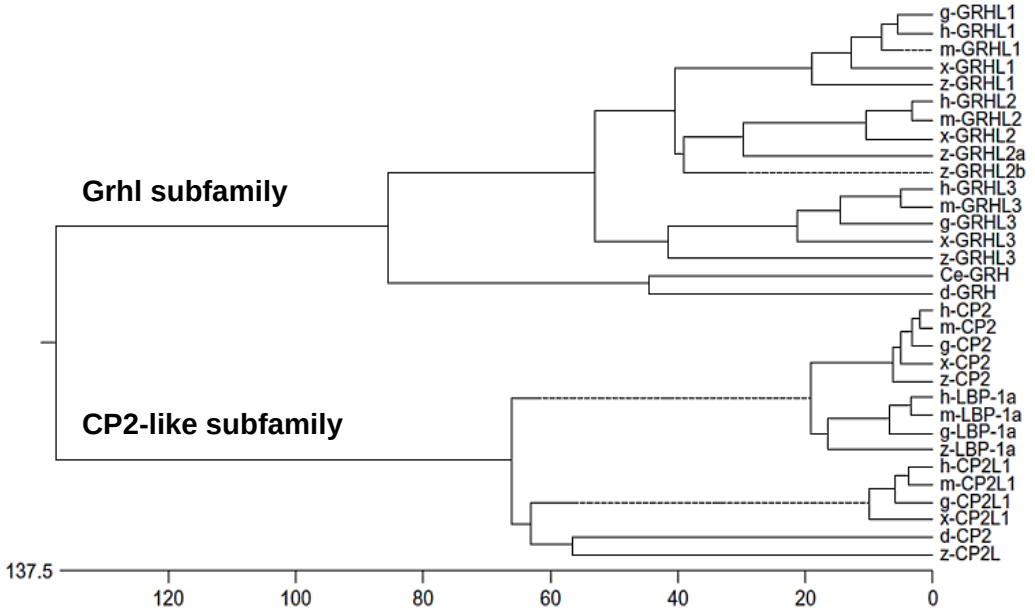

Figure 1-1. Phylogenetic tree of the Grh/CP2 family transcription factors. The Grh/CP2 transcription factor family is divided into two main branches: the CP2 subfamily (Tfcp2, Tfcp2l1 and Ubp1) and the Grainyhead-like (Grhl) subfamily (Grhl1, Grhl2 and Grhl3). *Ce, Caenorhabditis elegans; d, Drosophila melanogaster; g, Gallus gallus; h, Homo sapiens; m, Mus musculus, x, Xenopus laevis; z,* zebrafish*, Danio rerio*[6].

The Grhl subfamily has been well studied in terms of both function and structure. *Grhl* subfamily genes are predominantly expressed during embryogenesis, in the central nervous system, and cuticular tissues essential for epidermal development and regeneration and wound repair[20]. Several reviews have reported on the Grhl subfamily[6,21,22]. In 2018, Heinemann's group determined crystal structures of the DNA binding domain of Grhl1 and Grhl2, which share a similar structure with tumor suppressor TP53[23]. For the CP2 subfamily, many reports have focused on the biological function of its three members, Tfcp2, Tfcp2l1 and Ubp1[5,13,24,25]. The work described in this thesis is also focused on the CP2 subfamily proteins.

1.2 Transcription factor CP2 subfamily

1.2.1 Identification of the *CP2* gene in *Drosophila*

Unlike the Grhl subfamily, the first *Grh* gene was identified in *Drosophila*[17,26], and its homologous genes were subsequently found in animals ranging from nematodes to humans[4,27,28]. In CP2 subfamily, the *CP2* gene was identified after its homologous genes. Expression and functional analysis indicated that compared with the Grhl subfamily and *Drosophila* Grh protein, Tfcp2, Tfcp2l1, and Ubp1 displayed distinct functions. Wilanowski and colleagues predicted that there would be a gene in *Drosophila,* which was closely related to Tfcp2, Tfcp2l1, and Ubp1. By screening a cDNA library from *Drosophila* embryos, they identified the novel gene and named it *Drosophila CP2 (dCP2)*[4].

1.2.2 Mammalian CP2 subfamily

1.2.2.1 Location of mammalian *CP2* subfamily genes in the genome

Three members of *CP2* subfamily genes are located on the human genome:

- The *Tfcp2* gene resides on chromosome (Chr) 12q13 and contains 15 exons.
- The *Ubp1* gene maps to Chr 3q22 (16 exons).
- The *Tfcp2l1* gene is located on Chr 2q14 (15 exons)[29].

In mice, these three genes are also located on different chromosomes: mTfcp2 – Chr 15qF1 (16 exons), mUbp1 – Chr 9qF3 (16 exons), and mTfcp2l1 – Chr 1qE2.3 (15 exons). Like Tfcp2, Tfcp2l1 and Ubp1 have different synonyms in non-human mammals. The Tfcp2l1 is referred to as CP2 related transcriptional repressor-1 (CRTR-1)[30] or long terminal repeat binding protein-9 (LBP-9)[31].

Upstream binding protein 1 (Ubp1) is also known as LBP1, which contains two isoforms: LBP-1a and LBP-1b[11], and is also referred to as nuclear factor 2d9 (NF2d9)[32].

1.2.2.2 Expression of mammalian *CP2* subfamily genes

As previously described, Tfcp2 was first found in humans that could bind to a promoter element of SV40 gene[7] and orthologous mouse Tfcp2 could bind to a promoter element of the murine α-globin gene[33]. From early embryonic development to terminal cell differentiation, Tfcp2 is involved in regulating the expression of specific target genes[34]. Ubp1 was initially found at the human immunodeficiency virus type 1 (*HIV-1*) gene transcription initiation site and could repress *HIV-1* gene transcription[35]. Nevertheless, Ubp1 could interact with Tfcp2 to activate the α-globin gene transcription in erythroid cells that functions as a transcriptional activator. Both Tfcp2 and Ubp1 mRNA are ubiquitously expressed in fetal and all adult mouse tissue and all human cell lines[13,36,37].

Tfcp2l1 (Mouse Genome Informatics, referred to as LBP-9) was first discovered to bind to the promoter region of the gene *P450scc* (-155/-131)[31]. In human JEG-3 cells, Tfcp2l1 suppresses the transcriptional activation effect of Ubp1 on *P450scc* in a co-expression assay[31,38]. In mice, Tfcp2l1 was also named CRTR-1[30], expressed spatiotemporally in pluripotent epithelial cells and adult kidney distal convoluted tubules (DCT)[39]. In early mouse embryos, the expression of Tfcp2l1 was constant from 3.5 days *post coitum* (dpc) to 4.5 dpc, while the expression was decreasing from 4.5 dpc to 4.75 dpc, and after 5.0 dpc, Tfcp2l1 expression could not be detected[40]. The down-regulation of Tfcp2l1 expression during 4.5-4.75 dpc revealed a transient pluripotent cell population[40]. Although Tfcp2l1 is highly expressed in the embryonic epithelial monolayer derived from distal kidney tubules, Tfcp2l1 is not expressed in the proximal convoluted tubules of the kidney[39]. Overall, Tfcp2l1 is expressed in a developmental and tissue-specific manner.

1.2.3 Structure and function analysis of CP2 subfamily members

1.2.3.1 Functional domain of CP2 subfamily factors

The three members of the CP2 subgroup, Tfcp2, Tfcp2l1 and Ubp1, are closely related. Protein sequence alignment showed that Tfcp2l1 (Uniprot code Q9NZI6) and Tfcp2 (Q12800) share 74.4% residue identity, Tfcp2l1 and Ubp1 (Q9NZI7) have 63.3% identity. Mouse Tfcp2 and human Tfcp2 share 96% identity in amino acid sequence as do Tfcp2l1 and Ubp1. All three proteins have the same domain structure: a very N-terminal sequence: an intermediate DNA binding immunoglobulin fold, which is homologous to the TP53 core DNA binding domain (DBD), and a

sterile alpha motif (SAM) domain and a C-terminal domain (CTD), SAM and CTD might involve in protein tetramerization or oligomerization function[19] (Fig. 1-2).

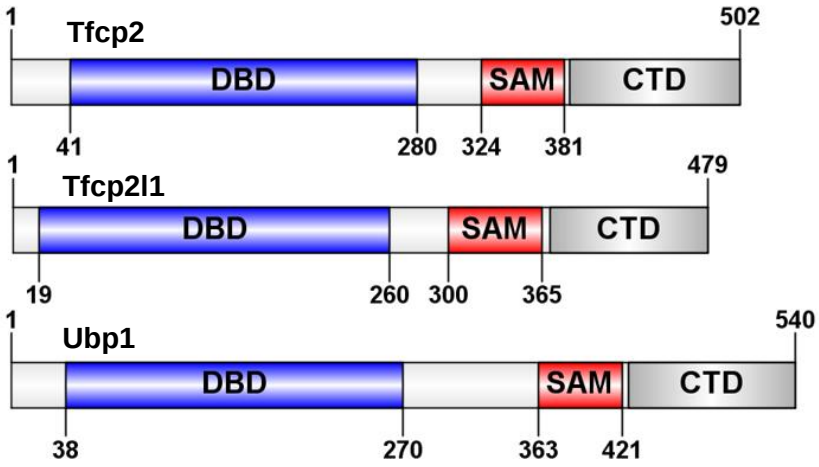

Figure 1-2. Schematic illustration of human CP2-subfamily transcription factor domain organization. Blue, red, and gray boxes indicate the DNA-binding domain (DBD), sterile alpha motif domain (SAM), and C-terminal domain (CTD), respectively.

1.2.3.2 Transcriptional repression or activation domain

Compared with the Grhl subfamily members, the CP2 subfamily factors lack an N-terminal transactivation domain (TAD). In general, Tfcp2 acts as a transcriptional activator, and the N-terminal 40 amino acids of Tfcp2 are sufficient to stimulate transcription[13]. Tfcp2 directly interacts with TATA-binding protein and TFIIB to increase TFIIB binding with the DNA sequence[41]. Tfcp2 can also act as a transcriptional repressor. For example, Tfcp2 represses the human immunodeficiency virus type-1 (HIV-1) long terminal repeat (LTR) transcription[42]. Amino acids 266-396 of Tfcp2 are sufficient for transcriptional repression[13]. Tfcp2 activates or inhibits the target gene's transcription in a context-dependent manner.

Tfcp2l1 was first discovered to be a transcriptional repressor, and the N-terminal 52 amino acids (AAs) of Tfcp2l1 are necessary and sufficient to maintain this activity[30]. Further studies reported that AAs 48-200 of LBP-9 conferred a suppressive response independent of the 52 AAs region at the N-terminus[43]. Gal4 transactivation assays showed that Tfcp2l1 lacks the classical activation domain to maintain the transactivation function[38,43]. The transcription activation activity may be attributed to post-translational modification or coordination with other factors.

Ubp1 also binds to the HIV-1 LTR region and regulates its transcription in a sequence-specific manner, either by activation or repression[35,44]. The 60 AAs at the N-terminus of Ubp1 share a high

(85%) similarity with Tfcp2, while AAs 266-396 share only 44% identity. So far, only few studies have reported on the classic transcriptional activation or repression on Ubp1.

1.2.3.3 DNA-binding domain

Tfcp2 was initially described to recognize the DNA sequence CNRGN$_6$CNRG (N = any nucleotide, R = purine nucleotide)[45]. A later compilation provides the more precise 16-mer target sequence (GCTGGTTTGTGCTTGC)[46]. And C and G are strictly conserved. Several studies have reported that the spacing between the (G)CTGG and CTTG(C) motifs is important, but the central 6 base pairs' identity is not so important[35,47,48]. The DNA recognition sequences of the CP2 and Grhl subfamilies are highly related. Ming et al. established that the Grhl subfamily binds to a symmetrical 12-mer DNA[23]. Based on ChIPSeq analysis, Tfcp2l1 will bind to a 14-mer DNA motif, while Tfcp2 will bind to 12-mer DNA fragments, which is highly similar to the Grhl subfamily[49] (Fig. 1-3). Further experiments are required to determine the correct target DNA sequence bound by Tfcp2.

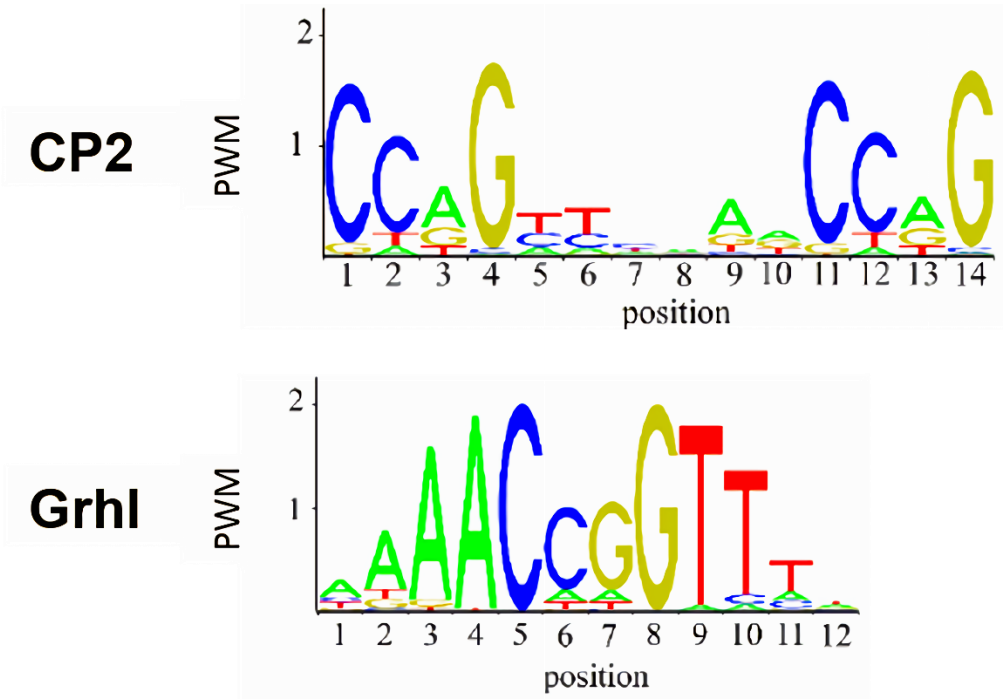

Figure 1-3. DNA recognition sequences for CP2 and Grhl subfamily members. DNA sequences are shown as position weight matrices.

The boundaries of the Tfcp2 DNA-binding domain (DBD) were first determined to extend from residues 63 to 270 which conveyed full DNA binding activity[10]. A Tfcp2 fragment including AAs 133-383 binds the target DNA motif in the form of a tetramer, since this region includes the DBD

and SAM domains[50]. The C-terminal boundary of Tfpc2 DBD is located near residues 272[5]. Two short conservative extensions comprising residues 205-216 and 233-246 of Tfcp2 are critical for the DNA binding activity. Mutations in either one of these two regions can disrupt the binding of Tfcp2 DBD to DNA[51].

DNA binding assays have approximately delineated the Tfcp2l1 DBD region, and the N-terminal region (AAs, 1-260) is predicted to be a CP2-like DBD[52]. The precise boundary of the Tfcp2l1 DBD has not been further tested. Among the Grh/CP2 family factors, the DNA binding domain is the most conserved. Although the two subfamilies share approximately 20% sequence identity, the DBD is highly conserved within the CP2 subfamily. Tfcp2 DBD and Tfcp2l1 DBD share 86% sequence identity, and Tfcp2 and Ubp1 DBD share 89% identity. For Ubp1, based on the protein sequence profiles, it has a similar DBD boundary as Tfcp2.

1.2.3.4 Oligomerization domain and C-terminal domain

The CP2 subfamily factors contain a SAM domain and CTD domain at the C-terminus, which might involve oligomerization functions[5]. The CTD domain of the CP2 subfamily only shares a 28% sequence identity to DD domain of the Grhl subfamily. In contrast, the CTD is highly conserved with 64% residue identity within the CP2 subfamily. Generally, a dimeric structure is essential for any DNA-binding protein that recognizes palindromic target sequences[19]. DBDs mediate the interaction between transcription factor and DNA in a dimeric arrangement in Grhl subfamily, but there is no structure evidence for the dimerization domain of the CP2 subfamily.

Initially, the oligomeric region of Tfcp2 was approximately delineated by Shirra *et al.*. They found that protein residues from 133 to 383 bind to DNA in the form of tetramers[51]. *In vitro* crosslinking revealed that oligomerization was mediated by the protein region from AA 266 to 403 outside the DBD, in a polypeptide region referred to as sterile alpha-motif (SAM) domain[13]. A SAM domain located between DBD and CTD is also present in Tfcp2l1 and Ubp1. SAM domains have previously been found in many different proteins, such as protein kinases, lipid metabolism regulators, and transcription factors[19]. SAM domains are involved in protein-protein homo-oligomerization or hetero-oligomerization, which plays an essential role in many biological processes[53].

It seems that molecular evolution reduced the SAM domain present in the CP2 subfamily to a flexible loop in the Grhl subfamily. The additional SAM domain inserted between DBD and CTD leads to a more flexible DNA sequence recognition pattern for the CP2 subfamily factors[19]. The SAM domain itself is involved in the higher oligomerization of CP2 subfamily proteins that bind DNA as tetramers. Apart from that, the SAM domain's general function in the CP2 subfamily

remains unclear, and the 3D structure of a SAM domain in the CP2 subfamily has not been reported.

1.3 Tfcp2 biological functions

1.3.1 Lineage-specific functions of Tfcp2

Tfcp2 has been reported to serve specific functions in three hematopoietic lineages: erythrocytes, T lymphocytes, and B lymphocytes[13]. Globin genes are expressed in a tissue- and development-specific manner in the erythroid environment where globin proteins are responsible for transporting oxygen. Tfcp2 was found to bind to the α-globin promoter region and stimulate globin gene transcription[54]. During mouse erythroleukemia cells differentiation, Tfcp2 is also required to induce the expression of both α- and β-globin genes[55]. Besides, Tfcp2 plays a critical role in regulating the uroporphyrinogen III synthase gene[56] and hemoglobin synthesis in erythroid cells[55].

In B lymphocytes, highly repetitive sequences located upstream of the constant region's coding sequences confer variability and specificity on antibodies[57]. Interestingly, the switch regions Sμ and Sα potentially form the Tfcp2 binding site. For example, the Sμ tandem repeat region: 5'-GAGCTGAGCTGGGGTGAGCTGAGCTGAGCTGGGGTGAG-CT-3' may form one and a half Tfcp2 DNA-binding sites[58]. Tfcp2 participates in regulating Ig heavy chain class switch recombination (CSR); disruption of the Tfcp2 DNA binding activity induced the IgM to IgA conversion in B cells[58]. In this process, Tfcp2 interacts with histone deacetylase and the repressor Sin3A to modify chromatin histone deacetylation to repress the occurrence of CSR[13,58].

Tfcp2 DNA-binding activity to cellular promoters in primary T cells is strikingly regulated by mitogenic signaling pathways during cell growth[59]. Mutational analysis demonstrated that pp44(ERK1) could specifically phosphorylate Tfcp2 at Ser29 both *in vitro* and *in vivo*, and the phosphorylation is a prerequisite for the activation of Tfcp2 DNA binding activity upon activation of resting T cells[60]. However, this phosphorylation is insufficient for activating Tfcp2 DNA-binding activity as detected in both *in vitro* and in mouse fibroblasts suggesting that in this cell type Erk phosphorylation of Tfcp2 is necessary but not sufficient and that an additional signaling event needs to cooperate with Erk to induce Tfcp2:DNA binding activity. Interleukin-4 (IL-4) was only expressed in actively dividing T cells, where Tfcp2 binds the IL-4 promoter and activates IL-4 gene expression[61]. This activation was co-regulated by other T-cell specific Tfcp2 partner proteins through a calcium-dependent signaling pathway[13,61].

1.3.2 Involvement of Tfcp2 in cancer

1.3.2.1 Role of Tfcp2 in hepatocellular carcinoma (HCC)

HCC is one of the five most common cancers worldwide[62]. Intensive studies of various aspects of HCC have been reported in recent years. Oncogene astrocyte elevated gene-1 (AEG-1) is significantly overexpressed in more than 90% of human HCC cases and plays a critical role in HCC pathogenesis[63,64]. The overexpression of Tfcp2 was significantly correlated to AEG-1[65]. AEG-1 targets the Tfcp2 promoter and enhances the expression of Tfcp2, which activates the thymidylate synthase (TS) gene to generate thymidylate, in response to treatment with 5-FU[63]. The up-regulation of thymidylate levels by AEG-1 leads to the strong resistance of HCC to chemotherapy with 5-FU[63,66].

Besides direct oncogenic mutations, the expression of Tfcp2 can also be mediated by related signaling pathways and ultimately induce human HCC. Analysis of liver cancer tissue from patients showed that the receptor protein Notch1 and Tfcp2 have a strong positive correlation in both their expression level and their biological functions[67]. Over-expression of the Notch1 intracellular domain could increase the expression of Tfcp2. Furthermore, blocking Notch signaling with its inhibitor could also repress the expression of Tfcp2. In summary, the research identified Tfcp2 as a crucial mediator of the Notch signaling pathway, and Tfcp2 mediates the Notch1 induced HCC carcinogenesis[67,68]. In addition, the cooperation between Notch and Ras signaling could also upregulate Tfcp2 expression, which again confirmed the function of Tfcp2 during human hepatocarcinogenesis[69].

Tfcp2 upregulates osteopontin (*OPN*) gene expression to mediate the aggressive progression of HCC and metastasis[65]. c-Met is a hepatocyte growth factor receptor, which plays an essential role in HCC[70,71]. Activated c-Met could initiate the epithelial-mesenchymal transition (EMT) and promote tumor cell migration and invasive growth[72,73]. As a downstream target of Tfcp2, the secreted OPN binds to the CD44 receptor, resulting in auto-phosphorylation of c-Met and subsequently in activating its main downstream PI3K/Akt signaling pathway[73,74]. Inhibition of c-Met phosphorylation disrupts Tfcp2 mediated tumorigenesis and metastasis[74].

As hallmarks of human cancer, matrix metalloproteinase (MMPs) are upregulated in almost all cancer types, and they play a major role in promoting cancer progression by stimulating the proliferation, migration, invasion and metastasis of cancer cells, and tumor angiogenesis[75,76]. Among all members of the MMP family, the function of MMP9 in regulating hepatocellular carcinoma cell migration and invasion has been well established[77]. Interestingly, Tfcp2 is also highly expressed in HCC and linked to stronger metastatic and angiogenesis potency of the

tumor[73]. The underlying molecular mechanism was investigated, and MMP9 was identified as a direct target gene of Tfcp2 through chromatin immunoprecipitation on chip experiment (ChIPSeq). Loss of function analysis showed that Tfcp2 binds to the promoter region of MMP-9, thus upregulating MMP-9 expression and facilitating angiogenesis in HCC[78]. These findings demonstrated a novel target of Tfcp2, which contributes to its carcinogenic properties.

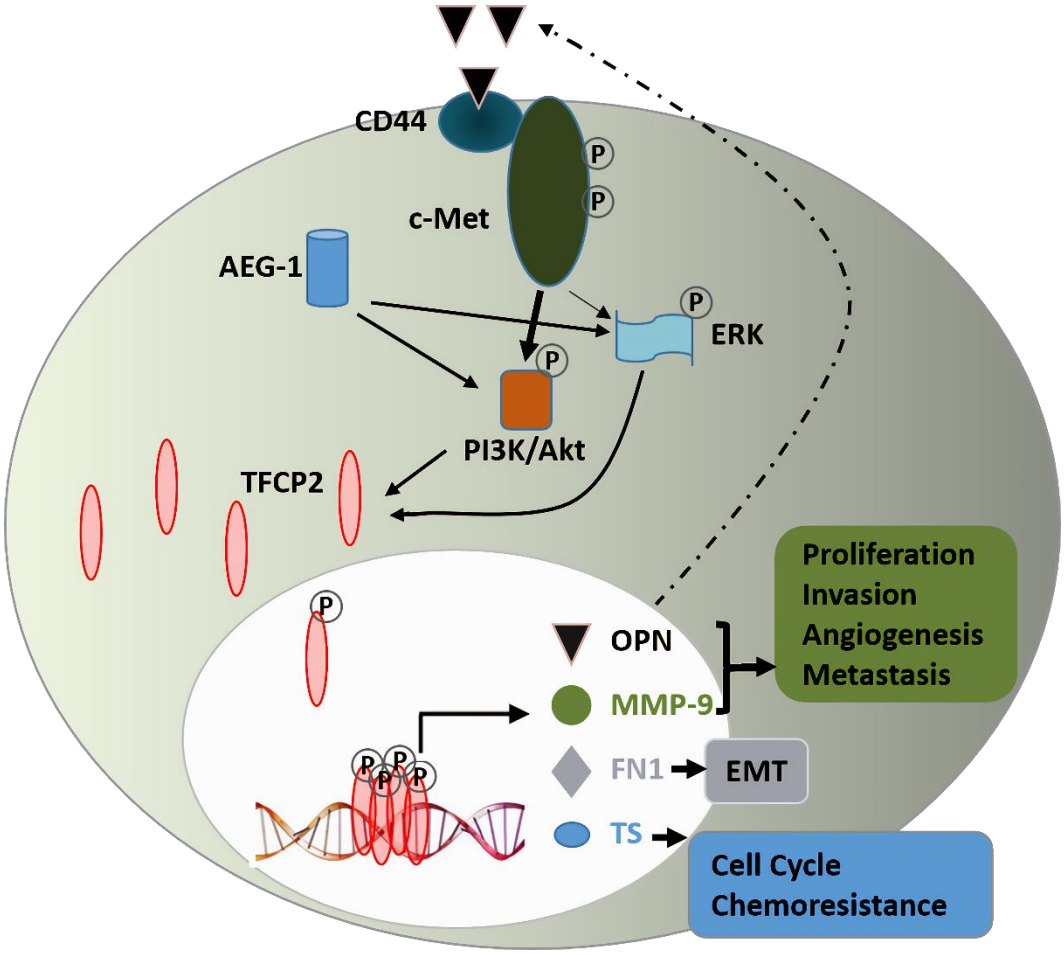

Figure 1-4. Proposed molecular mechanism of Tfcp2 involvement in hepatocarcinogenesis. Notch1 and AEG-1 promote the increase of the Tfcp2 expression level. AEG-1 also promotes Tfcp2 gene expression via PI3K/Akt and ERK signaling pathways. Tfcp2 forms a tetramer that upregulates osteopontin (OPN) and matrix metalloproteinase-9 (MMP-9), leading to cell proliferation, invasion, angiogenesis, and metastasis. The secreted OPN binds to the CD44 receptor, thereby contributing to auto-activation of c-Met, and then facilitating its downstream PI3K/Akt and ERK signaling pathways. Tfcp2 also upregulates fibronectin 1 (FN1), which is involved in the EMT. Furthermore, Tfcp2 promotes the expression of thymidylate synthase (TS), resulting in cell cycle progression and chemoresistance[73].

In addition to MMP9, during EMT the expression of fibronectin will also increase, and this increase is also a hallmark of mesenchymal cells[79]. Studies showed that Tfcp2 is involved in Snail1-induced fibronectin 1 (*FN1*) gene expression and EMT[9]. Similar to this discovery, genome-wide Tfcp2 target gene detection in HCC by ChIP-seq followed by microarray analysis for gene expression also identified *FN1* as a direct Tfcp2 target gene in HCC[80]. This identification broadened the understanding of how Tfcp2 contributes to the highly aggressive and metastatic phenotype during hepatocarcinogenesis.

In summary, Tfcp2 is a target of the oncogene product AEG1 while also being regulated by the tumor inducing Notch and Ras signaling pathways. Overexpressed Tfcp2 promotes expression of its downstream target genes including *TS, OPN, MMP-9, and FN1,* which directly or indirectly contributes to human hepatocarcinogenesis[24,73,81]. Therefore, Tfcp2 is a viable drug target, and inhibition of Tfcp2 in these signaling pathways might be exploited as a potential clinical therapeutic option for HCC (Fig. 1-4).

Even though various underlying regulatory mechanisms of Tfcp2 during human HCC have been identified as summarized above, there is still no available effective treatment directed against Tfcp2 for the chemotherapy of HCC[24,65]. Currently, as a tyrosine kinase and Raf inhibitor, Sorafenib is the only approved standard agent for the treatment of advanced HCC patients[82]. Tfcp2 has no ligand-binding domain. To directly inhibit its function requires small molecules that can specifically interfere with its protein:DNA interaction, and this is still a significant challenge. Therefore, such transcription factors are often considered as “undruggable”[19,83]. Using *in vitro* fluorescence polarization assay, large-scale screening of commercially available compounds that can repress the Tfcp2 DNA binding activity was carried out, followed by multiple electrophoretic mobility shift assays (EMSA) and *in vitro* luciferase reporter assays. Finally, factor quinolinone inhibitor 1 (FQI1) was identified as the most effective compound that can interact with Tfcp2 to suppress its DNA binding activity[82]. Further tests showed that FQI1 not only displayed anti-proliferative activity in cultured HCC cells. Furthermore, FQI1 can also dramatically inhibit the growth of HCC tumors in a mouse xenograft model without causing general tissue toxicity[24,82]. This interesting finding suggests that FQIs may be further developed into a potential drug for the treatment of HCC.

1.3.2.2 Role of Tfcp2 in breast cancer

Accounting for up to 25% of all cases, breast cancer is the leading type of cancer in women[84]. Several studies using cultured breast cancer cells indicated the important role of Tfcp2 in regulating these cells’ proliferation, invasion, and metastasis through different molecular

mechanisms. Ornithine decarboxylase (ODC) catalyzes the rate-limiting step of polyamine biosynthesis, which is essential for both cell proliferation and differentiation, and overexpressed ODC was detected in multiple types of tumors, including breast cancer[85]. Cultured MCF7 breast cancer cells showed that Tfcp2 was involved in estrogen-induced *ODC* gene expression through a cAMP-dependent pathway[85]. MicroRNAs (miRNAs) are non-coding RNA molecules; several studies indicated that miRNAs could regulate the development and metastasis of breast cancer by regulating their mRNA targets[86]. Previous studies demonstrated that miRNA-660-5p is upregulated in breast cancer patients[87]. Y.Shen *et al*. showed that Tfcp2 is a direct downstream target of the microRNA miR-660-5p[88]. A decreased expression of miR-660-5p can dramatically suppress MCF7 breast cancer cells proliferation, migration, and invasion through the repression of Tfcp2[88].

Except for cell culture studies, recent literature demonstrated the role of Tfcp2 as an oncogenic driver in basal-type and triple-negative breast cancer (TNBC) using *in vivo* xenografts and metastasis assays[89]. Gene expression profiling interactive analysis (GEPIA) was used to investigate the expression level of Tfcp2 in breast cancer tissues of 1085 patients; the results showed that Tfcp2 was overexpressed in breast tumors compared with normal tissues[90]. Gene set enrichment analysis (GSEA) suggested that Tfcp2 is well-correlated with aggressive basal type breast cancer[89]. In addition, Kaplan-Meier survival analysis indicated that increased expression of Tfcp2 demonstrated basal, luminal, and HER2 subtype patients have poor survival rates[91]. Further studies revealed a novel pathological regulatory mechanism according to which Tfcp2 could positively regulate gene expression of epidermal growth factor (EGF) and transforming growth factor-α (TGF-α) through direct binding to their promoter regions and then activating autocrine signaling through the EGF receptor (positive feedback loop), which ultimately enhanced EMT, metastasis, and stemness of breast cancer cells[89]. These findings suggested that for malignant breast cancer, Tfcp2 might be a new anti-metastatic treatment target for TNBC patients.

1.3.2.3 Role of Tfcp2 in oral squamous cell carcinoma (OSCC)

Due to its association with environmental factors, oral cancer is one of the most common cancers in Asia, with the highest incidence in South Asia[92]. Aurora kinase A (AURKA, also known as STK6) belongs to the serine/threonine kinase family and plays an essential role in centrosome function and duplication, therefore regulating the mitotic process of cell growth[93]. It was demonstrated that AURKA was overexpressed in patient tissue samples by analysis of both mRNA and protein levels[94]. As a first-line medication against type 2 diabetes, Metformin functions in repressing hepatic gluconeogenesis, therefore reducing hyperglycemia[95]. Interestingly, a current

investigation combining cell culture studies with xenograft animal model analysis showed that Metformin changed OSCC malignant behavior and suppressed its development through a Tfcp2/Aurora-A signaling pathway[96]. The investigation suggested that as a novel mediator of AURKA signaling, Tfcp2 also plays a pivotal role in human oral cancer tumorigenesis[94].

1.3.2.4 Role of Tfcp2 in other cancer types

There is evidence demonstrating that Tfcp2 is also involved in colorectal cancer (CRC), cervical cancer and ovarian cancer. Tfcp2 was also reported to link to melanoma[97] and pancreatic cancer[98] formation. However, the molecular mechanisms of Tfcp2 in these cancers currently is still unclear. Further studies are required to elucidate the molecular mechanism of Tfcp2 in these cancers.

CRC is one of the three most common types of cancer in the world, and it is characterized by a high mortality rate[99]. Primary tumor tissues isolated from CRC patients were used to investigate the expression level of Tfcp2 on both transcriptional and translational levels. Data analysis showed that increased Tfcp2 has a positive correlation with CRC tumor size and poor prognosis[100]. Compared with patients expressing Tfcp2 at low levels, the 5-year survival rates of patients with high Tfcp2 expression was dramatically reduced. The study suggested Tfcp2 is a critical mediator of CRC tumorigenesis and progression[100]. However, the pathological mechanism underlying the detected phenotype is still unclear and needs to be clarified by further studies.

Human papillomavirus infection (HPV) contributes to more than 90% of cervical cancer cases[101]. The tumor susceptibility gene 101 product (TSG101) works as a negative regulator of cell growth and differentiation, and decreased expression of the *TSG101* gene was detected in HPV positive cervical cancer cells[102,103]. Tfcp2 was detected to bind to the *TSG101* gene promoter region based on the specifically designed software Cis-element cluster finder (Cister)[103]. The expression of Tfcp2 and TSG101 was detected in patient tumor samples. Significantly increased Tfcp2 expression in HPV-positive cervical cancer cells leads to decreased expression of TSG101 and also to HPV-dependent cervical carcinogenesis[104].

Due to limitations of early diagnosis and treatment, ovarian cancer (OC) is one of the leading causes of cancer-related deaths in female patients[105]. So far, there are only a few markers available for early clinical tumor detection. Cancer antigen 125 (*CA125*) is expressed in more than 50% of early-stage ovarian cancer patients, and it is the most commonly used marker for OC diagnosis[106]. With the help of bioinformatics tools, re-analysis of the published database generated from OC patients identified multiple transcription factors that regulated OC-related gene expression. Among them, Tfcp2, which was overexpressed in OC, was detected, and this protein might be a new potential diagnostic biomarker for OC[107].

1.3.3 Biological functions of Tfcp2 in other areas

1.3.3.1 Tfcp2 in Alzheimer's disease

Alzheimer's disease (AD) is a chronic neurodegenerative disease that was identified more than one century ago[108]. The cause of the disease is poorly understood, and there is still no effective treatment to cure it. More than 30 years ago, the hypothesis of a deposition of β-amyloid (Aβ) peptide in plaques in patients' brains driving the pathogenesis of AD was proposed, and it was supported by subsequent research[109]. Amyloid precursor protein (APP) belongs to the type-I family trans-membrane proteins and can be cleaved into APP intracellular C-terminal domain (AICD) and β-amyloid (Aβ) peptide[109]. Aβ is responsible for plaque formation in AD[109]. Fe65 is a multi-domain transcription co-regulator adaptor protein[110]. A study reported that Fe65 interacts with AICD and is prevented from further nuclear translocation and induction of apoptosis[111]. Besides, it was also shown that Fe65 could interact with the Tfcp2 transcription factor with its protein-protein interaction domain, which suggested that Tfcp2 might also be related to AD[13,112]. Interestingly, the study could only show that Fe65 interacts with AICD, which will inhibit Tfcp2 transactivation of the thymidylate synthase (TS) gene followed by cell cycle progression in cultured fibroblasts but not in neuronal cell lines[113]. The study suggested that Tfcp2 has a different gene expression regulation program in neuronal cells.

In addition, it was also reported that APP plays a critical role in signal transduction, the aberration of which leads to neuronal cell apoptosis and the loss of neurons in AD brain[114]. A study with neuroblastoma cells showed that APP could effectively decrease apoptosis through activation of Tfcp2, which is a downstream target of the PI3K/AKT signaling pathway[112]. APP-mediated nuclear translocation and DNA binding activity of Tfcp2 is significantly enhanced by the increased PI3K/AKT signaling activity, and aberration Tfcp2 expression will result in neuronal loss in AD brain[112].

Epidemiological studies reveal that Tfcp2 is involved in Alzheimer's disease through dnTfcp2 (an allele of Tfcp2)[112]. The Tfcp2 allele induced lower expression and activity of Tfcp2, which promotes neuronal apoptosis[115]. The Tfcp2 mRNA 3' untranslated region is also linked to a predisposition to AD, but this phenomenon is observed in different human population cohorts[116].

1.3.3.2 Role of Tfcp2 in human immunodeficiency virus (HIV) infection

As mentioned above, Tfcp2 has an important lineage-specific function in both B cells and T cells of the human immune system[58,59]. Interestingly, a direct role of Tfcp2 in regulating HIV-1 transcription was also identified from both *in vivo* and *in vitro* assays[117,118]. *In vitro* plus-chase experiments showed that Tfcp2 prevents TFIID binding to a HIV-1 TATA promoter element and

inhibits HIV-1 transcription elongation[117]. Besides, it was reported that human transcription factor yingyang 1 (YY1) functions to inhibit HIV-1 production through repressing its long terminal repeat (LTR) transcription. *In vivo assay*, Tfcp2, together with YY1 and HDAC1, form a heteromeric nuclear protein complex that binds to the HIV-1 LTR transcription initiation site and inhibits LTR-directed gene expression and virus production[118]. The repression may result in a reservoir of latent virus in a pool of stably infected unproductive memory CD4+ cells, which might be reaching the limit of clinical detection and make virus eradication impossible[119].

1.3.4 Summary

Tfcp2 activity is involved in many cellular signal transduction pathways. These involvements may manifest itself in three ways: enhancement of Tfcp2 DNA binding activity, decrease of Tfcp2 DNA binding activity, and modification of Tfcp2 post-translational phosphorylation. Tfcp2 functions as a transcription activator mainly through an increase of Tfcp2 binding to target gene promoter regions. This up-regulation contributes to cell proliferation, which leads to various cancers. By decreased DNA-binding activity, Tfcp2 down-regulates target gene expression, which might be correlated to some diseases, such as AD and HIV. Through ERK/Akt or other signaling pathways, phosphorylation of Tfcp2 is required to activate Tfcp2, facilitating Tfcp2 tetramer formation to affect gene expression. In addition, Tfcp2 could synergistically interact with other co-factors to regulate gene expression.

1.4 Tfcp2l1 biological functions

1.4.1 Role of Tfcp2l1 in embryonic stem (ES) cells

ES cells have a remarkable ability to maintain self-renewal and pluripotency[120]. During the gene expression programs that sustain these capacities, transcription factors play crucial roles. To investigate transcriptional regulatory networks, ChIP-seq experiments with antibodies against specific transcription factors have been performed. Data analysis demonstrated how these transcription factors define the ES cell identity[121]. Among them, Tfcp2l1 was shown to interact with the core transcription factors Oct4, Nanog, Sox2, Esrrb, and Stat3 to form a transcriptional regulatory network, which acts as either activator or repressor depending on the target[122]. *Tfcp2l1* knockdown in cultured ES cells followed by qPCR analysis showed that Tfcp2l1 could repress lineage marker expression and increase expression of genes conferring pluripotency, which suggests that Tfcp2l1 plays a role in suppression of endoderm, mesoderm, and trophectoderm

specification and pluripotency maintenance of mouse ES cells[123]. It was also demonstrated that the N-terminal and CP2-like domain of Tfcp2l are necessary for ES cell self-renewal, and that Tfcp2l1 suppresses formation of the three germ layers through inhibition of Lymphoid Enhancer Binding Factor 1 (Lef1) expression, a component of the WNT signaling pathway. Besides, a study focused on a zinc finger transcription factor, Snai1, showed that Tfcp2l1 is a direct target of Snai1. Induced by retinoic acid, Snai1 binds to pluripotency gene promoters and represses their expression, thus promoting ES cell exit from pluripotency and initiation of differentiation[124,125].

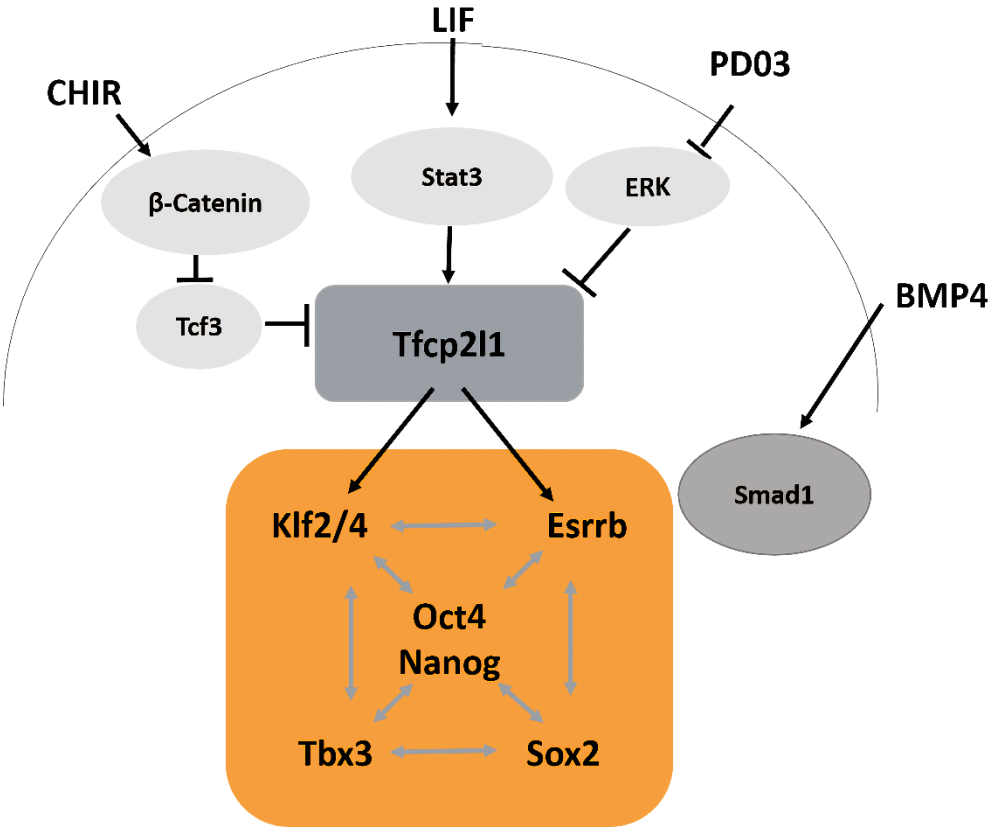

Figure 1-5. Transcriptional regulatory network in ES cells. Nanog, Oct4, Sox2, Klf4, Tbx3, and Esrrb form the core pluripotency gene regulatory network. Tfcp2l1 is the downstream target of LIF/Stat3, Wnt/β-catenin and ERK signaling pathways and is independently induced by LIF, CHIR, and PD03 through activation of Stat3, activation of Wnt/β-catenin and repression ERK pathways, respectively. BMP4 impacts the regulatory network through the Smad1-Sox2 regulatory pathway.

Besides the nuclear transcription factors, extrinsic growth factors from the environment, also play essential roles in the maintenance of the ES cells' pluripotency by activating specific signaling pathways. Among them, the leukemia inhibitory factor (LIF) and bone morphogenetic protein 4 (BMP4) signaling pathways were the first to be identified[126]. Interestingly, further study showed that these signaling pathways could integrate into an Oct4, Sox2, and Nanog regulatory network

via Smad1 and STAT3[127]. Transcriptome analysis of Stat3 null ES cells identified a group of targeted signaling genes. Among them, *Tfcp2l1* was the top regulated gene. Overexpression and knockdown of *Tfcp2l1* also showed a similar regulatory effect compared with LIF/Stat3 signaling. Therefore, it was reconfirmed that *Tfcp2l1* is a downstream target of the LIF/Stat3 signaling pathway; overexpression of *Tfcp2l1* promotes ES cell self-renewal[128,129] (Fig. 1-5).

Except for transcription factors and signaling pathways, the self-renewal and pluripotency ability of ES cells is also controlled by genetic factors and chromatin state[130]. Silencing histone H3 Lys9 (H3K9) dimethylation and trimethylation leads to ES cell exit from self-renewal and initiation of differentiation[130,131]. Jumonji domain containing 1A (Jmjd1a) is the main demethylase targeting H3K9Me2 and H3K9Me3. It was reported that Jmjd1a specifically demethylates H3K9Me2 at the promoter of Tfcp2l1, positively regulates its expression, and maintains the ES cells in a pluripotent state. The study demonstrated an important role of Tfcp2l1 in regulating ES cells self-renew from the chromatin modification level[132].

1.4.2 Role of Tfcp2l1 in the kidney

In the development of nephrons, mesenchymal stem cells (MSCs) give rise to renal epithelial cells (NREs) through the mesenchymal-epithelial transition (MET)[133,134]. In the reverse process, MSCs can be generated from NREs through the epithelial-mesenchymal transition (EMT)[134]. With aberrant EMT and adipogenic trans-differentiation, NREs develop into clear cell renal cell carcinoma (ccRCC)[135]. Tfcp2l1 and three additional transcription factors, GATA binding protein 3 (GATA3), transcription factor AP-2 beta (TFAP2B), and doublesex and Mab-3 related transcription factor 2 (DMRT2) are significantly down-regulated in ccRCC, which indicated that Tfcp2l1 might play a protective role in inhibiting ccRCC[135]. A further study using miRNA profiling in ccRCC, showed that miR-489 is notably upregulated and able to directly bind to the Tfcp2l1 3'-UTR and repress its expression, which suggested that the microRNA might be an upstream regulator of Tfcp2l1. Therefore, both Tfcp2l1 and miR-489 play an important role in ccRCC[136].

In addition to nephron development, Tfcp2l1 is required for kidney duct maturation. With the help of a mouse model, a truncated form of Tfcp2l1 with incomplete DNA-binding domain was generated by a homozygous gene-trap insertion in the *Tfcp2l1* locus. The *Tfcp2l* $^{tra/tra}$ mutant mice showed a defective maturation of the collecting duct and died within two days after birth[137]. The results demonstrated thatTfcp2l1 plays a critical role in maintaining the proper physiological function of the kidney. However, the underlying molecular mechanism is still unclear and needs

further investigation. Recent research showed that the collecting duct has different developmental stages with different cellular identities and that this tissue was composed mainly of two different cell types, the acid-base regulator intercalated (IC) cells and salt-water regulator principal (PC) cells[138]. Basically, at around E13 only PC marker proteins can be detected, followed by PC and IC double-positive cells from E15-E18. During postnatal development, these cells will acquire their identities and distribute in a rosette-like pattern. During this process, the deletion of Tfcp2l1 leads to an induction of PC cell markers expression. Further analysis showed Tfcp2l1 induces gene expression in ICs, which includes Jag1, a ligand of the Notch signaling pathway. Increased Jag1 expression stimulates Notch signaling in the adjacent PC cells; a combination of Tfcp2l1 with the Notch target gene *Hes1* defines these cells' identity. In summary, Tfcp2l1 induces gene expression in ICs and plays a critical role in regulating collecting duct progenitor plasticity. Deletion of Tfcp2l1 in the kidney led to the loss of IC and PC cell identities and pattern[138,139].

1.4.3 Role of Tfcp2l1 in breast cancer

Landemaine and colleagues reported that *Tfcp2l1,* together with five other genes (*DSC2, UGT8, ITGB8, ANP32E and FERMT1*), was a prognostic marker for lung metastasis of breast cancer[140]. However, further analysis of Landemanie's data by Aedin showed that the "six-gene" signature could be used to predict breast cancer subtypes, but not lung metastasis[140,141]. Using bioinformatics tools to compare the expression pattern of *Tfcp2l1* in a specifically induced WAP-T mouse model with human breast tumors showed that *Tfcp2l1* gene expression was correlated with a basal-like subtype of breast cancer[142].

1.4.4 Role of Tfcp2l1 in other epithelial carcinomas

Ducts are essential epithelial features of tubular organs, such as the salivary and mammary glands, the ovaries and thyroid[137]. Similar to renal ducts, salivary glands also require Tfcp2l1 to maintain their physiological functions[137]. A recent study reported that Tfcp2l1 works synergistically with five prognostic genes (*LRRC8D, BMBR1, EPOR, PARS2, and TTC30A*) to promote the differentiation of ovarian cancer stem cells. Therefore, Tfcp2l1 is a clinical marker for ovarian cancer[143]. DNA microarray analysis from five papillary thyroid cancer samples was performed by Hyun and colleagues, and it was reported that Tfcp2l1 was downregulated in thyroid cancer[144]. However, the mechanism by which Tfcp2l1 regulates or contributes to thyroid cancer is still unclear.

1.4.5 Post-translational modification of Tfcp2l1

Tfcp2l1 was initially characterized as a transcriptional repressor, while Sarah and colleagues reported that Tfcp2l1 could also act as a transcription activator. The activating activity is regulated by sumoylation of the its Lys30 residue in Tfcp2l1[43]. A recent study by Jinbeom and colleagues reported that Tfcp2l1 Thr177 could be phosphorylated by Cdk1, and that this modification is correlated with ES cell pluripotency and differentiation. In addition, the Tfcp2l1-CDK1 pathway was associated with bladder carcinogenesis, and wild-type Tfcp2l1 will impair the tumorigenic potency of bladder cancer cells[145].

1.4.6 Summary

The down-regulated expression of Tfcp2l1, a transcriptional repressor, facilitates cell proliferation, which contributes to ccRCC and thyroid cancer. Tfcp2l1 is a downstream target of multiple signaling pathways, and active Tfcp2l1 is involved in the core transcriptional regulatory network to maintain cell pluripotency and stemness. Interestingly, through Thr177 sumoylation, Tfcp2l1 may function as a transcriptional activator, thereby providing new insight into the function Tfcp2l1.

1.5 Ubp1 biological function

Compared to Tfcp2 and Tfcp2l1, the biological functions of Ubp1 are not well documented. Ubp1 has been reported to serve an important function in the regulation of extraembryonic angiogenesis[146]. *Ubp1* knockout mice displayed growth retardation at 10.5 dpc due to a deficiency in allantoic blood vessel development[146]. Probably due to the failure to connect the yolk sac vasculature with the surrounding vascular network, the *Ubp1*$^{-/-}$ mice died at 11.5 dpc[146].

As described before, Ubp1 acts as a transcriptional repressor or activator depending on promoter context[35,147]. Ubp1 binds strongly to the HIV-1 initiation site, which inhibits the core factor TFIID binding to the TATA box resulting in repression of HIV-1 transcription[35]. Ubp1 binds to the *P450scc* gene promoter region (-155/-133) and upregulates P450scc expression[31]. This up-regulation could be repressed by Tfcp2l1. In addition, Ubp1 can recognize the Tfcp2 consensus sites and compensate for the loss of Tfcp2 expression in erythroid cells[37].

Both Tfcp2 and Ubp1 are ubiquitously expressed. Furthermore, Tfcp2 and Ubp1 can bind the same target DNA motif. Ubp1 might synergistically interact with Tfcp2 to regulate α-globin gene expression. More studies are required to explain the function of Ubp1.

1.6 Aim of the thesis

In the CP2 subfamily of transcription factors, Tfcp2 is a pro-oncogenic factor involved in a wide variety of cancers. In addition, Tfcp2 plays a critical role in the regulation of cell cycle progress and is linked to several diseases. Meanwhile, Tfcp2l1 plays critical roles in embryonic stem cell pluripotency maintenance[128] and self-renewal[148]. Tfcp2l1 was also implicated in various kinds of cancers. Therefore, the CP2 family members could act as transcriptional activators and repressors depending on the promoter context. However, the molecular mechanisms that enable CP2 members to act as transcriptional repressors or activators in different tissues or development stages are not completely understood. Therefore, I aimed to analyze the three-dimensional structures of CP2 subfamily members by X-ray crystallography and to apply biochemical and biophysical methods to elucidate the molecular basis of their DNA interaction and target sequence recognition. The project will address the following topics.

Aim 1. Structure analysis of the Tfcp2l1 and Tfcp2 DNA binding domains.

Aim 2. Structure analysis of specific DNA motif binding by Tfcp2 and Tfcp2l1.

Aim 3. Structural basis for oligomerization of the intact Tfcp2 and Tfcp2l1 homologs.

Aim 4. How do mutations in the DNA-binding domain influence DNA binding?

2. MATERIALS AND METHODS

2.1 Materials

Antibodies

Antibody	Description	Manufacturer
Anti-penta-His (mouse)	Primary, 1: 1 000	Qiagen
Anti-mouse, IgG (H and L)	Secondary; 1: 10 000, HRP-linked antibody,	CST

Antibiotics

Antibiotic	Stock concentration	Working concentration
Carbenicillin (Carb)	100 mg/ml in 50% ethanol	100 µg/ml
Chloramphenicol	34 mg/ml in 100% ethanol	34 µg/ml
Kanamycin (Kan)	50 mg/ml in water	50 µg/ml

Bacterial strains

Bacterial strain	Genotype	Manufacturer
E. coli DH5α $T1^R$	F^- Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rk^-, mk^+) *pho*A *sup*E44 λ^- *thi*$^-$1 *gyr*A96 *rel*A1	Invitrogen, Carlsbad, USA
E. coli Rosetta 2 BL21 (DE3) $T1^R$	F- *ompT hsdS*B(rB- mB-) *gal dcm* (DE3) pRARE2 (CamR), containing the tRNA genes *argU, argW, ileX, glyT, leuW, proL, metT, thrT, tyrU and thrU*	Novagen, Darmstadt, D
*E.coli.*BL21(DE3) pLysS	F^- *omp*T *hsd*S_B($r_B^-m_B^-$) *gal dcm* (DE3) pLysS (Cam^R)	Thermo Fisher

Chemicals

Chemical	Description	Manufacturer
Agar	BD Difco agar, granulated	Thermo Scientific

Chemical	Description	Manufacturer
Ampicillin	≥ 99%	Roth
Bicine	≥ 98%	Roth
Bis-Tris	≥ 99%	Roth
$CaCl_2$	Dihydrate, pro analysis	Merck
Carbenicillin	≥ 88%, disodium	Roth
Chloramphenicol	≥ 98.5%	AppliChem
CHES	≥ 99%	Roth
Dioxane	≥ 99.8%	Sigma-Aldrich
DTT	DTT BioChemica	AppliChem
Ethanol	96%, Ph. Eur.	Roth
Glycerol	86% p.a. Rotipuran®	Roth
	99% p.a. Rotipuran®	Roth
HCl	Hydrochloric acid fuming 37%	Roth
HEPES	99.5% p.a.	Roth
Imidazole	≥ 99% p.a.	Roth
IPTG	≥ 99% (TLC), ≤0.1% dioxane	Sigma-Aldrich
Isopropanol	≥ 99.95% LC-MS-grade	Roth
Kanamycin	≥750 IU / mg	Roth
KCl	≥ 99% Ph. Eur.	Roth
KH_2PO_4	≥ 99% p.a.	Roth
K_2HPO_4	≥ 99% p.a.	Roth
LiCl	≥ 99%, p.a.	Roth
Methanol	≥ 99%	Roth
$MgCl_2$	Hexahydrate, ≥ 98% Ph. Eur.	Roth
MES	≥ 99%	Sigma-Aldrich
MOPS	≥ 99.5%	Sigma-Aldrich
NaAcetate	≥ 99.5% (NT)	Sigma-Aldrich
Na_3citrate	Dihydrate, Ph. Eur.	Merck
NaCl	≥ 99.8% p.a.	Roth
NaOH	≥ 99% p.a.	Roth
$(NH4)_2SO_4$	≥ 99.5%	Roth

Chemical	Description	Manufacturer
PEG400	Ph. Eur.	Merck
PEG1500	Ph. Eur.	Merck
PEG3350		Sigma-Aldrich
PEG4000		Sigma-Aldrich
PEG6000		Sigma-Aldrich
PEG8000		Sigma-Aldrich
PIPES	≥ 99%	Roth
TCEP	98% p.a.	Roth
Tris	Pure, pharma grade	AppliChem
Triton-X-100	non-ionic, aqueous solution	Roche
Tryptone	Tryptone BioChemica	AppliChem
Tween-20	viscous liquid	Sigma-Aldrich
Yeast extract	powdered, for bacteriology	Roth

Crystallization screen kits

Screen kit name	Manufacturer
Additive Screen HT	Hampton Research
AmSO4 Suite	Qiagen, Hilden, D
Basic HTS	Jena Bioscience, Jena, D
Cations Suite	Qiagen, Hilden, D
Classics Suite	Qiagen, Hilden, D
Classics II Suite	Qiagen, Hilden, D
Classics Lite Suite	Qiagen, Hilden, D
ComPAS Suite	Qiagen, Hilden, D
JBS-JCSG	Jena Bioscience, Jena, D
MPD Suite	Qiagen, Hilden, D
Nuc-Pro	Jena Bioscience, Jena, D
PACT Suite	Qiagen, Hilden, D
PEG Suite	Qiagen, Hilden, D
PEG II Suite	Qiagen, Hilden, D

Screen kit name	Manufacturer
pHClear Suite	Qiagen, Hilden, D
pHClear II Suite	Qiagen, Hilden, D
ProComplex Suite	Qiagen, Hilden, D

Enzymes

Enzyme	Usage	Manufacturer
BamHI/BahHI-HF	Cloning	NEB
DNase I	Protein purification	Roche
DpnI	Mutagenesis	NEB
NotI/NotI-HF	Cloning	NEB
Phusion HF polymerase	Initial mutagenesis and fusion PCR	NEB
Taq polymerase	Colony PCR	Roboklon/EUR$_X$
TEV protease	Protein tag cleavage	Heinemann's lab
T4 DNA ligase	Cloning	NEB
XhoI	Cloning	NEB

Instruments

Instrument	Type	Manufacturer
Agarose gel chamber	HG370, HG330	Savant
Agarose gel imaging system	GelDoc XR+	BioRad
Blotting device	Mini Trans-Blot ®Cell	Bio-Rad
Cap and vial	CrystalCap HT	Hampton Research
CD spectrometer	Chirascan	Applied Photophysics
Centrifuges	Avanti-J26 XP	Beckman Coulter
	Biofuge stratos	Heraeus
	5417R	Eppendorf
Chromatography columns	HisTrap FF 5ml	GE Healthcare
	HiTrap SP FF 5ml	GE Healthcare
	HiTrap Q HP 5ml	GE Healthcare
	Superdex 75 10/300 GL	GE Healthcare

	Superdex 200 10/300 GL	GE Healthcare
	Superdex S75 HiLoad 16/60	GE Healthcare
	Superdex S200 HiLoad 16/60	GE Healthcare
Chromatography system	Äkta Explorer	GE Healthcare
	Äkta Pure	GE Healthcare
Concentrators	Amicon Ultra	Millipore
Column matrix	Ni-NTA agarose matrix	Qiagen, Hilden, D
Cryo-loops	Mounted CryoLoop	Hampton Research
Crystal Clear sealing film	HR3	Hampton Research
Crystallization plates	Crystalquick plate, 96 well	Greiner Bio-one
	24-well hanging drop crystallization plate	MiTeGen
	In Situ-1 crystallization plate	MiTeGen
	INTELLI-PLATE, 96-3 well	Art Robbins Instruments
Crystallization robot	Gryphon	Art Robbins Instruments
Crystallization storage and observation system	Rock Imager 1000 (4 °C and 20 °C)	Formulatrix
Diffractometer	XcaliburTM Nova O	Oxford Diffraction
Disposable cuvettes	PMMA	Brand
Electrophoresis power supply	Power Pac 300	BioRad
Finescreen designer	Rock maker	Formulatrix
Finescreen dispensing system	Formulator 16sp	Formulatrix
Fluidizer	Microfluidizer	Microfluidics
Gel observation system (protein gels)	LAS 400	Fujifilm
Incubator	MIR-153 (1.5 ml, 2 ml)	SANYO
Isothermal titration calorimeters	VP-ITC	GE Healthcare
	PEAQ-ITC	Malvern Panalytical
Microscope	Wild M3C/Wild M420	Leica
Nanodrop	ND 1000 spectrophotometer	Peqlab
Nanodrop™ one	ND-ONE-W	Thermo Scientific
Native PAGE	NativePAGE Bis-Tris gels	Thermo Scientific

Instrument	Type	Manufacturer
Peristaltic pump	P-1	Pharmacia
pH-meter	FiveEasy	Mettler Toledo
Pipettes	10 µl, 20 µl, 200 µl, 1 ml	Gilson / Eppendorf
Power supply	Power RAC300	Bio-Rad
	Power Rack P25	Biometra
RALS system	VE3580 RI Detector	Viscotek
Rotator	neoLab Rotator	neoLab
Rotors	JA-25.50	Beckman Coulter
	JLA 8.1000	Beckman Coulter
Shaker incubator	Innova	New Brunswick Scientific
Small shaker incubator	HT	Infors
Sonicator	Typ GM 2200 (HD2200) with Sonotrode UW 2200 and titan plate TT13	Bandelin Sonoplus
Thermal block	Thermomixer 5437	Eppendorf
Thermocyclers	C1000 Touch	BioRad
	PTC-200	MJ Research
Vortex mixer	7-2020	NeoLAB
Water filtering system	Milli-Q® Academic	Millipore

Internet database and source

Name	Website
Uniprot	https://www.uniprot.org/
PSIPRED workbench	http://bioinf.cs.ucl.ac.uk/psipred/
XtalPred-RF	http://xtalpred.godziklab.org/XtalPred-cgi/xtal.pl
JASPAR	http://jaspar.genereg.net/
OligoEvaluator	http://www.oligoevaluator.com/LoginServlet

Ladders

DNA standard	Manufacturer
Perfect™ 100-1000 bp DNA Ladder	EURx

PerfectTM Plus 1kb DNA ladder	EURx

Protein standard	**Manufacturer**
Precision Plus Protein™ Unstained Protein Standards	BioRad
Precision Plus Protein™ Dual Color Standards	BioRad
Pierce™ Unstained Protein MW Marker	Thermo Fisher

Plasmids

Plasmid	**Property**	**Manufacturer**
pQlinkH	N-terminal polyhistidine tag	Heinemann's lab
pQlinkG	N-terminal GST tag	Heinemann's lab
pET28a-C-His	C-terminal polyhistidine tag	Heinemann's lab

Reagent kits

Kit	**Manufacturer**
GeneJET Gel Extraction Kit	Thermo Scientific
GeneJET Plasmid Miniprep Kit	Thermo Scientific
GeneJET PCR Purification Kit	Thermo Scientific
Thermofluor Fundament Kit CS-332	Jena Bioscience
JBScreen Buffers Kit CS-214	Jena Bioscience

Software

Function	**Name**	**Manufacturer**
Processing of X-ray diffraction data	CCP4	Rutherford Appleton Laboratory
Model refinement against X-ray diffraction data	Phenix	University of Cambridge, Duke University, LANL, LBNL
Protein purification	ÄKTA pure 25	GE Healthcare
Statistical analyses	GraphPad Prism 5	Prism
Image processing	Photoshop CS5	Adobe system

Function	Name	Manufacturer
Data analysis and text processing	Excel/Word/Powerpoint	Microsoft Office
Gene sequence operation	SnapGene	GSL
Structure visualization	Pymol	DeLano Scientific LLC

2.2 Molecular biological methods

A proper DNA construct is the prerequisite for recombinant expression of proteins for structural and biochemical analysis. The following experiments were performed to construct expression plasmids and bacterial strains based on the standard cloning protocol[149]. In the first step, primers were designed for the polymerase chain reaction (PCR) to amplify the target DNA. Secondly, both target DNA and the vector were digested by restriction enzymes. Then the sample was subjected to agarose gel electrophoresis and the pure digested DNA and vector were extracted from the gel and subsequently used for ligation. *Escherichia coli* cells were transformed with the recombinant plasmid, colonies were picked, and then plasmids were extracted and validated by sequencing.

2.2.1 Polymerase chain reaction (PCR)

Phusion High-Fidelity (HF) polymerase was used for PCR product amplification according to the supplier's instruction.

Reagent	**50 µl reaction**	**Final concentration**
Nuclease-free water	to 50 µl	
5X Phusion HF or GC buffer	10 µl	1X
10 mM dNTPs	2 µl	200 µM
10 µM forward primer	2.5 µl	0.5 µM
10 µM reverse primer	2.5 µl	0.5 µM
Template DNA	**	200 ng
Phusion DNA polymerase	0.5 µl	1.0 units/50 µl

**The volume of the template DNA solution is adjusted to the final concentration of 200 ng/50 µl.

Step	Temperature	Time	Cycle
Initial denaturation	98 °C	30 s	1
Denaturation	98 °C	10 s	
Annealing	55-62 °C ***	30 s	34
Extension	72 °C	1kb/30 s	
Final extension	72 °C	7 min	1
Hold	4 °C	∞	

***The primer annealing temperature is five degrees lower than the primer melting temperature.

2.2.2 Agarose gel electrophoresis

PCR products were analyzed by agarose gel electrophoresis. Based on the standard protocol, a 0.8% agarose gel was prepared by dissolving 0.48 g agarose into 60 ml TAE buffer (40 mM Tris, 20 mM Acetic acid, 1mM EDTA), supplemented with 0.5 µg/ml ethidium bromide. PCR product mixed with 6X DNA purple loading dye was loaded onto a 0.8% agarose gel. The electrophoresis was carried out 30 min at 120 V. The agarose gel was visualized by UV illumination, and PCR product size was determined by comparison to the standard DNA ladder.

2.2.3 DNA purification

The GeneJET Gel Extraction Kit (Thermo Scientific) was employed to purify the DNA fragment. After agarose gel electrophoresis, the DNA product band was excised from from the gel, and DNA was purified according to the supplier's instruction. DNA purification was performed after endonuclease digestion of DNA fragment and vector. The GeneJET PCR Purification Kit (Thermo Scientific) was used for DNA purification.

2.2.4 DNA digestion

The purified DNA fragment and the vector were digested by the proper restriction endonucleases. The reaction buffer for the double enzyme digestion was chosen to maximize enzyme activity. The digestion reaction was performed at 37 ºC for 2.5 h, and the digested DNA fragment was purified according to the DNA purification (Section 2.2.3).

2.2.5 Ligation

The digested DNA fragment and the vector were mixed at a molar ratio 4: 1 and ligated by T4 DNA ligase. The reaction volume was 10 µl, and the ligation reaction was performed at room temperature (RT) for 1 h according to the supplier's instruction.

2.2.6 Transformation

The ligation product (100-200 ng) or the plasmid (50-100 ng) was added to the 50 µl culture of competent cells for transformation. The cells were then incubated on ice for 30 min, followed by a heat shock at 42 °C for 90 s and immediately by incubating the cells again on ice for 2 min. 500 µl SOB medium was added to the cells before incubation at 37 °C for 1 h with shaking at 180 rpm. The incubated culture was centrifuged for 2 min at 4000 rpm, and 350 µl supernatant medium was removed. The resuspended cells were plated on agarose with appropriate antibiotics and incubated at 37 °C overnight.

2.2.7 Colony-PCR

Before purifying the plasmid from the transformed colony, colony-PCR was used to identify successfully transformed colonies. Taq polymerase (NEB) was applied for the PCR reaction. PCR reaction products were identified by agarose gel electrophoresis, and positive colonies were picked and cultivated for plasmid extract.

Reagent	**50 µl reaction**	**Final concentration**
Nuclease-free water	to 50 µl	
10X buffer	5 µl	1X
10 mM dNTPs	2.5 µl	200 µM
10 µM forward primer	2 µl	0.5 µM
10 µM reverse primer	2 µl	0.5 µM
Template DNA		##
Taq polymerase	0.5 µl	1.0 units/50 µl PCR

##: The template DNA is the transformed colony culture with 2 µl.

Step	**Temperature**	**Time**	**Cycles**
Initial denaturation	95 °C	3 min	1
Denaturation	95 °C	30 s	
Annealing	50 °C	30 s	34
Extension	72 °C	1 kb/min	
Final extension	72 °C	7 min	1
Hold	4 °C	∞	

2.2.8 Plasmid extraction and sequencing

The overnight culture was used for plasmid extraction using the GeneJET Plasmid Miniprep Kit (Thermo Scientific). Plasmid concentration was determined photometrically in the NanoDrop, and

quality was checked by agarose gel electrophoresis before sending the plasmid sample for sequencing.

2.2.9 Fusion PCR

Fusion PCR was applied to join two DNA fragments together[150]. There are three main steps: 1) General PCR is carried out to amplify DNA fragments F1 and F2. 2) DNA fragments F1 and F2 are fused to one gene. 3) The fusion gene is used as template DNA for general PCR. Four primers are needed to fuse two DNA fragments. The two fusion primers should overlap by at least 30 bp.

Step 1. According to general PCR as described in sections 2.2.1, 2.2.2, and 2.2.3, two DNA fragments, F1 and F2, were amplified and purified.

Step 2. F1 and F2 fusion to one gene.

Component	**30 µl reaction**	**Final concentration**
Nuclease-free water	31 µl	
5X Phusion HF buffer	10 µl	1X
10 mM dNTPs	2.0 µl	200 µM
Phusion DNA polymerase	0.5 µl	1.0 units/50 µl PCR
Fragment1		150-160 ng
Fragment2		150-160 ng

Note: No primers in this reaction.

Step	**Temperature**	**Time**	**Cycles**
Initial denaturation	98 °C	30 s	1
Denaturation	98 °C	10 s	
Annealing	65 °C	2.5 min	10
Extension	72 °C	5 min (1 kb/30 s)	
Final extension	72 °C	7 min	1
Hold	4 °C	∞	

The fusion PCR product was analyzed according to section 2.2.2. 8 µl fusion sample allowed detection of the fused gene band on an agarose gel. The concentration of the fusion gene was determined by comparison to the DNA marker.

Step 3. The fusion gene was used as template DNA and the final DNA construct was then obtained according to sections 2.2.1 to 2.2.8.

2.2.10 Site-directed mutagenesis

Site-directed mutagenesis was performed according to the general PCR in section 2.2.1. In order to obtain optimal results, the design of the mutant primers followed these rules: (1) each mutagenesis changed between one to three base pairs at one location; (2) the desired mutation should be included in both primers; (3) the 5'-primer end should be at least four base pairs away from the mutation site; (4) the 3'-primer end should be at least eight base pairs away from the mutation site; (5) the 3'-primer end should comprise at least eight non-overlapping bases[151,152].

After the PCR reaction, 1 µl DpnI restriction enzyme solution was directly added to the reaction tube. The reaction was incubated at 37 °C for 3 h or at 4°C overnight. The template DNA was digested by DpnI, and the mutation product was purified and transformed into competent DH5α cells. The mutation was validated by DNA sequencing.

2.3 Protein expression and purification

2.3.1 Recombinant protein expression test

A prokaryotic expression system was applied to produce recombinant protein by transformation of *E. coli* Rosetta 2 (DE3) $T1^R$ cells with the DNA construct. The transformation was done according to section 2.2.6. A single colony was picked and cultivated in 4 ml lysogeny broth (LB) medium with proper antibiotics at 37 °C overnight. The overnight pre-culture was injected into 100 ml LB medium and further cultivated until the OD_{600} reached the value of 0.6. 0.5 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to induce the culture at 18 °C. After growing the culture overnight it was centrifuged at 4 °C and pellets were stored at -80 °C.

The *E. coli* pellets were suspended in 1 ml lysis buffer supplemented with 20 µl lysozyme solution (final working concentration 1 mg/ml) and incubated at 4 °C for 30 min. The freeze-thaw method was applied to lyse the cells. The supernatant was separated from the cell debris by centrifuging 13000 rpm for 30 min at 4 °C. The supernatant was transferred to a new Eppendorf tube, and 20 µl equilibrated Ni-NTA agarose beads were added to pull down the target protein. The beads were collected by centrifugation at 500 rpm for 5 min at 4 °C. The target protein was visualized by SDS-PAGE after elution in elute buffer.

2.3.2 Protein expression

For the large-scale culture, cells were cultivated in 100-200 ml LB medium. The pre-culture was injected into 2 l of two-fold LB medium (volume ratio of 2.5%). After OD_{600} reached 1.0 in two-fold

LB medium, 0.5 mM of IPTG was added to induce protein expression. Cells were grown at 18 °C overnight the culture was centrifuged at 4 °C and the pellet stored at -80°C.

2.3.3 Nickel affinity chromatography

1 g of cell pellets were suspended in 5 ml His-lysis buffer (Appendix D, Table D1) at 4 °C and lysed by lysed by sonication. After high-speed centrifugation at 18000 rpm at 4 °C for 1 h the supernatant was collected for protein purification. The supernatant sample was loaded to a Ni^{2+}-nitrilotriacetic acid (NTA) column, which was pre-equilibrated with 10 column volumes (CV) of the buffer solution (Table D1). The column was washed out with wash buffer of increasing imidazole concentration (30 mM, 40 mM, and 50 mM imidazole), and the target protein was eluted with elute buffer. Samples from each step were tested by SDS-PAGE.

2.3.4 His-tag cleavage

The N-terminal His-tag was cleaved from the protein with tobacco etch virus (TEV) protease, a target sequence for which was introduced by cloning. Protein fractions eluted from the Ni^{+2}-affinity chromatography were collected, TEV protease was added at a ratio of 1: 5 (1 mg TEV for 5 mg His-tag protein), and the mixture was dialyzed against dialysis buffer at 4 °C overnight. The protein solution was then loaded onto a pre-equilibrated Ni-NTA column, and the target protein without His-tag was collected from the flow-through, while the TEV protease and residual protein with His-tag remain bound to the Ni^{+2} beads, separating the tag-free target protein from contaminating proteins and the TEV protease.

2.3.5 Ion exchange chromatography

The target protein was further purified by anion or cation-exchange chromatography to remove nucleic acid (DNA and RNA) contaminations. The salt concentration of the protein solution was adjusted to 100 mM with the dilute buffer (volume ratio 1: 1) and the diluted protein solution loaded to a pre-packed 5 ml SP or Q column at a speed of 2 ml/min. The column was washed by a linear increasing salt gradient. Protein fractions from the eluted peaks were tested by the SDS-PAGE.

2.3.6 Size-exclusion chromatography (SEC)

Homogeneous protein samples were separated by SEC. Protein fractions from anion or cation exchange chromatography were pooled and concentrated. The concentrated protein sample was loaded to a Superdex Hiload 200 16/60 (GE Healthcare) (molecular weight range 10 - 600 kDa) or Superdex Hiload 75 16/60 column (GE Healthcare) (molecular weight range 3 - 70 kDa) in SEC buffer (Table D1). Samples from each step were tested by SDS-PAGE. Finally, purified proteins were concentrated, flash-frozen in liquid nitrogen, and stored at -80 °C.

2.4 Biochemical and biophysical methods

2.4.1 Protein and protein:DNA complex concentration

Protein residues that contain aromatic rings (tryptophan and tyrosine) are the primary reason for UV light absorption at 280 nm[153,154]. Protein concentrations were determined in a Nanodrop spectrophotometer measuring the light absorbance at 280 nm according to Beer-Lambert's law $A = \varepsilon l C$[155] *(A, absorbance; ε, molar extinction coefficient; l, length of light path; C, concentration of sample)*. The molar extinction coefficient was calculated based on the online program Protparam (https://web.expasy.org/protparam/).

Nucleic acids have a strong absorption at 260 nm, and this property was employed to detect and quantitate DNA[156]. The pure protein without DNA contamination usually has an A260/A280 ratio of less than 0.6. Protein:DNA complexes have a different ratio (> 0.6) depending on the amount of DNA in the complex. Final protein:DNA complex concentrations were measured based on the extinction coefficient of the protein.

2.4.2 SDS-PAGE

Sodium dodecyl sulfate polyacrylamide-gel electrophoresis (SDS-PAGE) was used to assess protein quality and purity. The separation gel of 12% or 15% acrylamide was chosen depending on the protein molecular weight range to be analyzed. The protein samples were mixed with a four-fold SDS-PAGE loading buffer (Table E), and the electrophoresis was carried out at 240 V for 30 min.

After the electrophoresis, the gel was stained following a quick-staining protocol: 1) add 2-3 ml of staining solution I to the gel and microwave for 30 s; 2) rinse with water; 3) add 2 ml of staining solution II and supplement with 1 ml Coomassie Blue solution. The protein bands could be visualized after 15 min, and then the gel image was taken on a Fujifilm LAS 4000 system.

2.4.3 Western blot

Western blots were used to validate the identity of the recombinant protein[157]. Following section 2.4.2 on SDS-PAGE, the gel was transferred to the PVDF membrane to assemble the sandwich. The gel was placed towards the negative and the membrane towards the positive side of the sandwich. The assembled sandwich was run in the wet transfer buffer for 1 h at 60 V to transfer the protein to the PVDF membrane. The PVDF membrane was taken out and blocked in 5% skimmed milk in TBST buffer for 1 h at room temperature. Then the blot was washed for 5 min with TBST and incubated in TBST plus 5% skimmed milk plus primary antibody solution for 1 h at room temperature or overnight at 4 °C. After incubation with the primary antibody, the blot was

washed 3 times with TBST for 5 min incubated in TBST plus secondary antibody solution for 1 h at room temperature. This was followed by three wash steps with TBST, each time for 5 min. Finally, a detection reagent was added to the blot to detect the secondary antibody (a two-component HRP substrate). The PVDF membrane was imaged by the LAS 400 for secondary antibody detection.

2.4.4 Thermal shift assay (TSA)

Thermal shift assays were employed to measure protein stability under varying buffer conditions[158]. During sample heating, the thermal denaturation temperature was measured by real-time PCR. A higher denaturation temperature indicates that the protein is more stable in the chosen buffer condition. Protein concentrations were adjusted in the range from 0.1 – 1 mg/ml. Samples were prepared on ice.

The program setting were as follows: In the initial step the temperature is held constant at 20 °C for 2 min, then the temperature was ramped up in increments of 1 °C/s to 95 °C, and the fluorescence was read at the end of a 30 s hold at each temperature[158]. The program of fluorophore detection was set to the FRET option. The measured denaturation temperature was calculated by plotting the first derivative of the fluorescence signal as a function of temperature (−dF/dT)[158].

2.4.5 Mass spectrometry

The purified protein was validated by mass spectrometry[159]. Protein mass was measured at the mass spectrometry facilities of Dr. A. Schuetz and Dr. P. Mertins group at the Max Delbrück Center in Berlin. Protein sample concentration was adjusted to about 10 µM. Fresh matrix was prepared and mixed with protein solution (2 µl + 2 µl). Sample (1 µl) was spotted on the metal plate in duplicates, and the spot was allowed to dry completely.

2.4.6 Right-angle light scattering (RALS)

The RALS method to directly calculate the particle molecular weight by measuring the scattering light intensity at 90º to the incident beam was applied to determine the protein or protein complex molecular weight and homogeneity in solution[160]. The RALS unit containing a refractive index (RI) detector is coupled with the SEC and employed to analyze the eluate from the chromatography. The system was equilibrated with buffer overnight, and 100 µl of the sample was injected into the column. Fractions were collected and analyzed on SDS-PAGE. The RALS results were analyzed with the OmniSEC software.

2.4.7 DNA double strand preparation

Chemically synthesized single-stranded oligonucleotides were purchased from Eurofins Company in HPLC purified form. The oligonucleotides were dissolved in DNA buffer (Table D2) and the Nanodrop spectrometer was used to measure the concentration. Complementary strands were mixed at a molar ratio of 1: 1. DNA aliquots were placed in a 100 µl tube and the mixt solution was heated at 95 °C for 5 min, then slowly cooled down to room temperature over a period of around 2 h. The concentration of the double-stranded DNA sample was determined and the material stored at -20 °C.

2.4.8 Isothermal titration calorimetry (ITC)

ITC experiments were used to analyze the thermodynamic parameters of protein:DNA interactions, including the stoichiometry of the interaction (n), the dissociation constant (K_D), change in enthalpy (ΔH), and change in entropy (ΔS)[161]. From the ITC experiment, the binding affinity is given by the equilibrium dissociation constant (K_D). The greater the K_D value, the weaker is the binding affinity of the interaction. The experiment was carried out on the MicroCal PEAQ-ITC titration micro-calorimeter. Both protein and DNA samples were dialyzed against the ITC assay buffer (Table D2) overnight. Each ITC run comprised 19 injections. The sample concentration in the syringe is 20 fold higher than the sample concentration at the cell compartment. Binding curves were analyzed by the MicroCal-PEAQ-ITC program, which fitted the data in one of two possible modes: singe binding site or two binding sites.

2.5 Protein crystallization and structure determination

2.5.1 Protein crystallization

The sitting-drop vapor diffusion method was used for protein or protein:DNA complex crystallization. The protein concentration was between 5 mg/ml to 23 mg/ml for different crystallization screens. In all initial screening experiments, 200 nl protein solution were mixed with 200 nl reservoir solution by a Gryphon pipetting robot and equilibrated against 80 µl reservoir solution in 96-well microtiter plates. The commercial screening kits are listed in the Materials and Methods section. The plates were stored at either 4 ºC or 20 ºC in a Rock Imager storage system and automatically imaged by the system according to the standard schedule. After crystalline material was observed in the initial screen, fine screens were designed to improve crystal quality by varying the precipitant concentration and pH value of the reservoir solution starting from the initial crystallization conditions. Both 96-well plate sitting-drop and 24-well plate hanging-drop

vapor diffusion methods were applied in the fine screen. Crystals were cryo-protected with either 20% ethane glycerol (EG) or 20% glycerol, and flash-frozen in liquid nitrogen.

The Tfcp2l1 DNA binding domain (Tfcp2l1 DBD_{19-260}) was successfully crystallized in nine different conditions at 4 ºC of the initial screen. Three promising conditions were selected for further optimization. The Tfcp2 DNA binding domain (Tfcp2 DBD_{60-275}) was successfully crystallized in two conditions of the initial screen. The condition with ammonium sulfate as precipitant was optimized by both sitting-drop and handing-drop vapor diffusion methods in the fine screen. Another condition with dioxane was optimized by the sitting-drop vapor diffusion in the fine screen.

2.5.2 Protein:DNA complex crystallization

In order to obtain protein:DNA complexes for crystallization, the purified DNA binding domain was mixed and incubated with the DNA double strand. Based on the ITC analysis, the protein to 12-mer DNA molar ratio was kept at 2.1: 1, and the protein to 20-mer DNA molar ratio was kept at 4.1: 1.

The Tfcp2l1 DBD_{19-260}:ds12bpDNA complex was crystallized under three similar conditions, and the crystals were optimized in a fine screen by sitting-drop vapor diffusion. Crystals were cryo-protected with 20% EG and flash-frozen in liquid nitrogen.

The ds20bp DNA crystals were obtained in three conditions when co-crystalize the Tfcp2l1 DBD_{19-260} with ds20bpDNA. The crystals from the fine screen were validated only contain DNA inside.

2.5.3 Data collection

Single crystal X-ray diffraction data were collected at beamlines BL14.1 and BL14.2 at the BESSY II synchrotron facility (Helmholtz-Zentrum Berlin, Germany)[162]. BL14.1 was equipped with a PILATUS 6 M detector and BL14.2 with a MAR165 CCD detector. All diffraction experiments were carried out at a wavelength of 0.9184 Å. The iMosflm program was used to index the diffraction data[163,164], and XDSAPP 2.0 was used for data processing[165].

2.5.4 Molecular replacement

Phases for the Tfcp2l1 DBD_{19-260} diffraction data were obtained by molecular replacement using the crystal structure of Grhl1 DBD (PDB entry: 5MPI) as a template. Phases for the Tfcp2 DBD_{60-275} data were obtained by molecular replacement using the Tfcp2l1 DBD_{19-260} structure as template. Finally, phases for the Tfcp2l1 DBD_{19-260}:ds12bpDNA complex were obtained by molecular replacement using the Grhl1 DBD:DNA complex (PDB entry: 5MPF) as template. All phases were obtained using the program PHASER[166]. The structures were manually built and completed using COOT[167].

2.5.5 Model building and structure validation

The structure model of Tfcp2l1 DBD_{19-260} was refined using the CCP4_refmac program after each round of model building[168]. The structure model of Tfcp2 DBD_{60-275} was refined in the same way using TLS (translation-libration-screw) parameters calculated in CCP4. The structure model of the Tfcp2l1 DBD_{19-260}:ds12bpDNA complex was refined using the PHENIX_refine program and the CCP4_refmac program[169]. All refined structure models were validated by the Molprobity server[170,171].

3. RESULTS

3.1 Mouse Tfcp2l1 and human Tfcp2 constructs design and protein expression screen

Based on bioinformatics analysis, both mTfcp2l1 and hTfcp2 are predicted to consist of a CP2-like DNA-binding domain (DBD), a sterile alpha motif (SAM) domain and a C-terminal domain (CTD). Tfcp2 contains an extra Q-rich sequence compared to Tfcp2l1 (Fig. 3-1). To study the structural basis of gene expression regulation by mTfcp2l1 and hTfcp2, constructs containing the DBD, SAM, CTD, and all of them in one polypeptide, without unstructured polypeptide regions, were designed for protein expression.

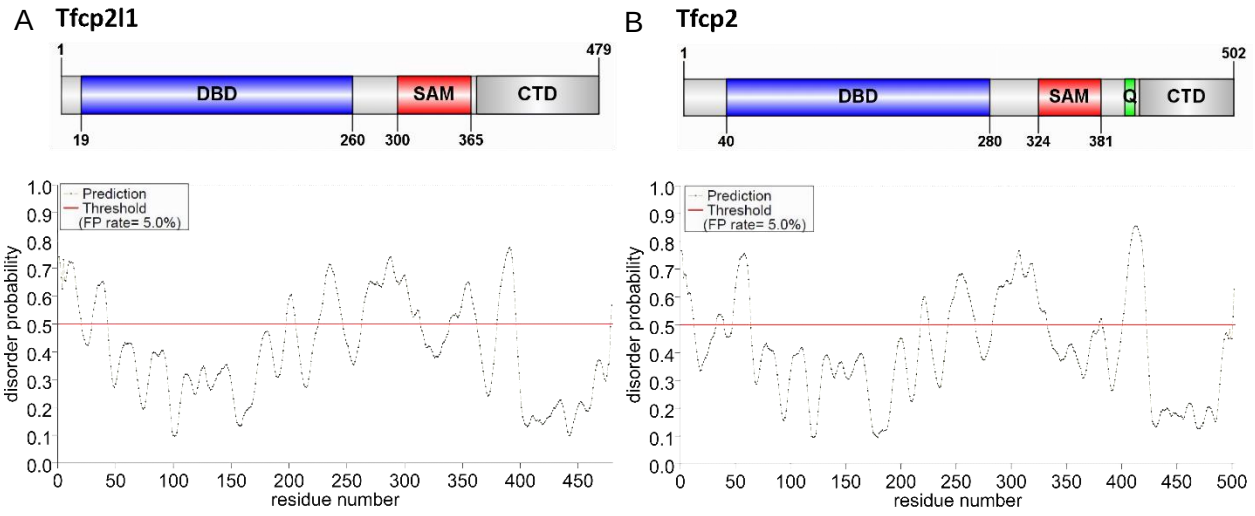

Figure 3-1. Cartoon domain organization and disorder prediction of Tfcp2l1 and Tfcp2 proteins. A, mouse Tfcp2l1 (UniProtKB: Q3UNW5) and B, human Tfcp2 (UniProtKB: Q12800) were assessed and predicted by the PrDoS online server.

The full-length cDNA sequence encoding mTfcp2l1 and hTfcp2 (isoform 1) were obtained from the Mammalian Gene Collection library. These constructs were transformed into the *E. coli* host strain BL21 DE3 Rosetta2 for protein expression. Before growing large-scale cultures, all constructs were screened by pre-expression tests in 20 ml LB medium. The soluble constructs were overexpressed in large-scale culture, then followed by protein purification. The constructs and protein expression tests are listed in Appendix A: Table A1 and A2.

3.2 mTfcp2l1 construct protein purification and biochemical assay

The Tfcp2l1 full-length construct could not be expressed, while the N-terminally truncated Tfcp2l1 Δ19 (amino acids (AAs), 19-479) could be expressed in the *E. coli* host strain BL21 DE3 Rosetta2. Tfcp2l1 Δ19 was purified by affinity chromatography, and the protein was validated by Western-blot based on the N-terminal 6-His-Tag (Fig. 3-2 A). In order to remove DNA/RNA contaminations, anion-exchange chromatography was applied. The eluted fractions from the Ni-NTA column were pooled, and the conductivity of the buffer was adjusted to corresponding to the low salt concentration buffer (Table B). The eluted fractions from the anion-exchange chromatography were pooled and concentrated, then loaded to the size exclusion chromatography (SEC) column. After the SEC, the Tfcp2l1 Δ19 protein samples were analyzed by SDS-PAGE, and the molecular weight was further validated by mass spectrometry (Fig. 3-2 B-D).

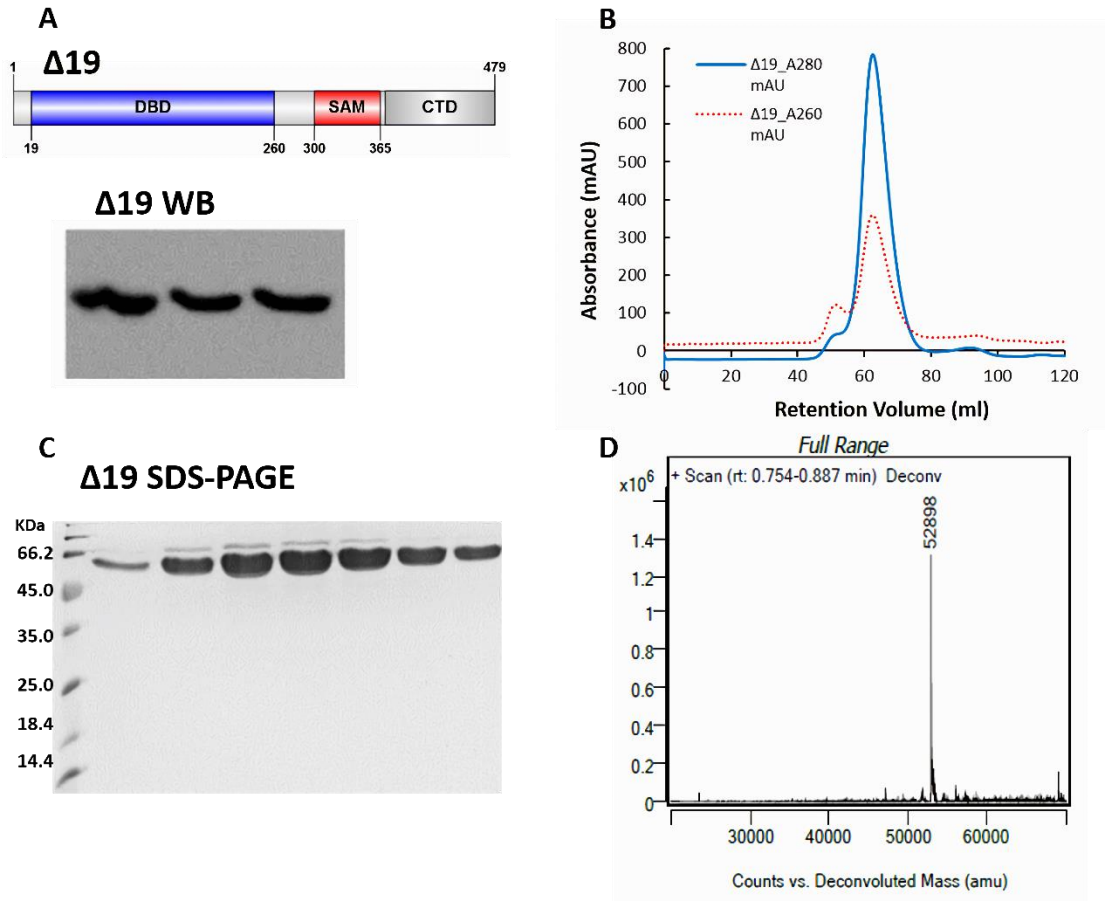

Figure 3-2. Purification of Tfcp2l1 Δ19 protein. A, Cartoon of the domain structure and Western blot of Tfcp2l1 Δ19 (with His-tag) after the Ni^{2+} affinity chromatography. B, SEC chromatogram of Tfcp2l1 Δ19 from the Superdex 200 16 60 column. C, SDS-PAGE for protein eluting from SEC. Mass spectrometry of Tfcp2l1 Δ19 from the final purification.

Tfcp2l1 Δ19 is stable in the solution without degradation at 4 ºC as shown by SDS-PAGE one week after the purification. No crystal was observed from the initial crystallization screening of Tfcp2l1 Δ19. As a potential problem regarding protein stability, one TEV protease cleavage site was identified in the C-terminal domain of Tfcp2l1. Therefore, a Q435A mutation was introduced to block this cleavage site and obtain the intact protein without N-terminal six-His tag. The new construct Tfcp2l1 Δ42 (AAs, 42-479) is shorter than Tfcp2l1 Δ19 and used for further crystallization.

After purification, Tfcp2l1 Δ42 was tested by the thermal shift assay (TSA) to identify the optimal buffer in which the protein is most stable, displaying the highest melting transition. The commercial screening kit CS-214 was applied for buffer optimization. As indicated by the TSA analysis, Tfcp2l1 Δ42 is most stable at pH values between 7.5 and 8.0 except in citrate buffer where pH 6.5 and 7.0 are optimal (Fig. 3-3 A). Simultaneously, the sodium ion concentration was screened at 50 mM, 150 mM, 250 mM, and 500 mM. The TSA result suggested that the purification buffer (25 mM HEPES, 250 mM NaCl, pH 7.8) is promising compared to other conditions (Fig. 3-3 B). Therefore, Tfcp2l1 Δ19 and Tfcp2l1 Δ42 were stored in this buffer for further analysis.

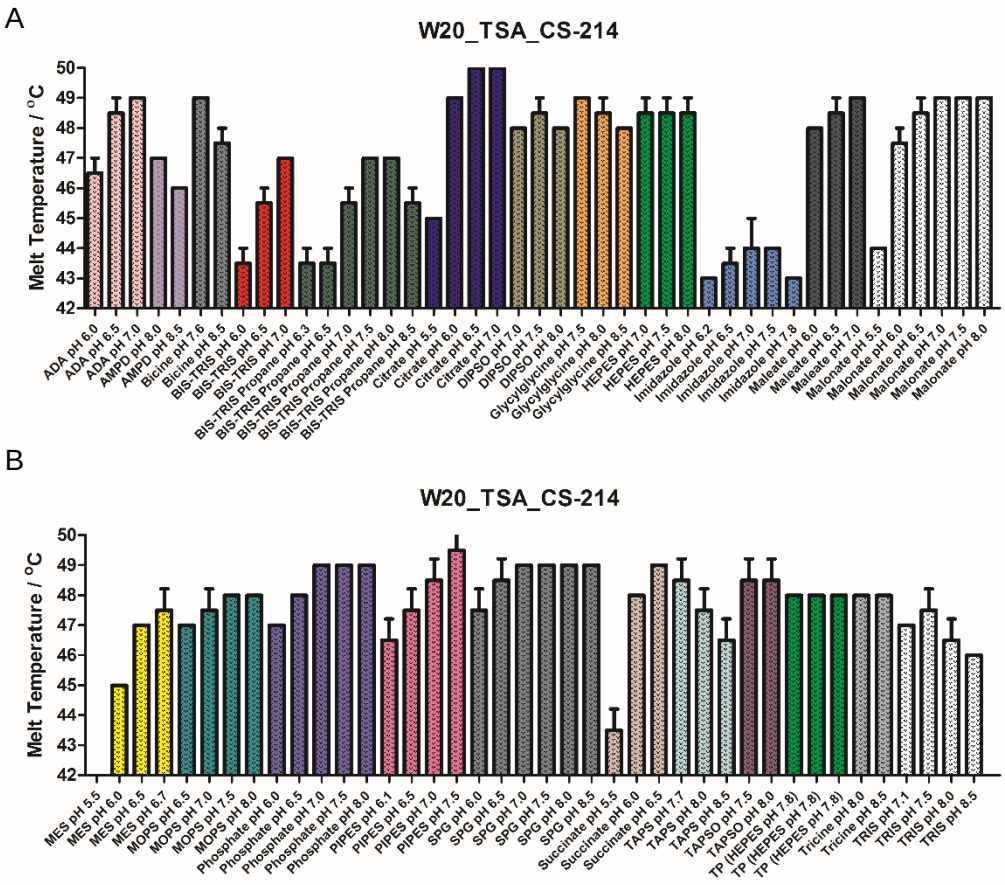

Figure 3-3. Histogram of buffers screening of Tfcp2l1 Δ42 (W20) by thermal shift assays (TSA). A and B, The kit CS-214 was applied for buffer optimization. The same buffer was grouped in one color. TP (target protein) buffer was used in the purification. Experiments were done in duplicates.

Both Tfcp2l1 Δ19 and Tfcp2l1 Δ42 were stable as tetramers in solution during the purification. Unfortunately, the crystallization screen (Fig. 3-4 B), did not yield any crystals of these two proteins. One strategy to overcome this problem may be to truncate flexible peptide regions to stabilize the protein for crystallization. Therefore, I generated the truncated constructs Tfcp2l1 Δ266-308 (AAs, 19-266, 308-479), Δ266-366 (AAs, 19-266, 366-479) and Δ364 (AAs, 19-364). Compared to Tfcp2l1 Δ19, the linker region (42 AAs) connecting the DBD and the SAM domain was absent in Tfcp2l1 Δ266-308, which contains the DBD and the CTD peptide region. Tfcp2l1 Δ364 extends from residue 19 to 364, containing the DBD and SAM domain.

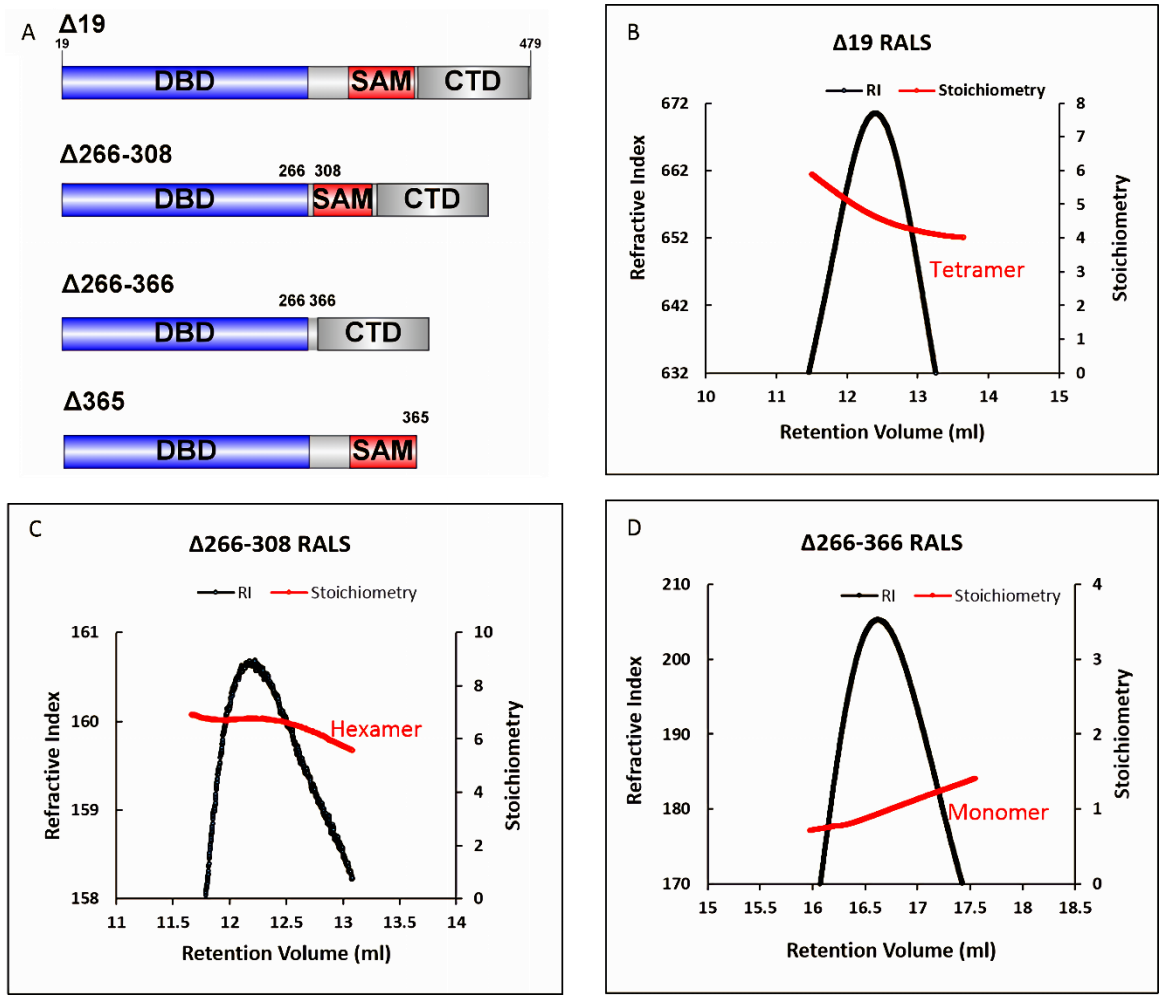

Figure 3-4. RALS study of Tfcp2l1 domain constructs. A, domain organization of the constructs tested in the RALS experiments. B-D, RALS analysis of proteins eluting from the Superdex 200 10/300 Hiload column. B, Tfcp2l1 Δ19 and C, Tfcp2l1 Δ266-308 at 1.5 mg/ml concentration, D, Tfcp2l1 Δ266-366 at 3 mg/ml concentration. The refractive index from RALS is shown in a black curve. The stoichiometry values are shown in a red line, which was obtained by dividing the measured molecular weight by the protein's theoretical molecular weight.

After the purification, the molecular weight Tfcp2l1 Δ266-308 was calculated by RALS. Tfcp2l1 Δ266-308 forms hexamers or species with molecular weight higher than hexamers (Fig. 3-4 C). This oligomerization behavior may indicate that the loop region (AAs, 266-308) plays a role in stabilizing the orientation of the Tfcp2l1 CTD. RALS analysis further showed that Tfcp2l1 Δ266-366 is present as a monomer in solution, suggesting that the SAM domain supports Tfcp2l1 oligomerization (Fig. 3-4 D). Interestingly, the CTD polypeptide region (AAs, 396-479) of Tfcp2l1 does not seem to be involved in oligomerization.

Tfcp2l1 Δ364, on the other hand, precipitated during anion exchange chromatography. Tfcp2l1 Δ364 is not stable and could not be purified. This is taken to imply that the CTD of Tfcp2l1 serves some function in stabilizing the protein.

Study of the Tfcp2l1 C-terminal domain

The Tfcp2l1 C-terminal fragment contains the SAM domain and the CTD. The SAM domain is flanked by two loop regions connecting the N-terminally adjacent DBD and CTD. The construct Tfcp2l1 Δ301 (AAs, 301-479) was created and expressed. Tfcp2l1 Δ301 is not stable and starts to aggregate at approximately 1 mg/ml concentration during the purification. A similar behavior was observed with the SAM domain itself. It has been reported that SAM domains in other proteins are not stable during purification, and it was suggested that these SAM domains could form polymer fibers by head-to-tail connection[172,173]. In one case, polymer formation could be blocked by a large domain adjacent to the SAM domain[173]. Similarly, Tfcp2l1 Δ19 and Tfcp2l1 Δ42, both containing the DBD and the Tfcp2l1 CTD, were stable in solution. In one study, a single-site mutation was reported that had the ability to disrupt the SAM domain oligomerization[174]. The Tfcp2l1 SAM domain has been predicted to form hexamers with the residues D345 and G355 involved in subunit interactions[175]. Therefore, D345A and G355E mutations were introduced to improve the stability of monomeric SAM domains in solution. The Tfcp2l1 protein variants carrying the D345A or G355E mutations precipitated during purification as did the corresponding wild-type protein fragments. The result suggested that these two mutations are not crucial for SAM domain oligomerization.

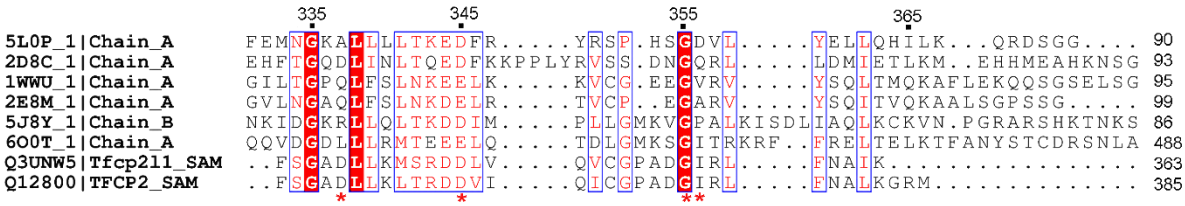

Figure 3-5. SAM domain sequence alignment. Strictly conserved residues are colored in red. Red asterisks indicate the positions of single-site mutations. Residue positions above the aligned sequences are for Tfcp2l1

Based on a set of known SAM template structures with PDB IDs 1WWU (chain A), 2D8C (A), 2E8M (A), 5J8Y (B), 5L0P (A) and 6O0T (A), three-dimensional structures the Tfcp2l1 and Tfcp2 SAM domains were obtained by homology m modeling using the SWISS-MODEL program[176]. Two further residues in Tfcp2l1, D337 and I356, were selected based on the structure models, and the mutations D337R and I356R were introduced to improve the Tfcp2l1 SAM domain stability (Fig. 3-5). Unfortunately, these four individual single-site mutations did not improve the protein stability, even though these four residues were predicted to locate at the subunit interface in SAM domain oligomers.

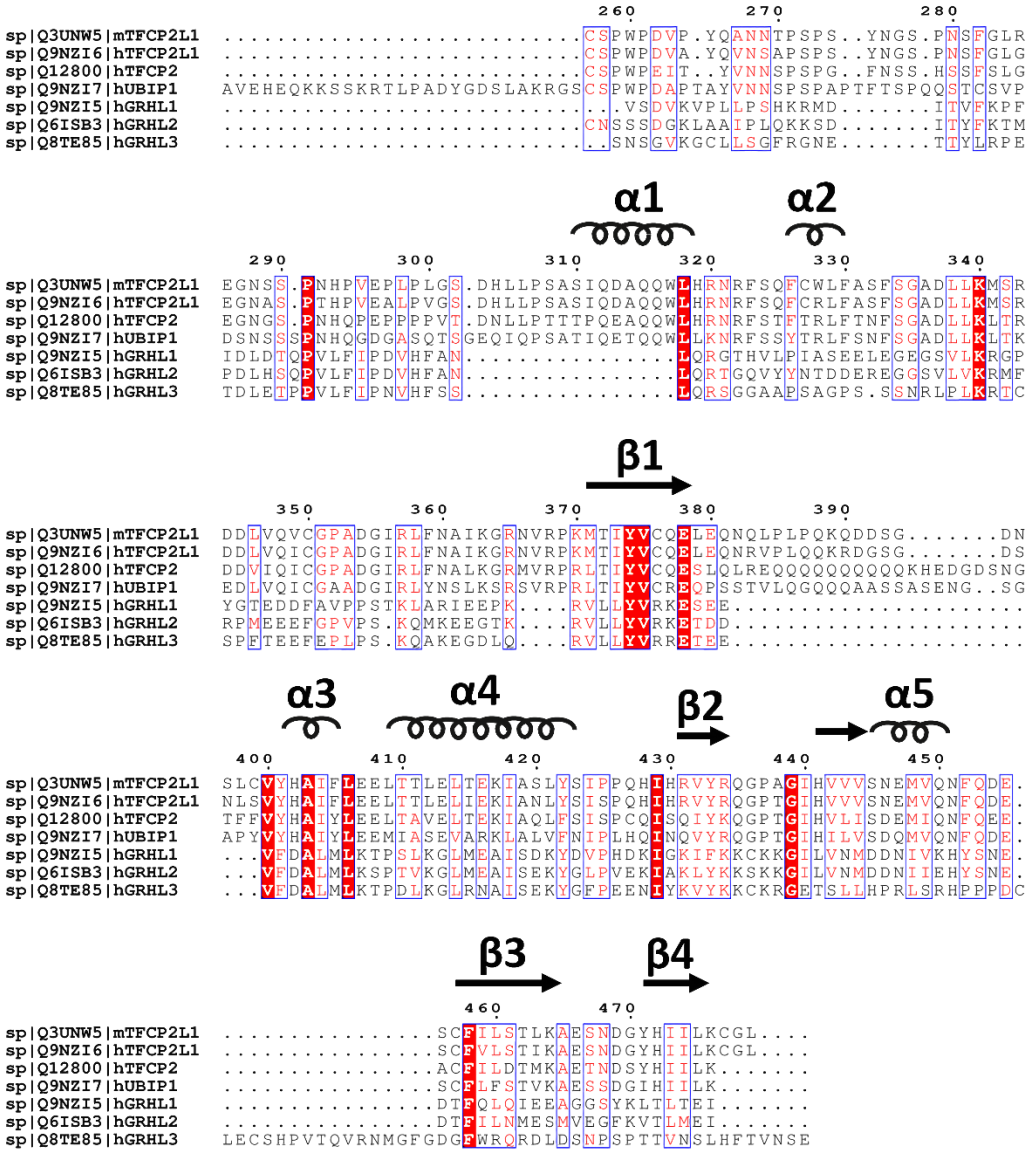

Figure 3-6. Sequences alignment of the C-terminal domains of homologs of the Grh/CP2 family. The conserved residues were colored in red. The predicted secondary structural domain organization (α helix and β strand) was labeled and showed above the sequence. Both mouse and human Tfcp2l1 were shown here.

As mentioned before, the CTD portion of Tfcp2l1 is not involved in protein oligomerization. Therefore, no single site mutation was performed in this region, but the CTD appears to contribute to Tfcp2l1 stability in solution. Protein sequence alignment shows that the C-terminal regions of Tfcp2l1 and Grhl1 share 18% residue identity (Fig. 3-6). In spite of the relatively poor sequence match, similar secondary structure is predicted for the C-terminal portions of Tfcp2l1 and Grhl1 (Fig. 3-6). However, a Tfcp2l1 construct covering AAs 395-479 proved unstable in solution during purification. This protein fragment gels after affinity chromatography at 500 mM sodium chloride. In low-salt buffer (150 mM NaCl), the CTD precipitated immediately. Neither the SAM domain itself nor the intact C-terminal region could be purified.

Study of the Tfcp2l1 DNA-binding domain

To study the Tfcp2l1 DBD, three different protein constructs were designed to determine the DNA-binding domain boundaries: Tfcp2l1 DBD_{47-283}, DBD_{19-283}, and DBD_{19-260} (Fig. 3-7 A). This protein fragment was purified and tested first, but the N-terminal His-tag of Tfcp2l1 DBD_{47-283} could not be cleaved. The hypothesis is that the TEV protease cleavage site is inaccessible in this protein because it may be too close to the core DBD structure. I therefore purified Tfcp2l1 DBD_{19-283} and Tfcp2l1 DBD_{19-260} and found that their N-terminal His-tags could completely be cleaved. Therefore, it is assumed that the Tfcp2l1 DBD N-terminal boundary is at or near residue Y19.

Both Tfcp2l1 DBD_{19-283} and Tfcp2l1 DBD_{19-260} were purified for further tests. After the SEC, SDS-PAGE showed that the Tfcp2l1 DBD_{19-283} sample contained two bands representing protein species which the Ni^{2+} affinity chromatography, the cation-exchange chromatography, and the SEC could not separate (Fig. 3-7 C). Mass spectrometry only gave one base peak for the Tfcp2l1 DBD_{19-283} sample, demonstrating that the lower band is not from the upper band's degradation. The RALS assay showed that Tfcp2l1 DBD_{19-283} is stable and monomeric in solution (Fig. 3-7 E).

For Tfcp2l1 DBD_{19-260}, SDS-PAGE showed that the protein was pure without any other bands (Fig. 3-7 D). The RALS assay showed that Tfcp2l1 DBD_{19-260} is a monomer in solution, and the molecular weight determined by mass spectrometry was corresponding to the theoretical molecular weight (Fig. 3-7 F and B). To determine if the loop region from residue 261 to 283 was involved in protein:DNA binding, an ITC experiment was employed to determine if the Tfcp2l1 DBDs with C-termini at either residue 261 or 283 differed in their DNA affinity. There is no significant difference between two K_D values of Tfcp2l1 DBD_{19-260} and DBD_{19-283} binding to ds20bpDNA (Fig. 3-7 G and H). DBD All these three Tfcp2l1 DBD proteins were set up for crystallization.

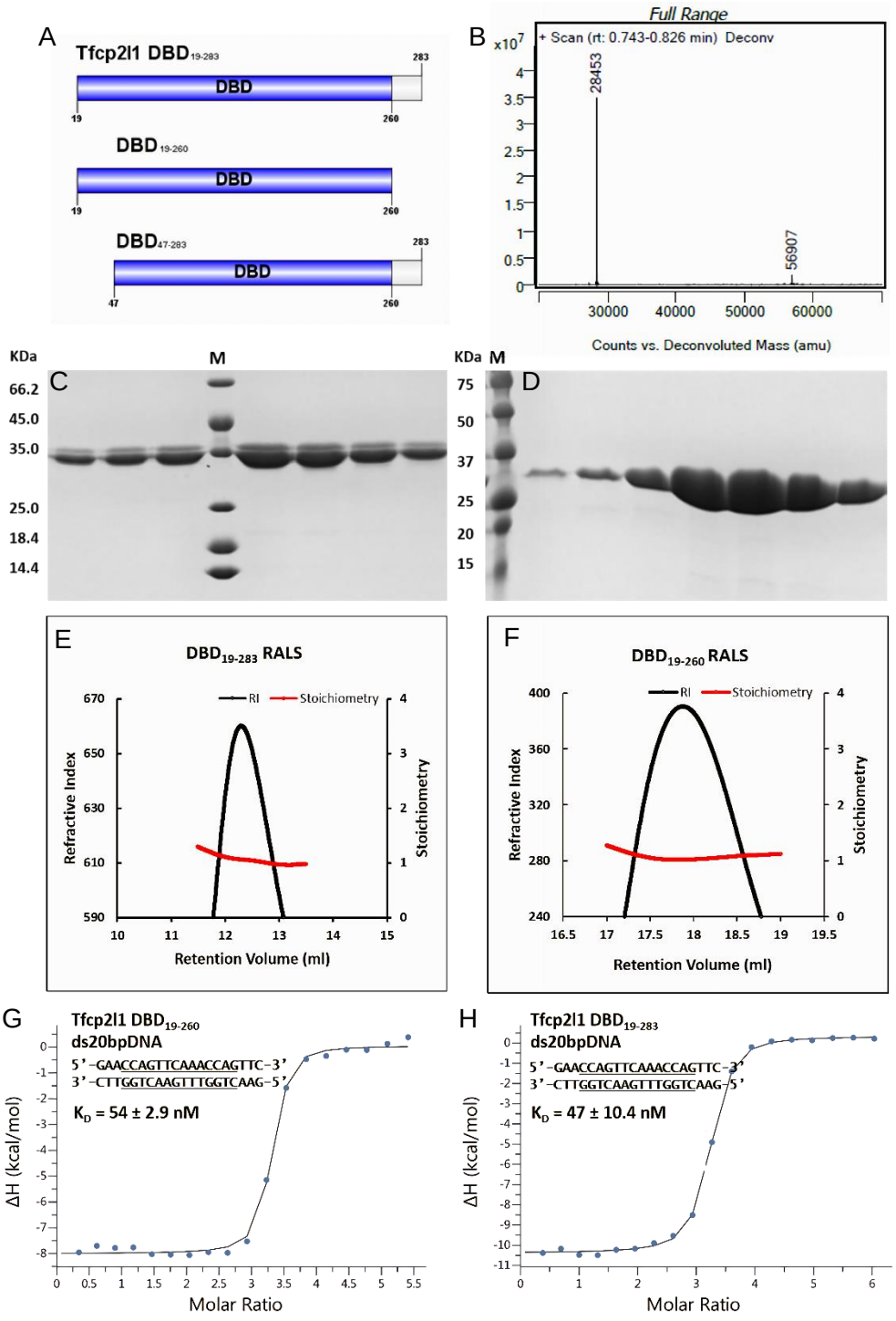

Figure 3-7. Characterization of Tfcp2l1 DBDs. A, domain organization of the constructs used in the purification. B, Mass spectrometry of Tfcp2l1 DBD$_{19\text{-}260}$. C and D, SDS-PAGE of Tfcp2l1 DBD$_{19\text{-}283}$ and Tfcp2l1 DBD$_{19\text{-}26}$ after SEC. E, RALS assay coupled with the Superdex 75 10/300 GL column to analyze Tfcp2l1 DBD$_{19\text{-}283}$ and F, RALS coupled with the Superdex 200 10/300 GL column to analyze Tfcp2l1 DBD$_{19\text{-}260}$. The refractive index from RALS is shown in the black curve. The stoichiometry values are shown as red lines, obtained by dividing the measured molecular weight into the protein's theoretical molecular weight. G, ITC measurement of DBD$_{19\text{-}260}$ and ds20bpDNA with K_D value 54 nm. H, measurement of DBD$_{19\text{-}283}$ and ds20bpDNA with K_D value 47 nM. Experiments were done in duplicates.

3.3 Tfcp2l1 DBD crystallization and structure determination

3.3.1 Initial screen

In the initial screen, Tfcp2l1 DBD_{47-283} was set up for crystallization in plates at a concentration of 16.2 mg/ml. For Tfcp2l1 DBD_{19-283} concentrations ranged from 6 mg/ml to 24 mg/ml, and for Tfcp2l1 DBD_{19-260} from 7.4 mg/ml to 13.6 mg/ml. Crystallization plates were set up both at 4 ºC and 20 ºC. Neither Tfcp2l1 DBD_{47-283} nor Tfcp2l1 DBD_{19-283} could be crystallized under these conditions. However, Tfcp2l1 DBD_{19-260} crystals were observed in 4 ºC in nine different conditions: **1)** on day 5 in 1 M imidazole, **2)** on day 7 in 0.2 M Na^+/K^+ tartrate, 20% w/v PEG 3350, **3)** on day 7 in 0.02 M $MgCl_2$, 0.15 M KCl, 0.05 M Tris pH 7.5, 15% w/v PEG 4000, **4)** on day 7 in 0.1 M Tris pH 8, 18% w/v PEG 8000, 0.2 M Mg^{2+} formiate, **5)** on day 9 in 0.1 M Na^+ succinate pH 7, 15% w/v PEG 3350, **6)** on day 9 in 0.01 M $CaCl_2$, 0.01 M Tris pH 7.5, 20% w/v PEG 8000, **7)** on day 10 in 0.1 M NaCl, 0.05 M MOPS pH 7, 20% w/v PEG 4000, **8)** on day 10 in 0.1 M NaCl, 0.1 M HEPES pH 7.5, 1.6 M $(NH_4)_2SO_4$, **9)** on day 14 in 0.1 M Tris pH 8, 20% w/v PEG 4000.

3.3.2 Fine screens

The three most favorable conditions were chosen for fine screens. In the fine screen condition of 0.2 M Na^+/K^+ tartrate, 18% w/v PEG 3350, crystals were observed on the third day and quickly formed clusters. A 24-well plate was chosen for optimization. 0.5 µl protein solution plus 1 µl mother liquor were loaded to a cover slip and suspended above a deep reservoir containing 200 µl mother liquor (0.1 M Na^+/K^+ tartrate, 15% w/v PEG 3350) (Fig. 3-8 B). The crystals grew for 35 days at 4 ºC. Crystals were mounted and protected with 20% EG and quick-frozen in liquid nitrogen. 1.93 Å resolution diffracted data were collected. For the fine screen condition of 0.05 M KCl, 0.05 M Tris pH 7.5, 12% PEG 3350, 2.4 Å data were collected (Fig. 3-8 A). Crystals from the

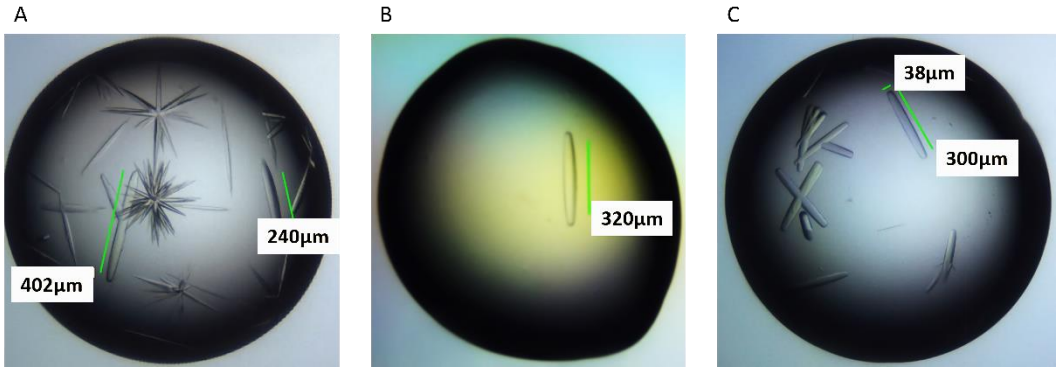

Figure 3-8. Crystals of Tfcp2l1 DBD_{19-260} from three fine screen conditions. A, Tfcp2l1DBD crystals from 0.55 M Tris pH 7.5, 0.05 M KCl, 15.5% w/v PEG 3350. B, Tfcp2l1 DBD crystal from 0.1 M Na^+/K^+ tartrate, 15% w/v PEG 3350. C, Tfcp2l1 DBD crystal from 0.1 M Na^+ succinate pH 6.6, 12.5% PEG 3350. All crystals grew at 4 ^{0}C. The size of the crystals was measured by the rocker program ruler.

screen condition of 0.1 M Na^+ succinate pH 6.6, 12.5% PEG 3350, were not further tested (Fig. 3-8 C).

3.3.3 Tfcp2l1 DNA-binding domain structure

Based on the protein sequence alignment, Tfcp2l1 DBD shares 64% identity with Grhl1 DBD. Therefore, the Grhl1 DBD structure (PDB: 5MPI) was chosen as the template for molecular replacement. Both CCP4 and Phoenix were applied for structure refinement. Finally, the Tfcp2l1 DBD domain structure was determined at 1.94 Å resolution with R-work of 17.69% and R-free of 22.34%. Based on these R-values, the conformation of the majority of protein side chains are revealed in this structure.

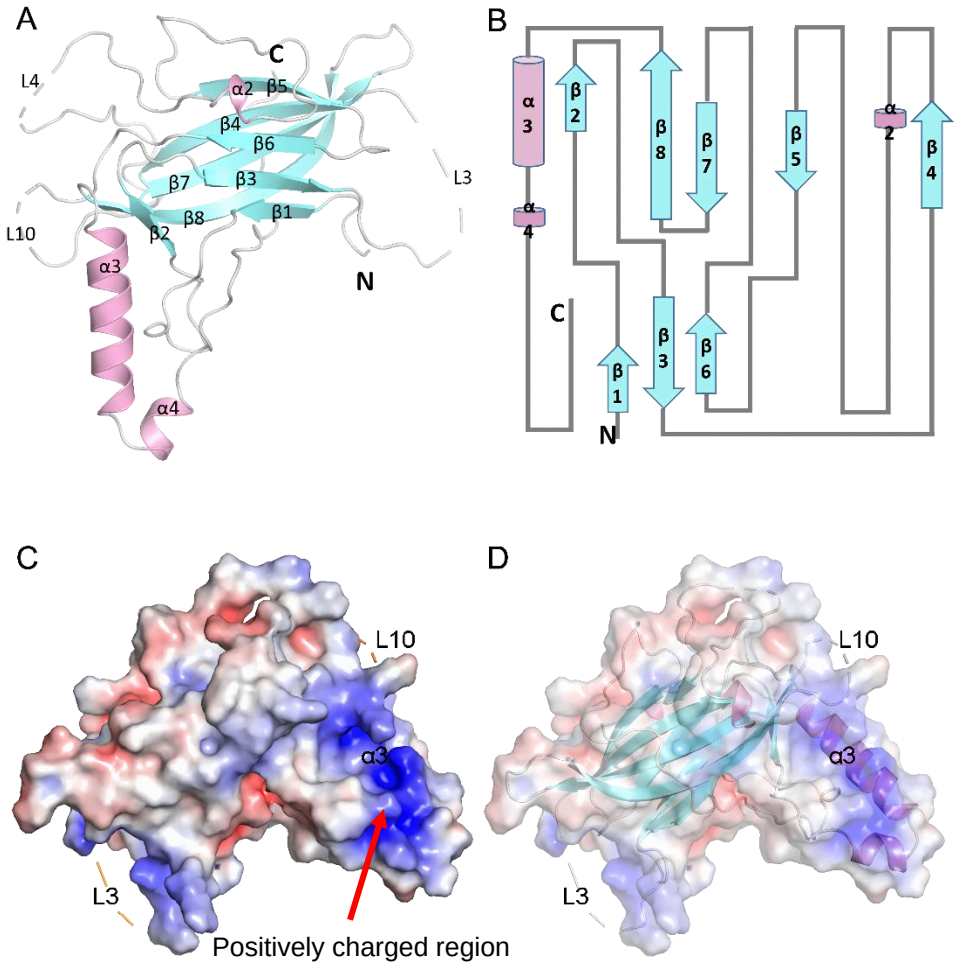

Figure 3-9. Overall structure of Tfcp2l1 DBD. A, crystal structure of Tfcp2l1 DBD presents in a cartoon drawing, α helices in pink and β strands in aquamarine. Dashed loops L3, L4 and L10 indicate flexible polypeptide segments without electron density, which were not modeled in the structure. B, topology diagram of the Tfcp2l1 DBD structure. C, electrostatic potential surface of Tfcp2l1 DBD. A prominent positively charged area is indicated by the red arrow. D, transparent view (-20%) of the electrostatic potential surface of Tfcp2l1 DBD highlighting helix α3, loops L3 and L10. The electrostatic potential surface is colored with positive potential (+5 kT) in blue and negative potential (-5 kT) in red.

Tfcp2l1 DBD crystals contained one protein molecule in each asymmetric unit. From the electron density map, the structure representing residues from 43 to 260 could be modeled (Fig. 3-9 A). Residues 86-93, 118-123 and 176-183 in loops L3, L4 and L10, respectively, were not represented in electron density. Tfcp2l1 DBD contains two twisted antiparallel β sheets comprising three strands (β1, β3, and β6) and five strands (β2, β4, β5, β7, and β8), which form an immunoglobulin-(Ig-) like core domain structure (Fig. 3-9 B). The three presumably flexible polypeptide regions without electron density (displayed as dashed polypeptide trace) are located in surface loops L3, L4, and L10. Together with the Ig-like domain helices α3 and α4 give the Tfcp2l1 DBD an L-shaped structure, which is completed by the C-terminal peptide segment connecting α3 with the β-stranded core domain.

Tfcp2l1 DBD electrostatic potential surface indicates a prominent positively charged regions are surrounding helices α3 and α4 and loop L10. This positively charged area is a potential site for DNA binding (Fig. 3-9, C-D).

3.4 hTfcp2 protein structure and biochemical assays

3.4.1 Characterization of full-length hTfcp2

As described, human Tfcp2 has a similar domain organization to murine Tfcp2l1. As observed with the mouse homolog, full-length human Tfcp2 could not be expressed in bacterial cells. The N-terminally truncated constructs $Tfcp2_{34-502}$ and $Tfcp2_{60-502}$ were expressed and purified. Unfortunately, these two proteins formed higher oligomers, presumably due to the presence of the SAM domain.

3.4.2 hTfcp2 DBD purification and crystallization

Based on the protein sequence alignment and secondary structure prediction, the constructs $CP2_{60-275}$ and $CP2_{60-288}$ of Tfcp2 DBD were created. These proteins were purified and validated by mass spectrometry. The Tfcp2 DBD variants were present as monomers in solution based on RALS analysis. The two proteins were purified, validated by MS and stored for further study.

$Tfcp2_{60-275}$ was set up for the crystallization at 20 mg/ml concentration. From the initial screen, crystals were observed at 4 ºC after 49 days in 1.6 M $(NH_4)_2SO_4$, 10% v/v dioxane, 0.1 M MES pH 6.5 (Fig. 3-10 A). These initial crystals diffracted to 3.3 Å resolution. During structure determination, the R-value remained high (~40%), and it proved challenging to build a complete structure model. Crystallization conditions were optimized by fine screening, yielding diffraction

data at 3.1 Å resolution (Fig. 3-10 B). The electron density map indicated the presence of three protein molecules per asymmetric unit, but the third molecule's electron density was of poor quality. A complete model of $Tfcp2_{60-275}$ could not be obtained.

Large crystals with good shape were obtained in the fine screen condition 1.45 M $(NH_4)_2SO_4$, 6.5% v/v dioxane, 0.1 M MES pH 6.5 (Fig. 3-10 C). A crystal was mounted, and 2.72 Å resolution diffraction data were finally collected. For the construct $CP2_{60-288}$, there is no crystal observed.

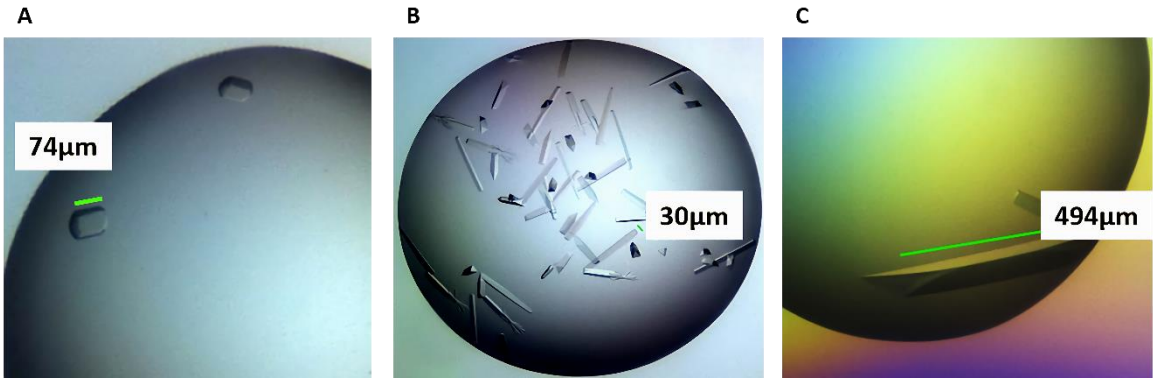

Figure 3-10. Crystals of $Tfcp2_{60-275}$ comprising the DBD. A, crystals from the initial screen. B and C, crystals from fine screens. The resolution of the resulting X-ray diffraction data was 3.3 Å, 3.1 Å, and 2.72 Å for (A), (B) and (C), respectively. All crystals grew at 4 ºC.

3.4.3 hTfcp2 DBD structure determination

The structure of Tfcp2l1 DBD was employed to be the template for molecular replacement phase analysis. I determined the Tfcp2 DBD structure at 2.72 Å resolution with R_{work} = 23.44% and R_{free} = 26.85%. With these R values, I am confident that the overall structure and the side chain conformations are correctly determined.

The Tfcp2 DBD crystal structure contains two protein molecules in each asymmetric unit. Residues 64 to 275 are revealed in the electron density map, but residues 128-142 in gap G1 and residues 197-203 in loop L8 are not part of the final model, presumably due to disorder (Fig. 3-11 A and B). As seen before for the Tfcp2l1 DBD, the Tfcp2 DBD also contains two twisted antiparallel β sheets, one comprising three strands (β1, β3, and β7) and the other comprising six strands (β2, β4, β5, β6, β8 and β9). Arrangement of these β-sheets atop one another forms an immunoglobulin- (Ig-) like core structure (Fig. 3-11 A). Together with the Ig-like domain, helices α3 and α4 give the Tfcp2 DBD an L-shaped structure, which is completed by the C-terminal peptide segment connecting α3 with the β-stranded core domain. The Tfcp2 DBD electrostatic

potential surface displays a prominent positively charged region surrounding helices α3 and α4. This positively charged area is a potential site for DNA binding (Fig. 3-11 C).

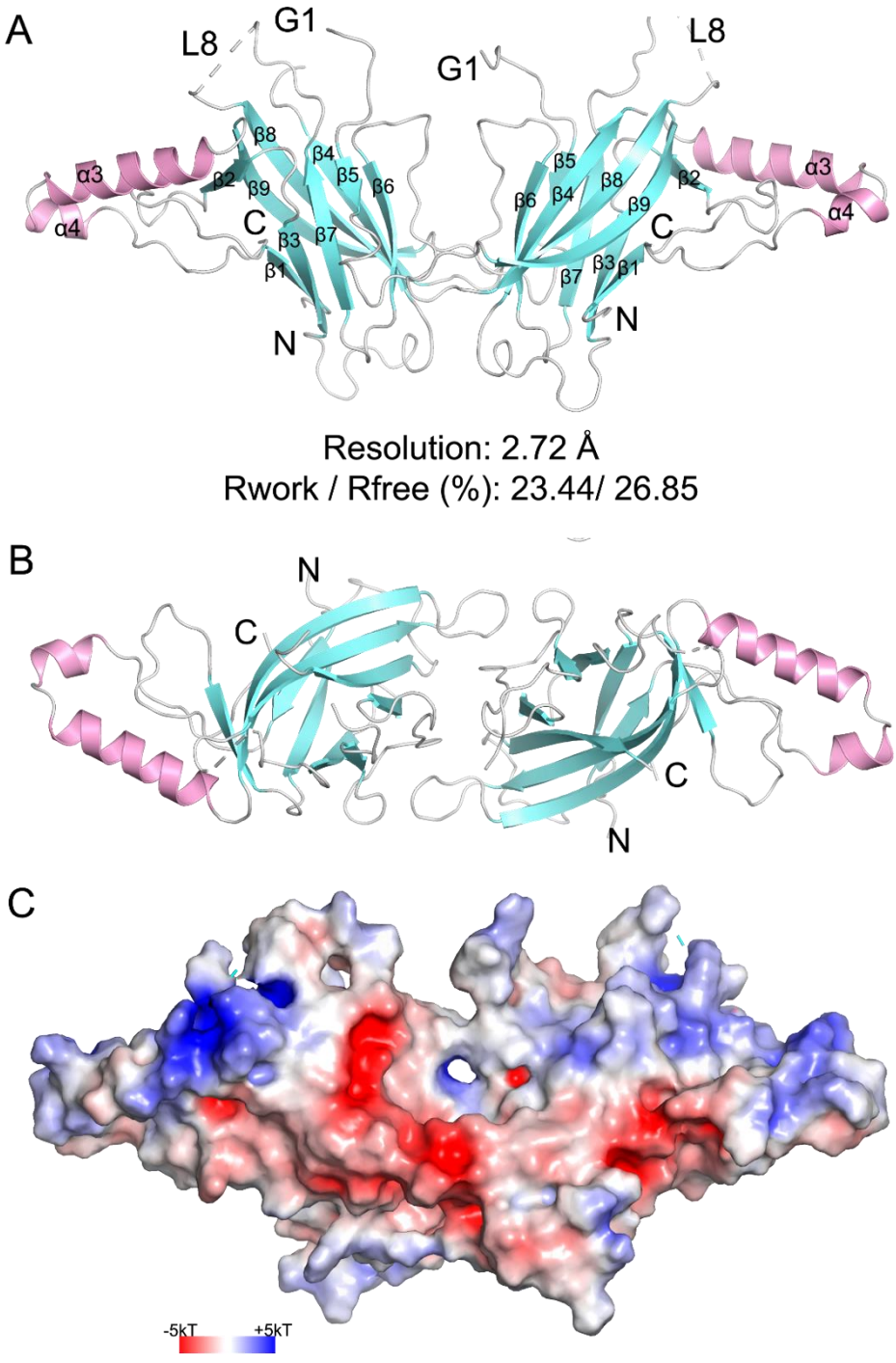

Figure 3-11. The overall structure of Tfcp2 DBD. A, cartoon drawing of Tfcp2 DBD in two orthogonal orientations. α helices are colored pink and β strands aquamarine. Regions with poor electron density where the structure could not be modeled are indicated by dashed line segments. B, cartoon view in rotation 90º. C, electrostatic potential surface of Tfcp2 DBD. The electrostatic potential surface is colored blue for potential (+5 kT) and red for negative potential (-5 kT).

3.5 DNA binding studies on CP2 subfamily members

3.5.1 Characterization of target DNA binding by Tfcp2l1 DBD

From published literature it is known that Tfcp2l1 binds specifically to the14-mer DNA motif CC$^{A}/_{G}$GTTCAAACC$^{A}/_{G}$G[49]. The double-stranded oligodeoxynucleotide CCAGTTCAAA-CCAG (ds14bpDNA) is one variant of this motif. ds14bpDNA was used to assay specific DNA binding by Tfcp2l1 *in vitro*. Another variant of the Tfcp2l1 target sequence, CCGGTTCAAACCGG, was predicted to form a hairpin structure by the *OligoEvaluator* online analysis tool and therefore excluded from binding studies with Tfcp2l1.

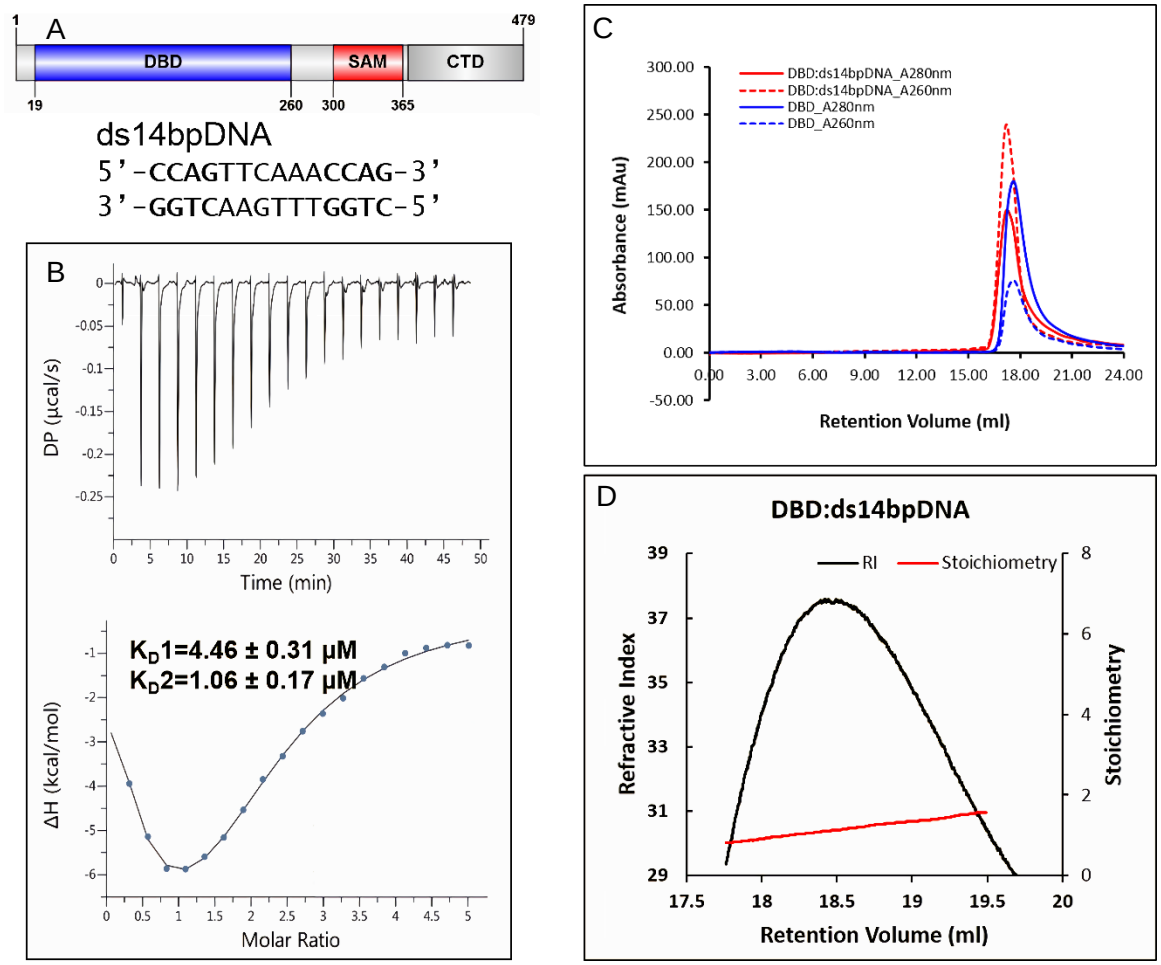

Figure 3-12. Characterization of the Tfcp2l1 DBD interaction with ds14bpDNA. A, domain organization of Tfcp2l1 and the ds14bpDNA sequence. B, ITC measurement of Tfcp2l1 DBD binding to ds14bpDNA. C, Comparison of elution profiles of Tfcp2l1 DBD (red) and Tfcp2l1 DBD in presence of ds14bpDNA (blue) from the analytical SEC with a Superdex 200 10/300 column. UV absorption at 280 nm is shown by solid lines and at 260 nm by dotted lines. D, RALS measurement showing monomeric Tfcp2l1 in solution. The refractive index from RALS is shown in the black curve. The stoichiometry values are shown in the red line, which was obtained by dividing the measured molecular weight into the protein's theoretical molecular weight.

DNA sequence of CCAGTTCAAACCAG and its complementary DNA strand were synthesized, ds14bpDNA was prepared by annealing (Fig. 3-12 A) and used for isothermal titration calorimetry. In the ITC assay, the DNA was titrated with Tfcp2l1 DBD_{19-260} at 10 ºC. The titration curve was fitted using a two-set binding model, revealing that DBD_{19-260} binds ds14bpDNA with a K_D1 of 4.46 µM and a K_D2 of 1.06 µM (Fig. 3-12 B). For ds14bpDNA, it is difficult to assess whether two Tfcp2l1 DBD molecules bind as an apparent dimer to one ds14bpDNA duplex, or four molecules bind ds14bpDNA as an apparent tetramer, even though the titration fitted into one-set binding model. More experiments need to be done to determine the molar ratio of Tfcp2l1 DBD binding to the DNA.

The Tfcp2l1:ds14bpDNA complex was prepared *in vitro* based on the stoichiometry of four protein molecules to one DNA duplex. However, analytical SEC provided no evidence for a complex of Tfcp2l1 DBD dimer or tetramer bound to the DNA (Fig. 3-12 C). The overlay of the elution profiles from two SEC experiments with Tfcp2l1 DBD and the Tfcp2l1 DBD in presence of ds14bpDNA showed no difference in retention volumes (RVs). It is concluded that the low-affinity protein:DNA complex indicated by ITC dissociated during the SEC in the ITC buffer. RALS results showed that Tfcp2l1 DBD in the presence of ds14bpDNA is monomeric in the solution (Fig. 3-12 D). Tfcp2l1 DBD could not strongly bind to this ds14bp DNA sequence, even though this sequence was identified as Tfcp2l1 target by ChIP-Seq[49]. The best target DNA sequence has to be determined to get a stable protein DNA complex.

Grhl1 DBD from the Grhl subfamily binds tightly and specifically to the 12-mer DNA AAAACCGGTTTT with a K_D of 91 nM[23]. To identify the optimal DNA motif which Tfcp2l1 targets, DNA sequences of different length were designed and synthesized. To extend the ds14bpDNA sequence, the nucleoside adenosine (A) was added to 5’ and Thymine (T) was added to 3’ ends to yield ds16bpDNA and ds18bpDNA. ds16bpDNA contains one A in both 5’ and 3’ flanking sequence around the ds14bpDNA sequence, and ds18bpDNA contains two As in both 5’ and 3’ flanking sequences. ds19bpDNA and ds20bpDNA were designed by further extending the ds18bpDNA sequence to create 5’-overhangs or blunt ends, respectively. ds20bpDNA was designed to have two similar sequence of GAACCAGTTC, while the ds19bpDNA lacks one C at the 3’-end compared to ds20bpDNA. ds21bpDNA carries an additional cytosine at the 5’-end compared to ds20bpDNA, and ds22bpDNA has two more base pairs A compared to ds20bpDNA. The sequences used in the following ITC experiments are shown in Table_3.1.

As determined in ITC assays, the K_D value of Tfcp2l1 DBD_{19-260} binding to the DNA is dramatically decreasing from 4.46 µM to 54 nM, going from ds14bpDNA to ds19bpDNA (Fig. 3-13 A-C, Fig. 3-

14 A). There seems to be a pattern that the shorter the 5' and 3' flanking sequences around the consensus DNA motif, the weaker is the affinity. With the longer fragments from ds20bpDNA to ds22bpDNA, the affinity does not change significantly (Fig. 3-13 D-F). Three nucleotides flanking of core 14-mer DNA sequence appears sufficient to support the protein:DNA binding, longer flanking sequences around the consensus DNA motif which do not contribute to the affinity.

Table_3.1 Sequence of the oligonucleotide variants

Double-stranded DNA	Sequence
ds14bpDNA	5'-CCAGTTCAAACCAG-3' 3'-GGTCAAGTTTGGTC-5'
ds16bpDNA	5'-ACCAGTTCAAACCAGT-3' 3'-TGGTCAAGTTTGGTCA-5'
ds18bpDNA	5'-AACCAGTTCAAACCAGTT-3' 3'-TTGGTCAAGTTTGGTCAA-5'
ds19bpDNA	5'-GAACCAGTTCAAACCAGTT-3' 3'-TTGGTCAAGTTTGGTCAAG-5'
ds20bpDNA	5'-GAACCAGTTCAAACCAGTTC-3' 3'-CTTGGTCAAGTTTGGTCAAG-5'
ds21bpDNA	5'-CGAACCAGTTCAAACCAGTTC-3' 3'-CTTGGTCAAGTTTGGTCAAGC-5'
ds22bpDNA	5'-GAAACCAGTTCAAACCAGTTTC-3' 3'-CTTTGGTCAAGTTTGGTCAAAG-5'

The binding isotherm from the titration of Tfcp2l1 DBD with ds20bpDNA is fitted well with the one-set binding model. The ds20bpDNA sequence GAA**CCAG**TTCAAA**CCAG**TTC contains two similar sequence G(A)AA**CCAG**TTC, and each one could bind two Tfcp2l1 DBD molecules (Fig. 3-13 D). The titration curve of Tfcp2l1 DBD binding to ds19bpDNA is best fitted with the two-set binding model, which implies that the lack of one 3'-terminal cytosine affects the protein DNA binding pattern. Therefore, the optimized length of the DNA sequence was determined to be 20 base pairs.

For ds14bpDNA and ds16bpDNA, Tfcp2l1 DBD is likely to bind the duplex DNA in 2: 1 stoichiometry as suggested by the non-integer stoichiometry value N = 2.5. The deviation of N from an integer value might be caused by the relatively weak protein:DNA binding. For oligonucleotides ranging in length from 18 to 22 -mer, the stoichiometry values are closer to three, suggesting the binding of three Tfcp2l1 DBD molecules to the DNA duplexes (Fig. 3-14 B). In both regimes, the stoichiometry of the Tfcp2l1 DBD to DNA double strands is not integer of two or

fouras initially expected. Therefore, RALS analysis was applied to determine the molecular weight of the protein:DNA complexes and to elucidate the binding stoichiometry.

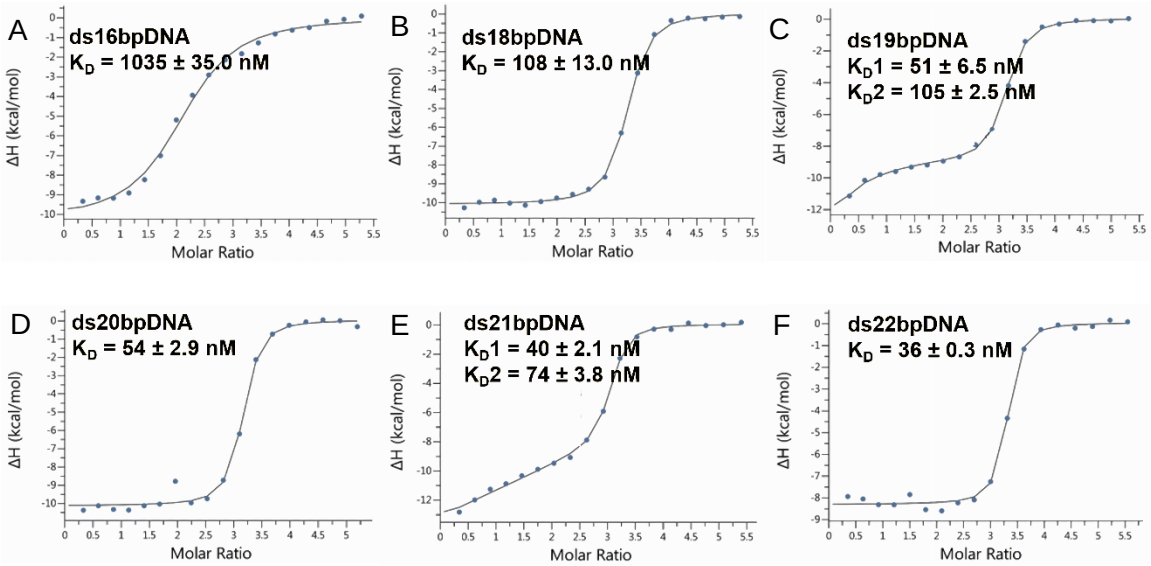

Figure 3-13. Quantitative analysis of the binding affinity of the Tfcp2l1 DBD with DNA variants of different length. A-F, ITC measurements of Tfcp2l1 DBD_{19-260} with A: ds16bpDNA, B: ds18bpDNA, C: ds19bpDNA, D: ds20bpDNA, E: ds21bpDNA and F: ds22bpDNA. The titration curves of DBD with ds19bpDNA (C) and ds21bpDNA (E) were fitted with the two-binding-sites model, the isotherms in A, B, D and F with the one-binding-site binding model. Experiments were done in duplicates.

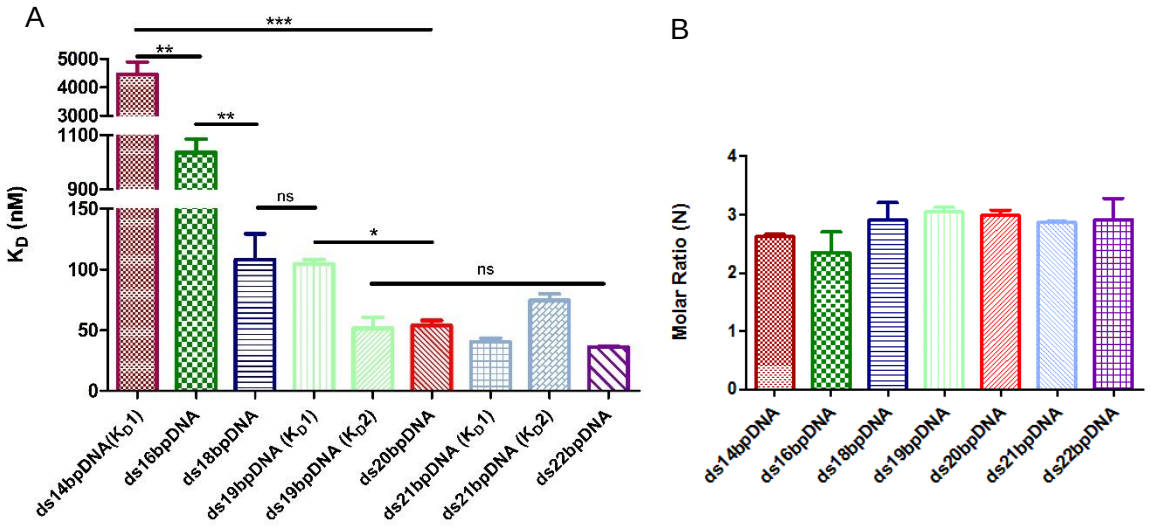

Figure 3-14. Histogram of affinity and stoichiometry of Tfcp2l1 DBD_{19-260} binding to DNA variants of different length. A, The binding affinity changes significantly when the DNA increases in length from 14-mer to 19-mer, while there is no significant change between 19-mer and 22-mer. For ds19bpDNA and ds21bpDNA K_{D1} and K_{D2} have the same color, respectively. B, stoichiometry of Tfcp2l1 DBD_{19-260} binding to DNA variants is constant. K_D value of single-site mutated DBD versus wild-type DBD in One-way ANOVA test. *P*<0.05*; *P**<0.01*; *P***<0.001*. Experiments were done in duplicates.

3.5.2 Tfcp2l1 DBD:DNA complexes analyzed by RALS

The ITC analyses of Tfcps2l1 DBD:DNA complexes did not yield the expected 2: 1 or 4: 1 stoichiometries of protein monomer to DNA duplex. To validate the molar ratios and explain the observed patterns of protein:DNA binding, Tfcp2l1 DBD:DNA complexes were prepared *in vitro* and their molecular mass was determined by RALS.

In agreement with the ITC data, the RALS measurements indicated that Tfcp2l1 DBD binds ds16bpDNA as an apparent dimer (Fig. 3-15 A). For ds18bpDNA, ds19bpDNA and ds21bpDNA, the RALS data do not allow to clearly discriminate between Tfcp2l1 DBD dimer or tetramer binding to DNA (Fig. 3-15 B, C, and E). Because these DNA variants were not considered promising targets for further analysis, the molar ratio of protein to DNA was not further analyzed. The RALS results showed that Tfcp2l1 DBD bound both ds20bpDNA and ds22bpDNA complexes as apparent tetramer, i.e. four Tfcp2l1 DBD molecules bind to one DNA double strand (Fig. 3-15 D and F).

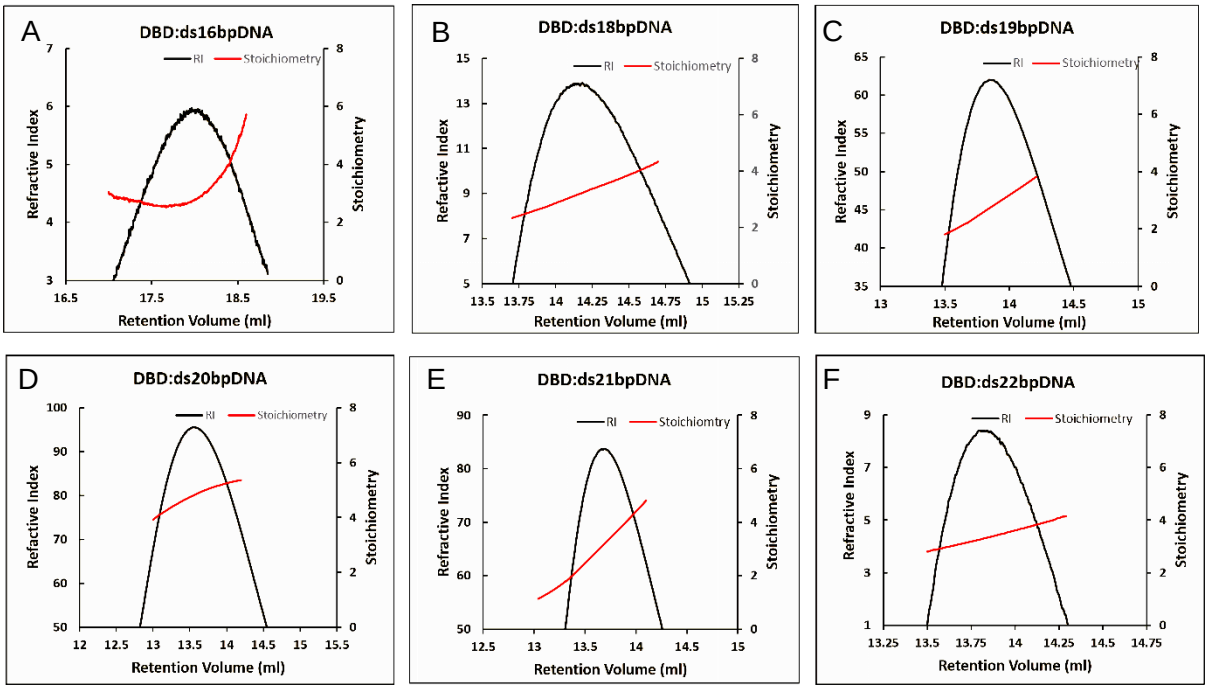

Figure 3-15. Characterization of Tfcp2l1 $DBD_{19\text{-}260}$ binding to different length DNAs. A-F, RALS analyses of Tfcp2l1 DBD in complexes with A: ds16bpDNA, B: ds18bpDNA, C: ds19bpDNA, D: ds20bpDNA, E: ds21bpDNA and F: ds22bpDNA. The RALS system coupled with Superdex 200 Increase 10/300 GL column. The refractive index from RALS is shown in the black curve. The stoichiometry values are shown in the red line indicating protein molecules to duplex DNA in the protein: DNA complex.

3.5.3 Tfcp2 DBD:DNA complex study

As described before, it is known that hTfcp2 specifically binds to the 12-mer DNA motif AAAACCGGTTTT[49]. The construct Tfcp2 DBD_{60-288} was employed in binding tests to elucidate target DNA recognition by Tfcp2 DBD.

In ITC assays, Tfcp2 DBD_{60-288} binding to both ds12bpDNA (AAAACCGGTTTT) and ds12bpAG DNA (AAAACCAGTTTT) could be fitted using the one-set binding model. Compared to Tfcp2l1 DBD binding to ds12bpDNA and ds12bpAG DNA, Tfcp2 shows weaker binding to these DNA sequences with K_D values of 247 nM and 362 nM, respectively (Fig. 3-16 A and B). Tfcp2 DBD_{60-288} also could bind to ds20bpDNA with a K_D value of 224 nM after fitting the binding isotherm by the one-set binding model (Fig. 3-16 C).

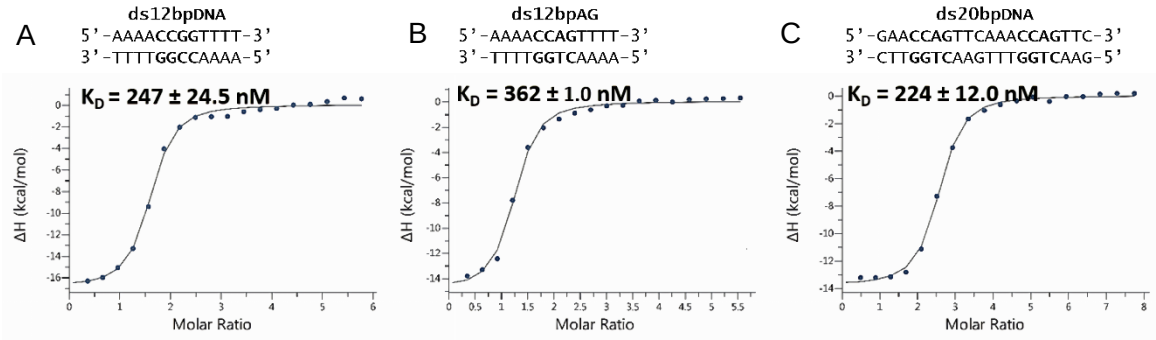

Figure 3-16. Quantitative analysis of Tfcp2 DBD binding to specific DNA motifs. A-C, Tfcp2 DBD_{60-288} titrated to ds12bpDNA, ds12bpAG and ds20bpDNA. The titration curves are fitted in the one-set binding model. Experiments were done in duplicates.

The RALS results showed that Tfcp2 DBD binds to both ds12bpDNA and ds12bpAG (Fig. 3-17 A and B). According to the chromatography, the RALS analysis shows that the complexes are not homogeneous, but rather a mixture of DNA-bound Tfcp2 monomers and dimers or dimers and tetramers. One possibility is that the protein:DNA complex solution is too viscous during the SEC forming the tailing peak. The same situation is observed for the Tfcp2 DBD:ds20bpDNA complex (Fig. 3-17 C), where a mixture of DNA-bound Tfcp2 dimers and tetramers are likely. Apparently, Tfcp2 binding to the target DNA is much weaker than DNA binding by other members of the Grh/CP2 family. It is unclear if the DNA sequence AAACCGGTTT is not the true Tfcp2 target sequence or Tfcp2 binds its specific target sequence only with 300 nM affinity.

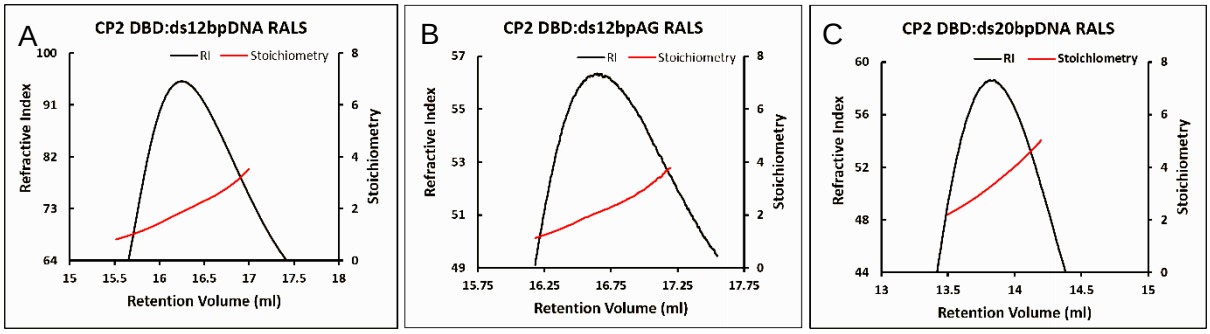

Figure 3-17. Characterization of Tfcp2 DBD: DNA complexes. RALS analysis of Tfcp2 DBD_{60-288} (CP2 DBD) in complexes with A: ds12bpDNA, B: ds12bpAG, and C: ds20bpDNA. The refractive index from RALS is shown in the black curve. The stoichiometry values shown as red lines indicate the number of protein molecules bound to the duplex DNA in the protein: DNA complex.

3.6 Tfcp2l1 DBD:DNA co-crystallization

3.6.1 Co-crystallization Tfcp2l1 DBD with ds20bpDNA

Based on the ITC data and RALS experiments, the ds20bpDNA was employed in co-crystallization with Tfcp2l1 DBD_{19-260}. The protein:DNA complex was prepared assuming a 4: 1 stoichiometry of protein to DNA and using a small molar excess of protein over DNA with the molar ratio of 4.1: 1. The solution with this complex was incubated at 4 ºC for 30 min. The sample was then loaded onto the equilibrated Superdex 200 Hiload 16/60 column. The excess protein was separated from the protein:DNA complex in the gel filtration. The protein:DNA complex peak showed a smaller retention volume and a larger A260/A280 ratio value compared to pure protein. The protein:DNA complex sample was collected and concentrated to 8 - 9 mg/ml, then the sample was set up for crystallization screening.

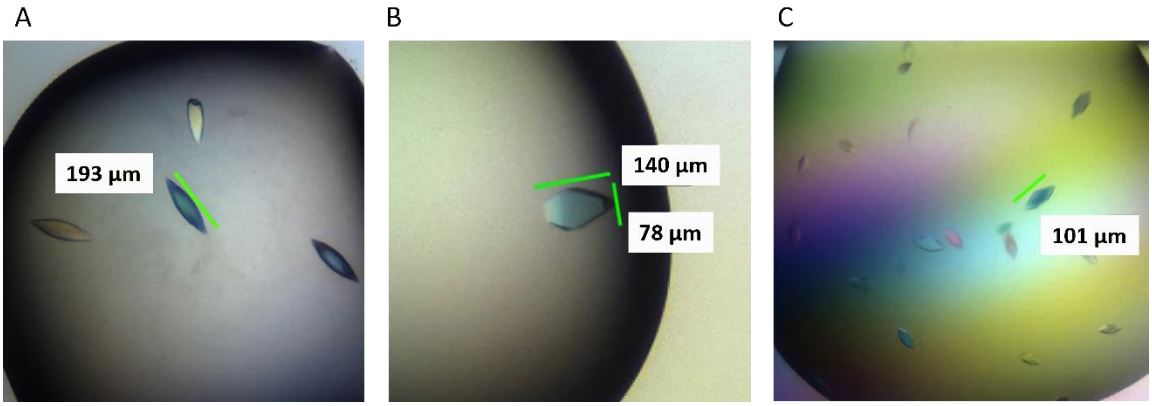

Figure 3-18. Co-crystallization of Tfcp2l1 DBD with ds20bpDNA. A, crystals from 0.55 M Na acetate, 0.1 M imidazole, pH 6.8. B, crystals from 1.0 M Na^+/K^+ phosphate, pH 6.3. C, crystals from 1.0 M di-sodium malonate, pH 7.0. All crystals grew at 4 °C.

After 21 days, crystals were observed at 4 ºC in the screen condition: 0.55 M sodium acetate, 0.1 M imidazole pH 6.8. Larger and well-shaped crystals were obtained after fine screening (Fig. 3-18 A). All crystals have a smooth rhombus shape. X-ray diffraction only extended to 5 - 6 Å resolution. Meanwhile, crystals were observed in two further screen conditions: 1) 1.0 M sodium-potassium phosphate pH 6.3; 2) 1.0 M di-sodium malonate pH 7.0 (Fig. 3-18 B and C).

A crystal from the fine screen based on the Na^+/K^+ phosphate condition yielded X-ray diffraction to 5.82 Å which was considered insufficient for structure determination. The fine screen based on the disodium malonate condition did not yield any larger crystals. In these three crystallization conditions, the fine screen window is narrow; a slight change either in salt concentration or pH value in the mother liquor is sufficient to significantly affect crystal growth. Finally, the optimized condition which yielded the largest crystals was selected and set up in one whole 96-well plate.

3.6.2 Crystal dehydration

Preliminary analysis of the X-ray diffraction data showed that the crystals did not diffract well. A dehydration procedure was employed to improve the quality of crystals from the fine screen condition with sodium acetate. Because the fine screen was set up in sitting drops, there is a danger that the crystal will be harmed during transfer to another condition. The method was therefore modified to change the mother liquor in the deep-well plate without touching the crystal.

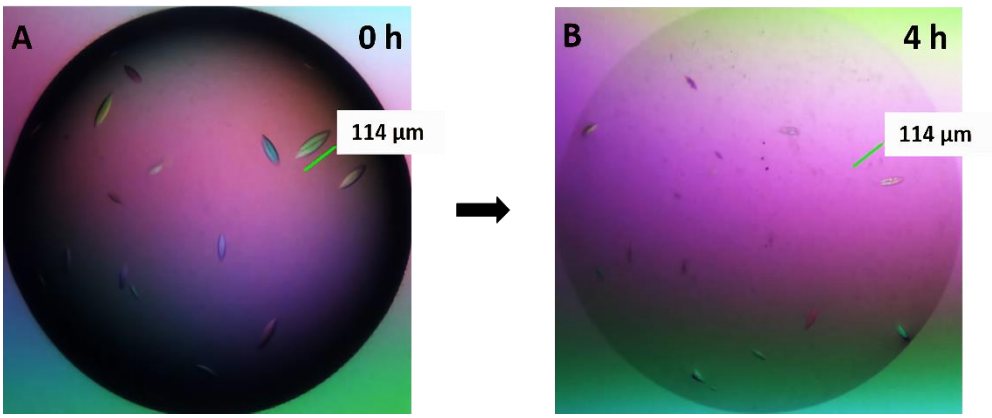

Figure 3-19. Crystal dehydration of Tfcp2l1 DBD in the presence of ds20bpDNA. A-B, the mother liquor of the 96-deep-well plate was replaced with a solution with higher salt concentration (1.8 M Na acetate, 0.1 M imidazole, pH 6.75). Images were taken after 0 h and 4 h.

Five different concentrations of Na acetate were prepared: 0.8 M, 1.0 M, 1.2 M, 1.8 M, and 3.0 M for the crystal dehydration. The solution pH value was kept constant at 6.75. The crystals shrank dramatically during the dehydration in the high concentration of mother liquor (1.8 M Na acetate, 0.1 M imidazole, pH 6.75), and the crystals even disappeared after six-hour dehydration (Fig. 3-19). The surviving crystals from the dehydration condition were mounted and tested for X-ray

diffraction. Unfortunately, the dehydrated crystals showed weaker diffraction and poorer resolution than the original crystals. It is concluded that crystal dehydration is not an effective way to improve the diffraction quality and resolution of these crystals.

3.6.3 *In situ* diffraction

In situ X-ray diffraction data collection is a developing method to analyze crystals as they grow[177]. In other words, for *in situ* or in plate diffraction there is no need to harvest and manually handle crystals, thus avoiding damage or loss of protein crystals[178]. Therefore, using *in-situ* diffraction might improve crystal diffraction and resolution. During the diffraction experiment, the crystal growth medium and the crystallization plate contribute to the scattering background[178]. The unique plate or classical loop were applied to increase diffraction signal-to-noise ratio (SNR) and minimize background effects [179].

The Tfcp2l1 DBD:ds20bpDNA complex was screened in the In Situ-1™ crystallization plate from *MiTeGen* in the HZB-MX BioLab. The In Situ-1™ crystallization plate has two different chambers: a narrow chamber and a broad chamber. 35 µl mother liquor was pipetted into the narrow chamber, and multiple drops of 0.2 µl each were placed into the large chamber. Crystals were observed after one month, but they were smaller than crystals from the original optimization. Therefore, the plate was not measured at the BESSY beamline.

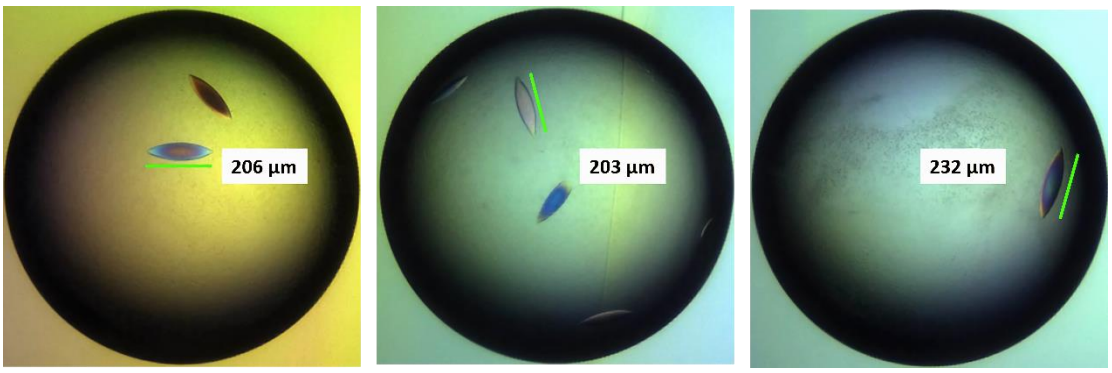

Figure 3-20. Crystals of ds20bpDNA grown in presence of Tfcp2l1 from the optimized fine screen. All crystals have a similar rhombus shapes. Crystals grew at 4 ºC for three weeks before testing at BESSY.

An alternative method to perform *in situ* diffraction uses the classical loop to mount the crystals and test the crystals in the condition in which they grew. The Tfcp2l1 DBD:ds20bpDNA complex was set up in the same condition over one whole plate, and crystals were observed (Fig. 3-20). The plate was brought to the beamline, and the crystals were mounted without cryoprotectant and

tested immediately. Ten crystals were tested, and two images (0º and 19º) were collected from each crystal. In total, we collected about 20 images with 5.2- 5.5 Å resolution and tried to analyze the structure model of the Tfcp2l1 DBD:ds20bpDNA complex. However, electron density was observed for the ds20bpDNA molecule only. It is concluded that these kinds of crystals only contain DNA double strands.

However, according to the ITC assays and RALS analysis, the complex with Tfcp2l1 DBD and ds20bpDNA is stable. It seems that the high salt concentration of 1) 0.5 M Na acetate, 2) 1.0 M sodium malonate, and 3) 1.0 M Na^+/K^+ phosphate, dissociated the protein:DNA complex during the crystallization.

3.6.4 ITC test of Tfcp2l1 DBD binding to DNA in high salt concentration buffer

As the application of Le Chatelier's principle shows, the stability of DNA double-strands increases as the salt concentration increases, but high salt concentration destabilizes DNA-protein complexes[180,181]. Therefore, a high-salt ITC buffer of 25 mM HEPES pH 7.2, 500 mM NaCl, and 0.5 mM TCEP, was employed to analyze the affinity of the protein to DNA *in vitro*.

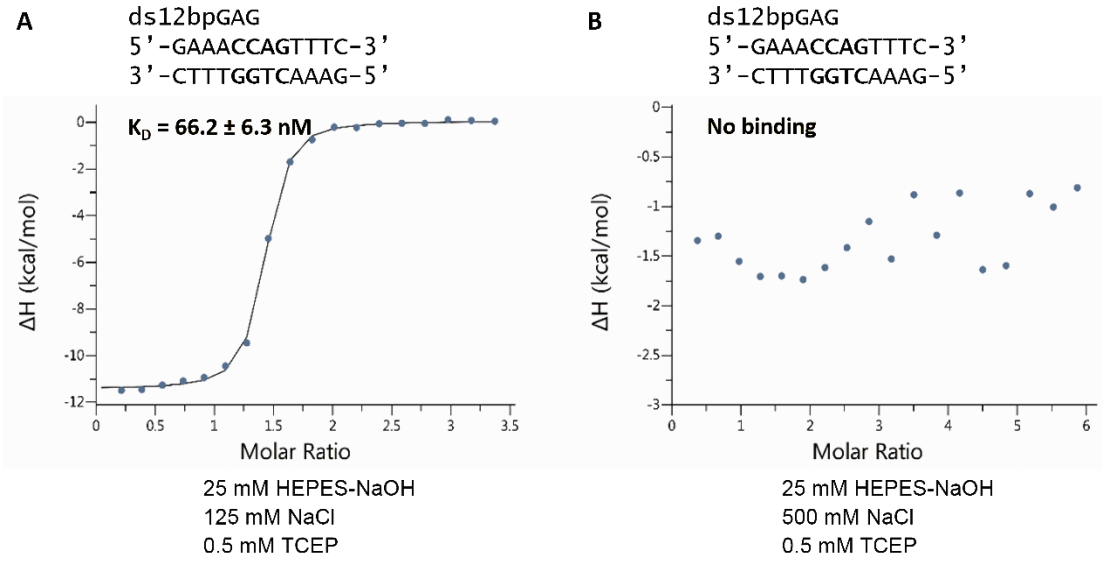

Figure 3-21. Characterization of Tfcp2l1 DBD binding to ds12bpGAG in different buffers. A, Tfcp2l1 $DBD_{19\text{-}283}$ tightly binds to ds12bpGAG DNA with K_D = 66.2 nM in 125 mM [Na]. B, Tfcp2l1 $DBD_{19\text{-}283}$ could not bind to ds12bpGAG DNA in 500 mM [Na].

To quickly determine the protein:DNA binding affinity and the binding pattern, ds20bpDNA was replaced by ds12bpGAG (GAAACCAGTTTC) to perform the ITC experiment. Both Tfcp2l1 DBD and ds12bpGAG were dialyzed into two different buffers: the high salt buffer (500 mM NaCl) and

the standard ITC buffer (125 mM NaCl). According to the ITC experiments, Tfcp2l1 DBD tightly binds to ds12bpGAG with K_D = 66.2 nM in low-salt buffer. Surprisingly, Tfcp2l1 DBD binding to ds12bpGAG could not be detected in the high-salt condition (Fig. 3-21). The ITC results demonstrated that Tfcp2l1 DBD could tightly bind to the unique DNA sequence ds12bpGAG, which could be perturbed by high salt concentration.

3.7 Co-crystallization of Tfcp2l1 DBD with DNA variants

3.7.1 Co-crystallization Tfcp2l1 DBD with DNA variants in 125 mM [Na] buffer

Based on the ITC experiments, complexes formed by Tfcp2l1 DBD_{19-260} with DNA variants were prepared *in vitro* with a molar excess of protein over DNA double strands at the anticipated stoichiometries. These protein:DNA complexes: DBD_{19-260}:ds16bpDNA, DBD_{19-260}:ds18bpDNA, DBD_{19-260}:ds19bpDNA, DBD_{19-260}:ds21bpDNA, and DBD_{19-260}: ds22bpDNA, were purified by size exclusion chromatography to remove the excess protein. The complexes were concentrated and set up the crystallization. However, all these complexes failed to yield any crystals in the initial screen.

3.7.2 Co-crystallization Tfcp2l1 DBD with ds12bpDNA and ds12bpAG

A member of the Grh subfamily of transcription factors, Grhl1 DBD, could be co-crystallized with a 12-mer DNA fragment[23]. As mentioned before, CP2 subfamily members were reported to bind to a tandem repeat consensus sequence: GAACC$^{A}/_{G}$GTTTGAACC$^{A}/_{G}$-GTTC. By analogy, it was therefore attempted to co-crystallize Tfcp2l1 DBD_{19-260} with 12-mer DNA duplexes centered on the core motif CC$^{A}/_{G}$G under low-salt buffer conditions. ds12bpDNA (AAAACCGGTTTT) and ds12bpAG DNA (AAAACCAGTTTT) were synthesized and applied for co-crystallization. It should be pointed out that the ds12bpDNA sequence is strictly self-complementary, whereas the ds12bpAG sequence is not.

Before co-crystallizing Tfcp2l1 DBD_{19-260} with the new DNA variants, ITC experiments were applied to measure the binding affinities in 125 mM [Na] concentration buffer. DBD_{19-260} could tightly bind to both ds12bpDNA with a K_D value of 66.1 nM and the ds12bpAG DNA with a K_D value of 167 nM (Fig. 3-27 A and D). The K_D value of DBD_{19-260} binding to ds12bpAG is two-fold that of ds12bpDNA, suggesting Tfcp2l1 prefers to bind the AAAACCGGTTTT over the AAAACCAGTTTT DNA sequence.

Both the Tfcp2l1 DBD_{19-260}:ds12bpDNA and DBD_{19-260}:ds12bpAG complexes were prepared with a molar protein (monomer) to DNA (duplex) ratio of 2.1: 1. Excess protein was removed by SEC (Fig 3-22 A). The Tfcp2l1 DBD_{19-260}:ds12bpDNA complex was concentrated to three different concentrations: 8.5 mg/ml, 12 mg/ml, and 17.5 mg/ml. The DBD_{19-260}:ds12bpAG complex was concentrated to two different concentrations: 8.7 mg/ml and 11 mg/ml. All samples were subjected to initial screens both at 4 ºC and 20 ºC with the screen suites: pH Clear I, pH Clear II, JBS Basic, ProComplex, and Nuc-pro suite.

Tfcp2l1 DBD_{19-260}:ds12bpDNA complex crystals were observed after 49 days at 4 ºC at 8.5 mg/ml in two conditions: 0.1 M Na^+ citrate pH 4.0, 20% v/v isopropanol and 0.1 M Na citrate pH 5.0, 30% v/v isopropanol (Fig. 3-22 B). Crystals from the initial screen were mounted and tested at BESSY in remote control mode. 2.3 Å resolution X-ray diffraction data were collected. A crystal from the fine screen yielded 1.73 Å resolution diffraction data.

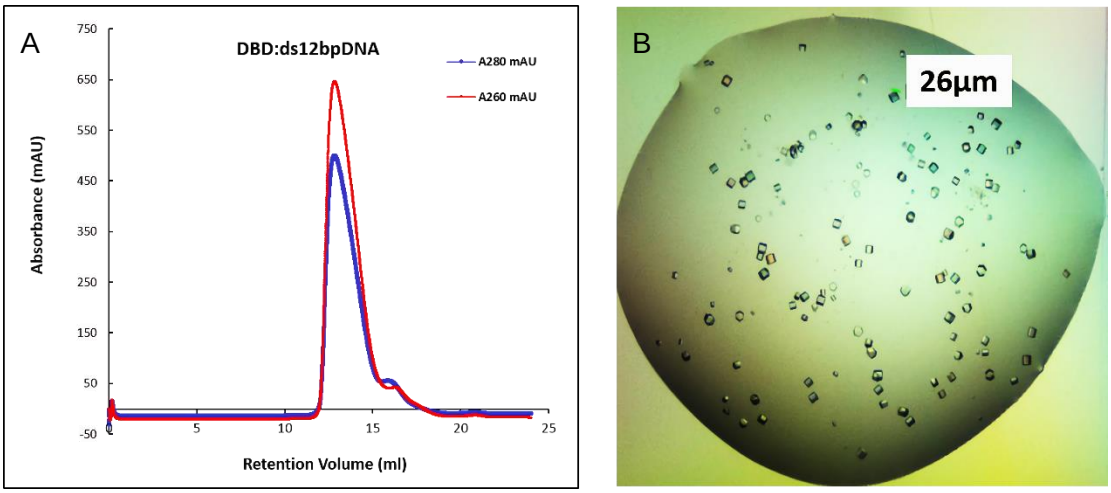

Figure 3-22. Complex of Tfcp2l1 DBD_{19-260} bound to ds12bpDNA. A, protein: DNA complex was purified by SEC. B, complex crystals from the initial screen: 0.1 M Na^+ citrate pH 4.0, 20% v/v isopropanol. Crystals grew at 4 °C.

3.7.3 Structure of the Tfcp2l1 DBD:DNA complex

The published structure of the Grhl1 DBD:DNA complex was used as the template for the Tfcp2l1 DBD_{19-260}:ds12bpDNA complex phase determination by molecular replacement. The anisotropy of the diffraction data was firstly checked by the sfcheck program from CCP4 suite. Using these data, I determined and refined the crystal structure of Tfcp2l1 DBD_{19-260} bound to ds12bpDNA (AAAACCGGTTTT) at 1.73 Å resolution with R_{work} = 16.55% and R_{free} = 19.83% (Fig. 3-23 A). Diffraction data and model refinement statistics are summarized in Table 3.2.

The Tfcp2l1 DBD_{19-260}:ds12bpDNA crystal structure contains two DBD molecules and one duplex 12-mer DNA in each asymmetric unit (Fig. 3-23 A). The two Tfcp2l1 DBD molecules are arranged almost symmetrically to opposite faces of the 12-mer DNA duplex, which adopts a standard B-form geometry (Fig. 3-23 A). The two DBDs form a dimeric interface with helix α3 and loop L10 interacting with the DNA major and minor groove, respectively, both protein elements corresponding to positively charged regions (Fig. 3-23 B).

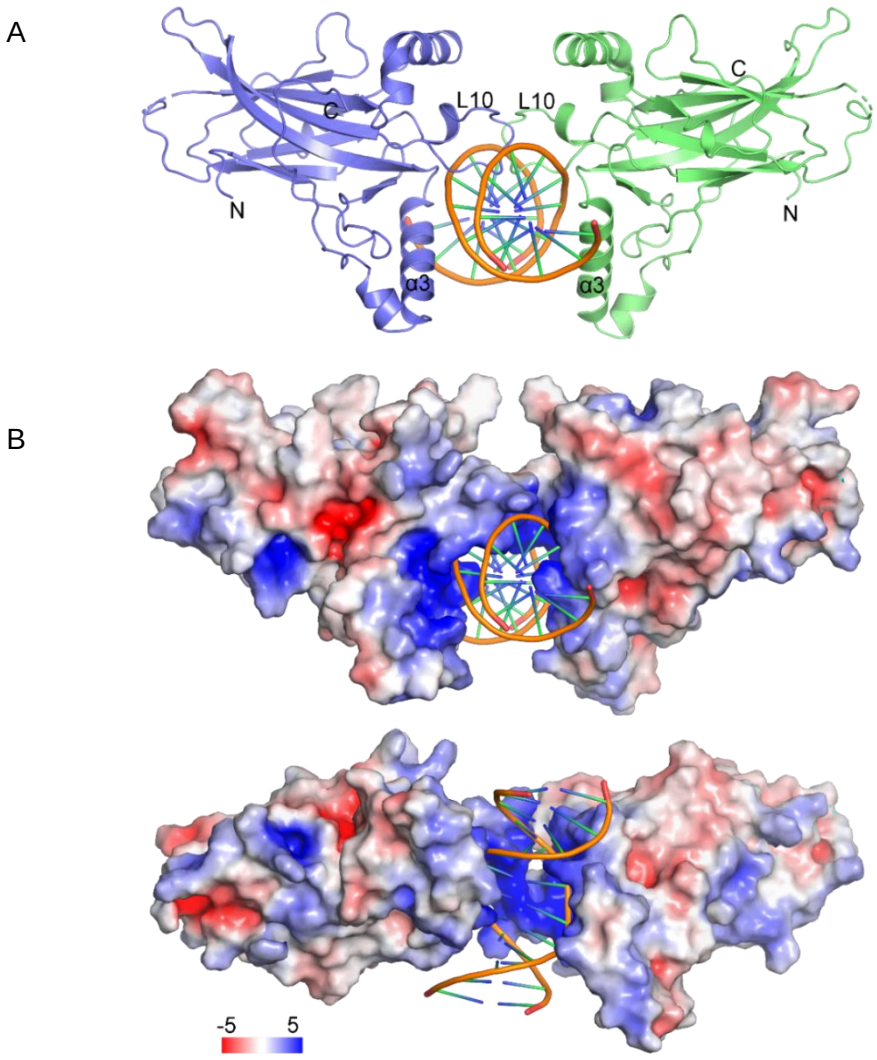

Figure 3-23. Overall structure of the Tfcp2l1 DBD:ds12bpDNA complex. A, cartoon drawing of Tfcp2l1 DBD bound to ds12bpDNA. Two Tfcp2l1 DBD molecules (blue and lime) interact with duplex DNA mainly through helix α3 located in the major groove of the DNA and loop L10 in the minor groove. B, orthogonal views of the electrostatic potential surface of DNA bound Tfcp2l1 DBD molecules (0º and 90º), colored with positive potential (+5 kT) in blue and negative potential (-5 kT) in red.

Table 3.2 Data collection and refinement statistics

	Tfcp2l1$_{19-260}$: ds12bpDNA	Tfcp2l1$_{19-260}$	Tfcp2$_{60-275}$
Data collection			
Beamline	BESSY 14.1	BESSY 14.1	BESSY 14.1
Wavelength (Å)	0.91841	0.91841	0.91841
Space group	$P3_1$	$P4_32_12$	R32
Cell dimensions			
a, b, c (Å)	101.4, 101.4, 45.8	67.4, 67.4, 110.3	157.2, 157.2,151.5
α, β, γ (°)	90.0, 90.0, 120.0	90.0, 90.0, 90.0	90.0, 90.0, 120.0
Resolution (Å)*	45.79 - 1.73 (1.83 - 1.73)	47.69 - 1.93 (2.05 - 1.93)	48.73 - 2.72 (2.88 - 2.72)
R_{meas}* (%)	8.9 (193.0)	9.3 (188.8)	9.7 (264.3)
$< I / \sigma(I) >$*	10.62 (0.83)	16.63 (1.08)	15.30 (0.97)
$CC_{1/2}$*	99.8 (39.3)	99.9 (42.3)	99.9 (39.1)
Completeness* (%)*	99.9 (99.7)	99.9 (99.8)	99.7 (99.4)
Multiplicity	5.2	7.5	8.4
No. unique reflections*	54,975 (5,463)	19,811 (1,913)	19,520 (3097)
Refinement			
R_{work} / R_{free} (%)	16.55 / 19.83	17.69 / 22.34	23.44 / 26.85
No. atoms			
Protein	4,037	1,615	3,092
DNA	242	-	-
Ligand	16	12	0
Water	380	174	7
Mean *B* factor (Å^2)	38.8	53.8	139.1
R.m.s deviations			
Bond lengths (Å)	0.010	0.006	0.020
Bond angles (°)	1.092	0.833	1.730
Mol/AU	1	1	2
Ramachandran (%)			
favored	98.13	98.43	98.10
allowed	1.87	1.57	1.90
outliers	0.00.	0.00	0.00

* Data in highest resolution shell are indicated in parenthesis. Result files generated for CCP4, CNS, and SHELX.

The C-terminus of the two Tfcp2l1 DBD molecules is oriented towards the same direction of the duplex DNA, perpendicular to the DNA axis. It is assumed that the SAM domain would extend from thereon and mediate interactions between the monomers, leading to Tfcp2l1 dimer- or oligomerization.

The L10 loops are accommodated in the minor groove of DNA. The two bound DBD chains contact each other. Residues K180 and H181 at the L10 loop and residue T174 mediates a possibly significant DBD-DBD contact for stabling DNA binding through polar interactions. (Fig. 3-24).

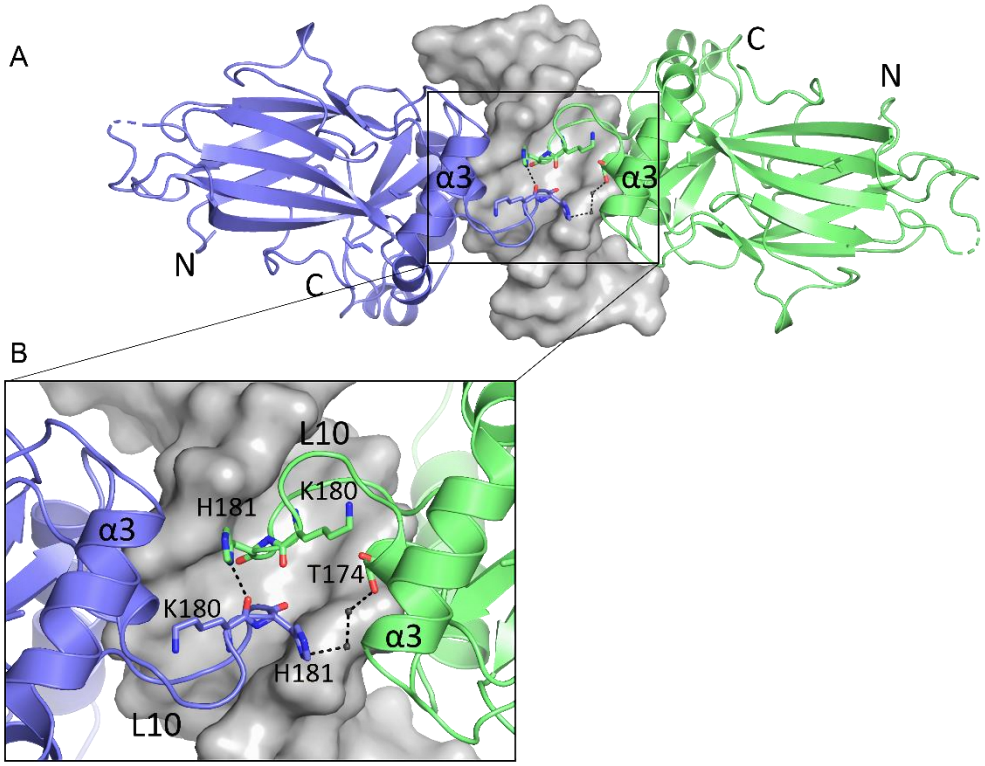

Figure 3-24. Overall view of the Tfcp2l1 DBD-DBD interface region. A, The loops L10 from two Tfcp2l1 DBD monomers are located in DNA minor groove. B, orthogonal close-up views of residues from loop L10 forming the polar interaction to stabilize the Tfcp2l1 DBD dimer based on the residues K180, H181, T174.

The protein:DNA interfaces are decorated with residues from Tfcp2l1 helix α3 interacting with the DNA major groove, and the loop L10 interactions with the DNA minor groove, involving a positively charged surface with the total size of 1562.5 Å. Both specific and unspecific contacts between protein and DNA molecules contribute to the protein:DNA interface.

The specific transcription factor-DNA interactions are formed by two conserved Tfcp2l1 residues, G183 and R225, which contact one conserved guanine base in the consensus DNA binding motif

(Fig 3-25). The R225 guanidino group plays the primary role in anchoring helix α3 to the DNA major groove via hydrogen bonding to the C6 carbonyl and N7 imine groups of guanine G8. Moreover, the G183 carbonyl oxygen provides additional selectivity by binding in the DNA minor groove via hydrogen bonding to the N2 amino group of guanine G8. It is concluded that only the four nucleotides $C_6G_7G_8T_9$ of the bound DNA double strand are involved in binding Tfcp2l1 DBD (Fig. 3-25 C).

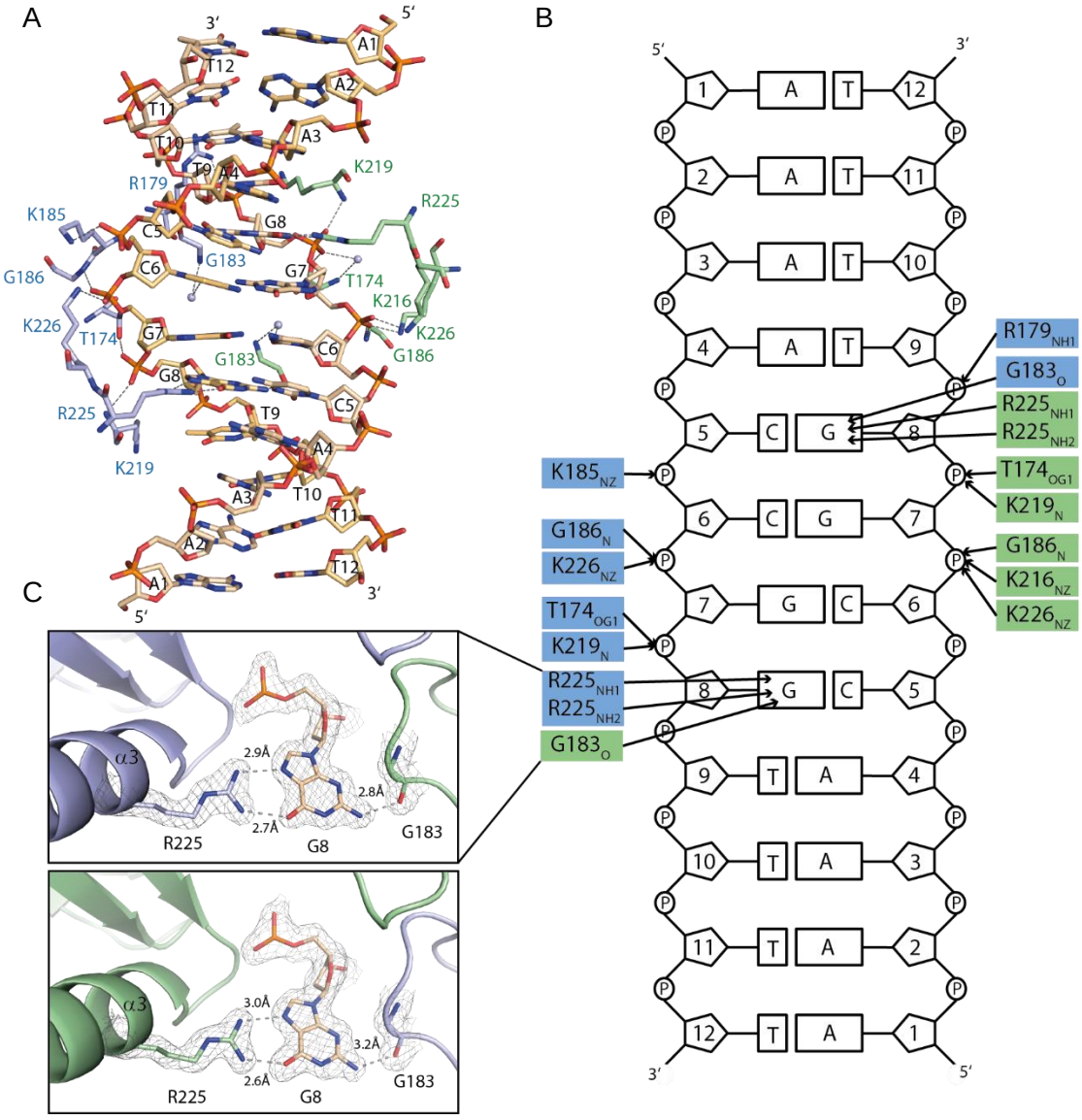

Figure 3-25. Overall view of protein: DNA interactions of Tfcp2l1 DBD:DNA complex. A, the polar interactions and hydrogen-bonding interactions were showed in stick presentation (3.5Å cut-off). Tfcp2l1 DBD residues from two chains are color slate (chain A) and lime (chain B). B, scheme view of protein: DNA interaction corresponding to the A were showed in arrows. C, close-up views of the specific interactions of R225 and G183 with guanine G8 from two chains. *2Fo-Fc* electron density map was contoured at 0.8 σ (gray).

Unspecific interactions, including hydrogen bonds and Coulomb contacts, are mediated by residues T174, R179, K185, G186, K216, K219 and K226 of Tfcp2l1 DBD (Fig 3-25 A and B). However, these residues' interaction with the DNA backbone are not identical in the individual Tfcp2l1 DBD chains. Residues T174, G186, K219, and K226 from the two protein chains of the complex (chain A blue, chain B lime) form contacts with phosphate groups of the DNA backbone. T174 and K219, G186 and K226 interact with phosphate groups of guanosine nucleotides G8 and G7, respectively. However, chain A residue R179 contacts the opposite thymidine nucleotide T9 phosphate group and K185 interacts with the C6 phosphate group. In chain B, both R179 and K185 do not interact with the DNA backbone. Another difference is that K216 from chain B interacts with the G7 phosphate group, an interaction which does not exist in chain A.

3.8 Biochemical studies based on the Tfcp2l1 DBD:DNA structure

3.8.1 Mutations in the Tfcp2l1 DBD affect ds12bpDNA binding

The Tfcp2l1 DBD:DNA complex structure reveals fine details of DNA binding and target site recognition by the transcription factor. To evaluate the contributions of both specific and non-specific interactions to the complex structure, single residue mutants of Tfcp2l1 DBD were generated, which were compared with the wild type Tfcp2l1 DBD regarding the DNA binding affinity.

The mutation R225A completely abrogated the interaction between Tfcp2l1 DBD and ds12bpDNA (AAAACCGGTTTT). This mutation confirmed the prediction from the crystal structure that R225 binding to guanine G8 plays the central role for protein:DNA binding. Mutations of T174, R179, K185, K219 and K226 to alanine, respectively, affected the unspecific protein binding to DNA which is weaker than with the wild-type protein (Fig. 3-26 A-I). Interestingly, the G183A and H181A mutants have a significant effect on the binding affinity. The G183A mutation is associated with a more than twenty-fold decrease in binding affinity, and the H181A mutation with an eight-fold reduction in binding affinity compared to the wild-type. Both H181 and G183 are located within the loop L10 region, confirming the conclusion from the crystallographic analysis that the loop L10 plays an important role in stabilizing protein binding to the DNA in a dimeric arrangement (Fig. 3-24 and Fig. 3-26 C, F and J).

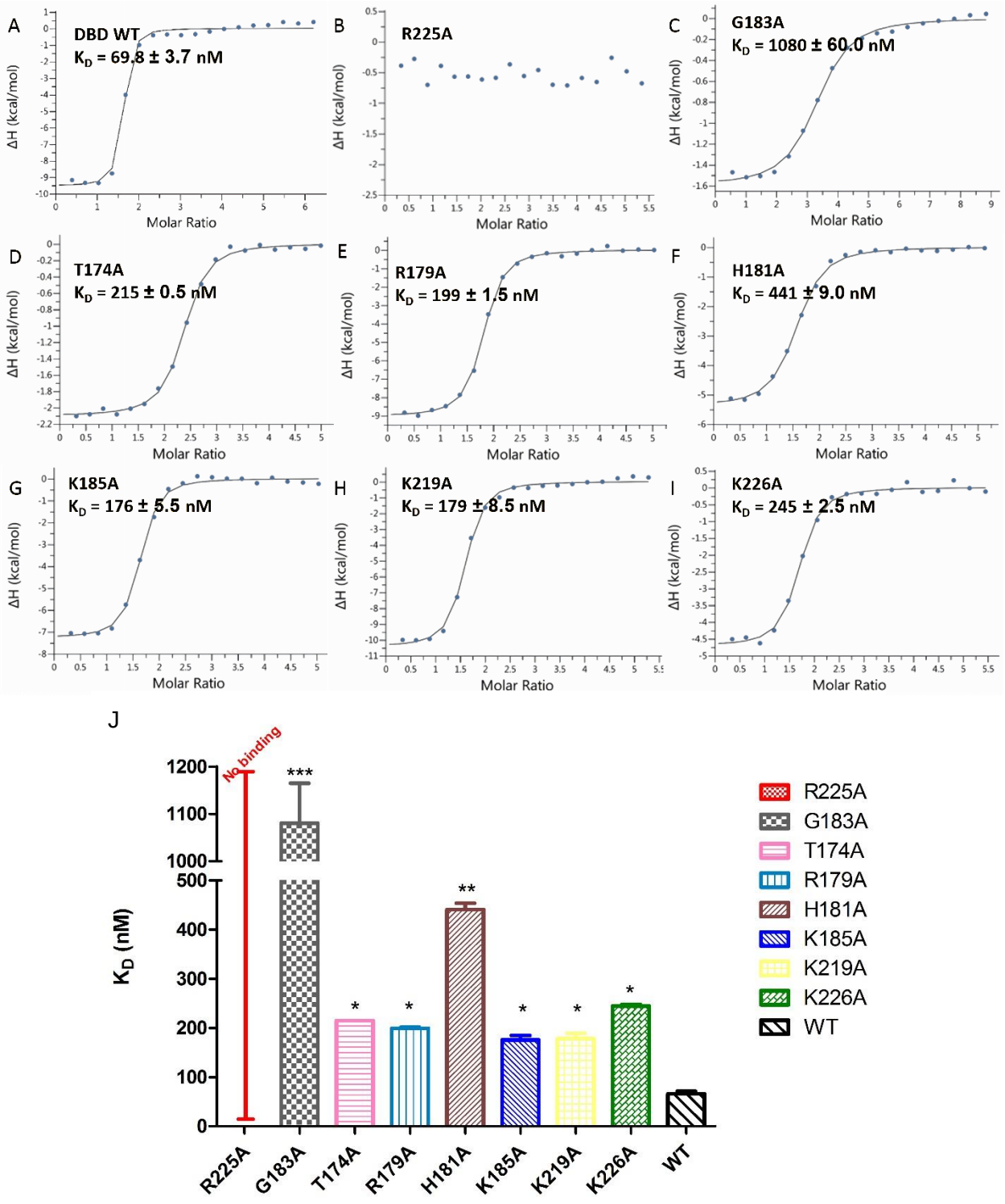

Figure 3-26. Characterization of Tfcp2l1 DBD mutation effects on protein: DNA interactions by ITC. A, DNA binding by wild-type Tfcp2l1 DBD. B-I, ITC measurements of DNA binding by Tfcp2l1 DBD mutants with ds12bpDNA sequence. J, Histogram of K_D values from ITC assays. A K_D value for DNA binding by the Tfcp2l1 R225A mutant could not be determined. K_D value of single-site mutated DBD versus wild-type DBD in One-way ANOVA test. *P*<0.05*; *P**<0.01*; *P***<0.001*. Experiments were done in duplicates.

3.8.2 Tfcp2l1 DBD binds to the specific core DNA sequence

Cytosine C5 and guanine G8 in the 12-bp DNA sequence AAAAC$_5$C$_6$G$_7$G$_8$TTTT are highly conserved as concluded from a previous ChIP-Seq analysis[49]. This is in agreement with the Tfcp2l1 DBD:DNA crystal structure, which identified G8 as the only nucleotide with direct hydrogen bonds to the protein. Mutating C5 and G8 to T5 and A8 completely abolished protein binding to DNA (Fig. 3-27 B). Just mutating G8 to A8 but retaining the C5 nucleotide gives rise to the ds12bpCA DNA which is still bound by Tfcp2l1 DBD but with significantly reduced affinity and requiring a different binding model in evaluating the ITC data (Fig. 3-27 C). It should be pointed out that the ds12bpCA sequence is not strictly self-complementary compared to ds12bpDNA and ds12bpGA sequences.

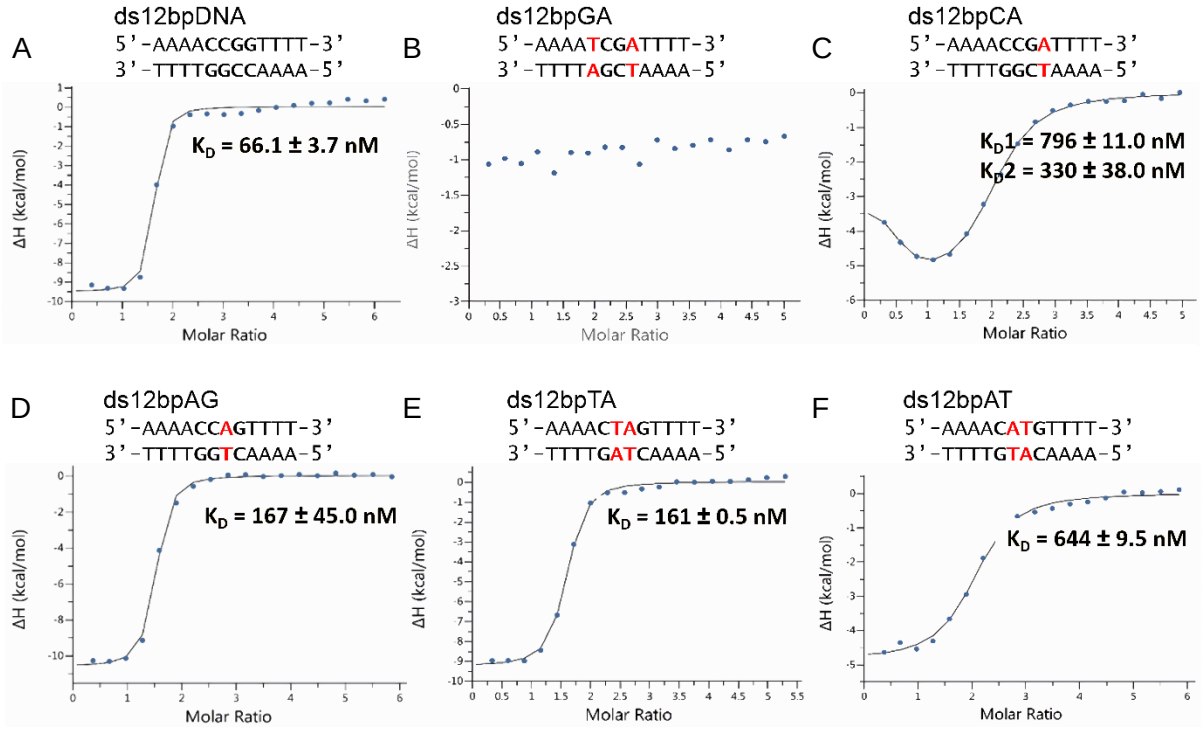

Figure 3-27. Analysis of Tfcp2l1 DBD$_{19\text{-}260}$ binding to mutant DNAs. A and D-F, titration curves fitted with the one-set binding model. C, titration curve fitted with two-set binding model. B, no binding was detected. Experiments were done in duplicates.

To elucidate the preference of Tfcp2l1 DBD in contacting the core DNA sequence, two more DNA duplexes, ds12bpAT and ds12bpTA, were designed and synthesized. As mentioned before, Tfcp2l1 prefers to bind to the C$_5$C$_6$G$_7$G$_8$ core motif rather than the C$_5$C$_6$A$_7$G$_8$ motif. The K$_D$ value increased three-fold as a consequence of the G7 to A7 mutation as measured by ITC assays. Upon double mutation of C6G7 to T6A7 to generate ds12bpTA, the K$_D$ value increased three-fold.

However, in the A6T7 double mutant, the affinity decreased more than ten times compared to the wild-type C6G7 DNA sequence. These data demonstrate that both C5 and G8 are strictly conserved and play a critical role during DNA sequence readout by the transcription factor. At position 6 of the binding site, pyrimidine is preferred and at position 7 a purine base.

3.8.3 DNA motifs bound by Tfcp2l1 are not always separated by six base pairs

Previous reports suggested that Tfcp2l1 binds to the core DNA sequence $CC^{A}/_{G}GN_6CC^{A}/_{G}G$ and N is exactly six base pairs[49]. It was also reported that Tfcp2l1 binds to the *Klf4*, *Esrrb* and *Foxi1* gene promoter regions[138,148]. For example, it was reported that Tfcp2l1 binds to the human *Klf4* gene upstream promoter region (-1101 bp to -1088 bp). However, a sequence search shows that the 50 kbp region upstream of the gene contains only one DNA sequence motif matching the core DNA sequence as defined above. It is the same situation in the *Esrrb* promoter region. It is commonly assumed that transcription factor binding to control gene expression is not a single event involving one specific target DNA sequence. Therefore, multiple Tfcp2l1 binding sites in the regions upstream of the *Klf4* or *Esrrb* promoters are expected. There is a fairly large copy number of the single-core motif of $CC^{A}/_{G}G$, but with one exception these motifs are not separated by six base pairs. To reconcile this observation with the established regulatory role of Tfcp2l1 for these genes, one has to assume that either Tfcp2l1 binding to half-sites of the core DNA sequence is sufficient or that Tfcp2l1 can bind to sites where CCA/GG motifs are spaced by more or less than six base pairs.

To test this latter hypothesis, three more DNA sequences, N5 ($AAACCAGN_5CCAGTTT$), N6 ($AAACCAGN_6CCAGTTT$) and N7 ($AAACCAGN_7CCAGTTT$) were designed and synthesized. The Tfcp2l1 Δ19 (AAs, 19-479) was used to assay Tfcp2l1 binding to these DNA variants. As confirmed before, Δ19 was present as tetramer in solution. Therefore, during the analysis, the molecular mass of the tetrameric Δ19 molecule was used in analyzing binding isotherms.

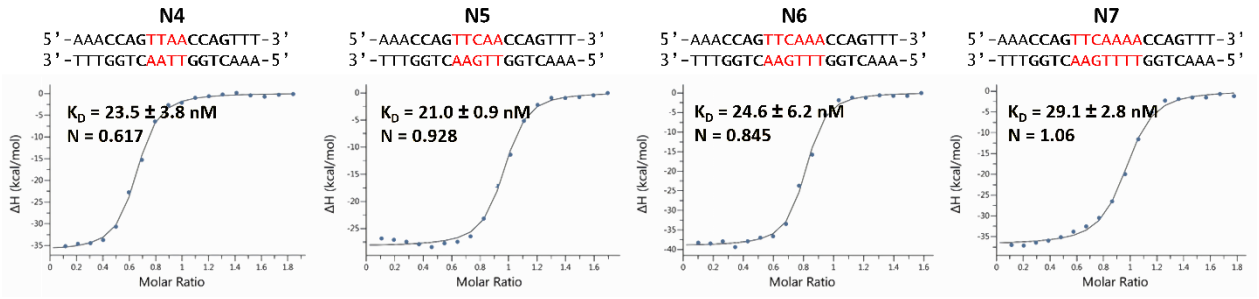

Figure 3-28. ITC measurements of Δ19 binding to DNA variants. Titration curves are fitted with the one-set binding model. Δ19 bind to N4 DNA with a K_D value of 23.5 nM, to N5 DNA with a K_D value of 21.0 nM, to N6 DNA with a K_D value of 24.6 nM, and to N7 DNA with a K_D value of 29.1 nM. Experiments were done in duplicates.

According to the ITC measurements, Δ19 tetramers could bind to N5 and N7 in a similar pattern as to standard N6. The binding stoichiometry in these three assays is close to one (Fig 3-28). The Δ19 tetramer binding to N4 is also best fitted with the one-set binding model, but the stoichiometry drops to 0.617, suggesting substoichiometric binding of Δ19 tetramer to N4 DNA. It is concluded that the number base pairs separating the CCAG core motifs is not strictly limited to six. Tandem motifs with five or seven separating base pairs could also be targeted by the Tfcp2l1 tetramer. Moreover, there is not enough space for Tfcp2l1 tetramers binding to tandem sites with the shorter spacers such as N4 (Fig. 3-28) and N3 (not shown here).

Tfcp2l1 is suggested to binds to consensus tandem DNA motifs in a geometry duplicating Tfcp2l1 DBD binding to ds12bpDNA (AAAACCGGTTTT). The SAM domains perform the tetramerization function as reflected in the ITC titration of tetrameric Δ19 to N6 DNA (Fig. 3-28). The long loop region between the DBD and the SAM domain of Tfcp2l1 provides some conformational freedom to Tfcp2l1 DBD binding to DNA (Fig. 3-28).

4. DISCUSSION

4.1 Special sequence features inside CP2 subfamily proteins

4.1.1 TEV protease cleavage site in CP2 subfamily members

There is one internal tobacco etch virus (TEV) protease cleavage site in the mouse Tfcp2l1 protein sequence. This is also the case in the homologous Tfcp2 and Ubp1 proteins. For purification of affinity-tagged Tfcp2l1 and Tfcp2 proteins, all constructs containing the C-terminal domain were modified by a Q435A mutation to block this TEV cleavage site and allow for proteolytic removal of the tag using TEV.

In contrast to the CP2 subfamily, the three members of the Grhl subfamily do not contain this TEV cleavage site. One hypothesis is that C-terminal residues of CP2 subfamily proteins was cleaved by TEV during purification, and the observed oligomerization was mediated by the SAM domain in the absence of these residues. Tfcp2l1 Δ266-366 is monomeric in solution, and Tfcp2l1 Δ365 is not stable during the purification, demonstrating that the SAM domain performs the oligomerization function. C-terminal residues apparently play a role in stabilizing the Tfcp2l1 protein. However, a structure including the Tfcp2l1 C-terminus was not determined. More evidence is needed to support this hypothesis.

4.1.2 N-terminal peptide

In spite of considerable effort, full-length Tfcp2l1 could not be expressed in *E. coli*. Several different cell lines were employed for the protein expression optimization; target protein expression could be detected in none of them. Interestingly, the N-terminally truncated proteins Tfcp2l1 Δ19 (without residues 1-19), and Tfcp2l1 Δ42 (without residues 1-42) could be expressed and purified in the *E. coli* host strain BL21 DE3 Rosetta2. Further analysis found that within the N-terminal 19 residues of Tfcp2l1 a proline codon CCC and a serine codon UCU are used. These codons are rarely used in *E. coli*, and their presence may have impeded protein expression even though the host strain BL21 DE3 Rosetta2 supplies rare tRNAs to enhance protein expression. This may explain why full-length Tfcp2l1 was expressed weakly or not at all, while N-terminally truncated constructs could be highly expressed in *E. coli*.

In contrast to Grhl subfamily members, CP2 subfamily members do not contain a N-terminal transactivation domain. They have a short N-terminal polypeptide sequence of 19 residues in Tfcp2l1, about 40 residues in Tfcp2 and Ubp1. The CP2 subfamily split from the larger Grh/CP2 family about 700 million years ago[5]. One hypothesis is that the N-terminal transactivation domain

(NTD) was lost during this event and the CP2 subfamily proteins acquired new functions. For example, Tfcp2l1 was initially found to act as a transcription repressor which down-regulates gene expression[30]. Furthermore, Grhl1 isoform 2 has been reported to function as a repressor, lacking the NTD[4]. Although the post-translational modified Tfcp2l1 has been reported to activate the following gene transcription (sections: 1.4.5), it did not involve the N-terminal region[43].

4.1.3 DNA binding region

Based on the secondary structure prediction, Tfcp2l1 DBD starts from L47. Consequently, the first protein construct was designed to start with residue L47. However, the N-terminal His-tag of this recombinant protein could not be cleaved, suggesting that the N-terminal region of Tfcp2l1 DBD may extend beyond residue L47. Therefore, new constructs were designed with N-terminus at residue Y19. The C-terminus of Tfcp2l1 DBD was placed at residue G283. The constructs Tfcp2l1 DBD_{47-283} and Tfcp2l1 DBD_{19-283} contain an extended loop region (AAs, 261-283). As SDS-PAGE showed, Tfcp2l1 DBD_{19-283} preparations always contained an impurity which could not be removed. Sequence alignment of Grh/CP2 family members indicated that the loop region of Tfcp2l1 could be shortened to terminate at residue W260, yielding a protein capable of completing the Ig-like fold. The new construct Tfcp2l1 DBD_{19-260} was therefore designed. After the purification, SDS-PAGE showed the protein preparation to be homogeneous without any impurity. Both Tfcp2l1 DBD_{19-260} and Tfcp2l1 DBD_{19-283} were employed the ITC measurements of *in vitro* DNA binding. ITC assays showed that Tfcp2l1 DBD_{19-260} and Tfcp2l1 DBD_{19-283} could bind to the same DNA double strands with the same affinity (Fig. 3-7 G and H). It is concluded that the C-terminus of Tfcp2l1 DBD ends with residue W260, not G283.

The peptide region from Y19 to G41 remains without electron density in the Tfcp2l1 DBD crystal structure, which starts from residue R42 and extends to W260. The Tfcp2l1 DBD crystal structure contains three more alanine residues after W260, which are from the expression vector. These three alanines neither influence protein DNA binding nor are they required for maintaining the native structure.

4.2 DBD structures are conserved in the Grh/CP2 family

4.2.1 Tfcp2l1 DBD and Tfcp2 DBD structures are similar

The asymmetric unit of the Tfcp2 DBD crystal contains two protein chains. Least-squares superimposition of these two chains yielded a root-mean-square deviation (RMSD) of 0.39 Å between Cα atoms demonstrating closely similar conformation (Fig. 4-1 A). Tfcp2l1 DBD (AA, 19-260) and Tfcp2 DBD (AAs, 40-280) protein sequence share 84% identity which led to the

expectation that the two structures are similar. It was confirmed that the two structures are highly similar based on the RMSD value of 0.603 Å between Tfcp2l1 DBD and Tfcp2 DBD chain A after structure alignment (Fig. 4-1 B).

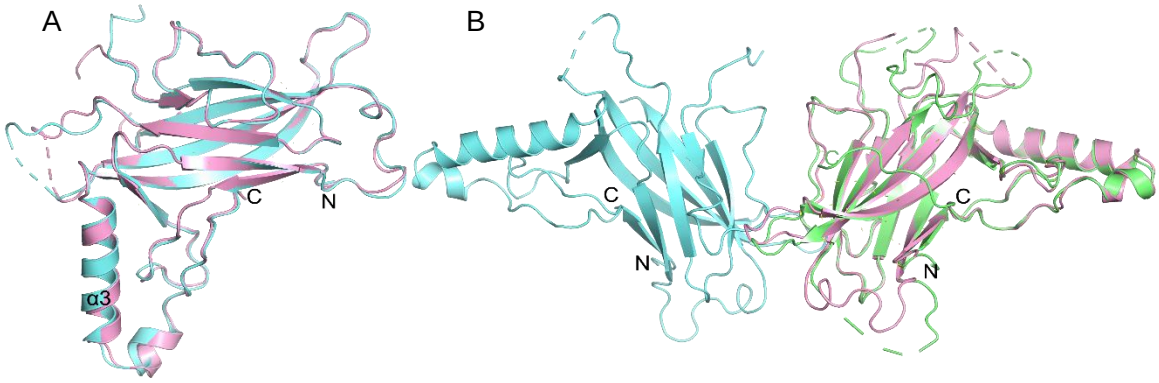

Figure 4-1. Tfcp2l1 DBD and Tfcp2 DBD structure alignment. A, least-squares superposition of the two Tfcp2 DBD chains present in the asymmetric unit yielded a RMSD value of 0.39 Å for 162 aligned Cα atoms (models colored in pink and aquamarine, respectively). B. Tfcp2l1 DBD (colored in lime) superimposed onto Tfcp2 DBD chain A (pink) yielded an RMSD value of 0.603 Å for 139 aligned Cα atoms.

4.2.2 Tfcp2l1 and Tfcp2 DBD structures are similar to Grhl1/2 DBD structures

Sequence alignment shows that the DNA-binding domains (DBD) of Grh/CP2 family members are highly conserved. The homology between these domains is confirmed by DBD structure alignment. Superimposing the Tfcp2l1 DBD onto the Grhl1 DBD (PDB: 5MPI) yields an RMSD value of 1.828 Å for 157 matching Cα atoms in Coot[167] (Fig. 4-2 A). This demonstrated that the Tfcp2l1 DBD and Grhl1 DBD structures are similar. In the Grhl1 DBD structure, thirteen β strands are forming an immunoglobulin (Ig)-like domain. Three helices, α1, α2 and α3, are decorating the Ig-like domain. In the Tfcp2l1 DBD structure, the Ig-like domain is formed by eight β-strands, and helices α3 and α4 complete the domain structure. In Tfcp2l1 DBD, a helix α1 as present in Grhl1 DBD is not seen in the electron density, presumably because this helix is flanked by two floppy peptide regions. Tfcp2l1 DBD has one extra helix α4 following helix α3, which switches the helix α3 orientation and determines the C-terminal loop orientation. Tfcp2l1 DBD lacks five β strands present in Grhl1 DBD, which are annotated as polypeptide loops, the first strand following strand β1 of Grhl1 DBD, the second following helix α1, the third following α2, and two more strands following Grhl1 DBD helix α4.

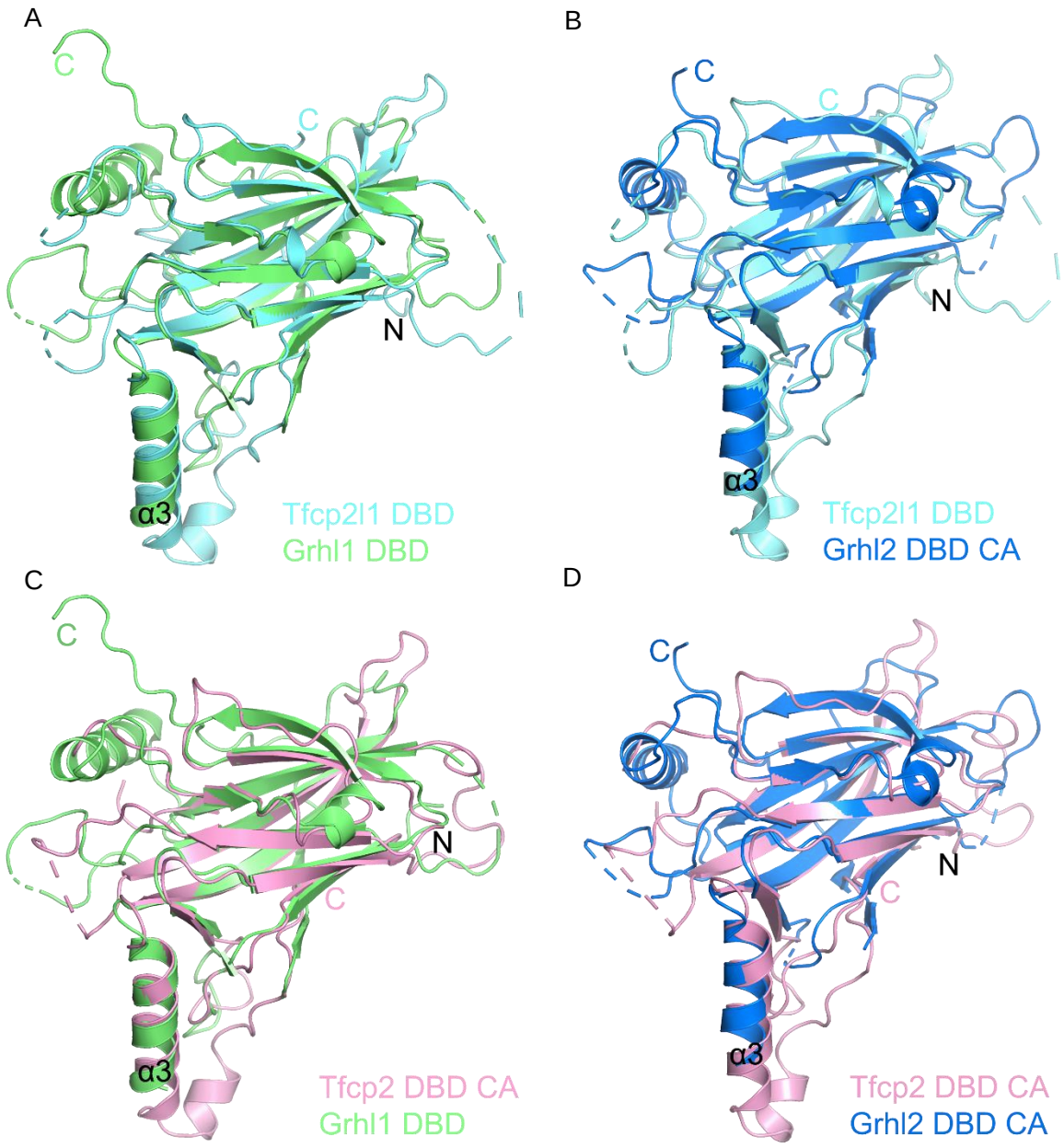

Figure 4-2. DBD structure alignment within the Grh/CP2 family. A, structure alignment of Tfcp2l1 DBD with Grhl1 DBD. B, structure alignment of Tfcp2l1 DBD with Grhl2 DBD chain A. C, structure alignment of Tfcp2 DBD chain A with Grhl1 DBD. D, structure alignment of Tfcp2 DBD chain A with Grhl2 DBD chain A.

Structure matching of the Tfcp2l1 DBD and the Grhl2 DBD (PDB: 5MR7) yielded an RMSD value of 1.347 Å (142 Cα atoms aligned in Coot). Structurally, Tfcp2 DBD was almost identical to Tfcp2l1 DBD (Fig. 4-1), and structure alignment of Tfcp2 DBD with Grhl1/2 DBD resulted in RMSD values of 1.828 Å and 1.528 Å, respectively (Fig. 4-2 C-D). It is concluded that Tfcp2l1 DBD and Tfcp2 DBD share a conserved structure with the DNA-binding regions of the homologous transcription

factors Grhl1 and Grhl2 and one may predict that these proteins share a conserved mode of DNA target-site recognition, although the target-site sequences differ in detail.

4.2.3 Tfcp2l1 DBD and Tfcp2 DBD resemble TP53 family structures

Above, it has been confirmed that the Tfcp2l1 DBD structure is similar to the Grhl1 DBD structure. Earlier, it was shown that the Grhl1 DBD structure resembles the TP53 DBD core structure[23]. It is therefore reasonable to assume that the Tfcp2l1 DBD structure is also similar to the TP53 DBD core structure although sequence alignment shows only 9.7% sequence identity between the Tfcp2l1 DBD and the TP53 core DBD.

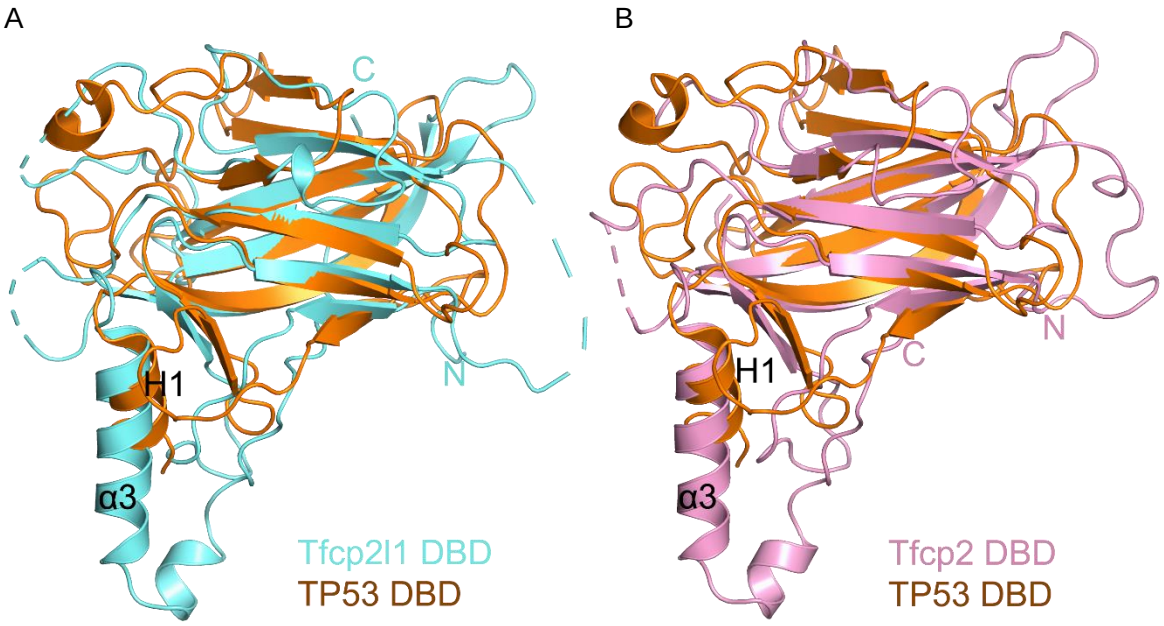

Figure 4-3. Structure alignment of Tfcp2l1 DBD and Tfcp2 DBD with TP53 DBD. A, structure alignment of Tfcp2l1 DBD with TP53 DBD. B, structure alignment of Tfcp2 DBD chain A with TP53 DBD chain A. Tfcp2l1 DBD colored in aquamarine, Tfcp2 DBD pink and TP53 DBD orange.

Superimposing the mouse Tfcp2l1 DBD structure on the TP53 DBD core structure (PDB entry: 1HU8, A) yielded a RMSD value of 2.35 Å for 129 aligned Cα atoms (Fig. 4-3 A), and human Tfcp2 DBD superimposed onto TP53 DBD (chain A) yielded a RMSD value of 2.54 Å for 123 aligned Cα atoms (Fig. 4-3 B). These results demonstrate that Tfcp2l1 DBD and Tfcp2 DBD have significant similarity with the TP53 DBD core structure. In the TP53 DBD core structure, eleven β-strands form the Ig-like domain, which is decorated by two helices, α1 and α2. Helix α1 of the TP53 DBD core structure, is missing in Tfcp2l1 DBD, probably due to disorder. Tfcp2l1 DBD helix α3 corresponds to the TP53 DBD helix α2. There are two more helices in the Tfcp2l1 DBD structure, α2 following α1 and α4 following α3. Tfcp2l1 DBD lacks three β strands present in TP53 DBD, the

first following β1, the second following α1, and the third following helix α2. The absence of these secondary structure elements may loosen the Ig-like core domain to some extent. The extra C-terminal loop wrapping around the Ig-like core domain is thought to stabilize the conformation.

Grhl1 DBD and the TP53 DBD core share 12.7% sequence identity (Fig. 4-4). Least-squares superposition of the two crystal structures yielded a RMSD of 2.67 Å for 132 aligned Cα atoms. Except for the C-terminal region containing two β strands decorating the Ig-like domain, Grhl1 DBD has 11 β strands[23]. Helices α1 and α3 of Grhl1 DBD correspond to TP53 DBD helices α1 and α2, while helix α2 of Tfcp2l1 DBD has no equivalent in TP53. The comparison therefore demonstrates that the Grhl1 DBD structure resembles the TP53 DBD structure in a similar way as Tfcp2l1 DBD resembles TP53 DBD, suggesting that the Grh/CP2 DBDs share a common molecular ancestor with TP53 DBD.

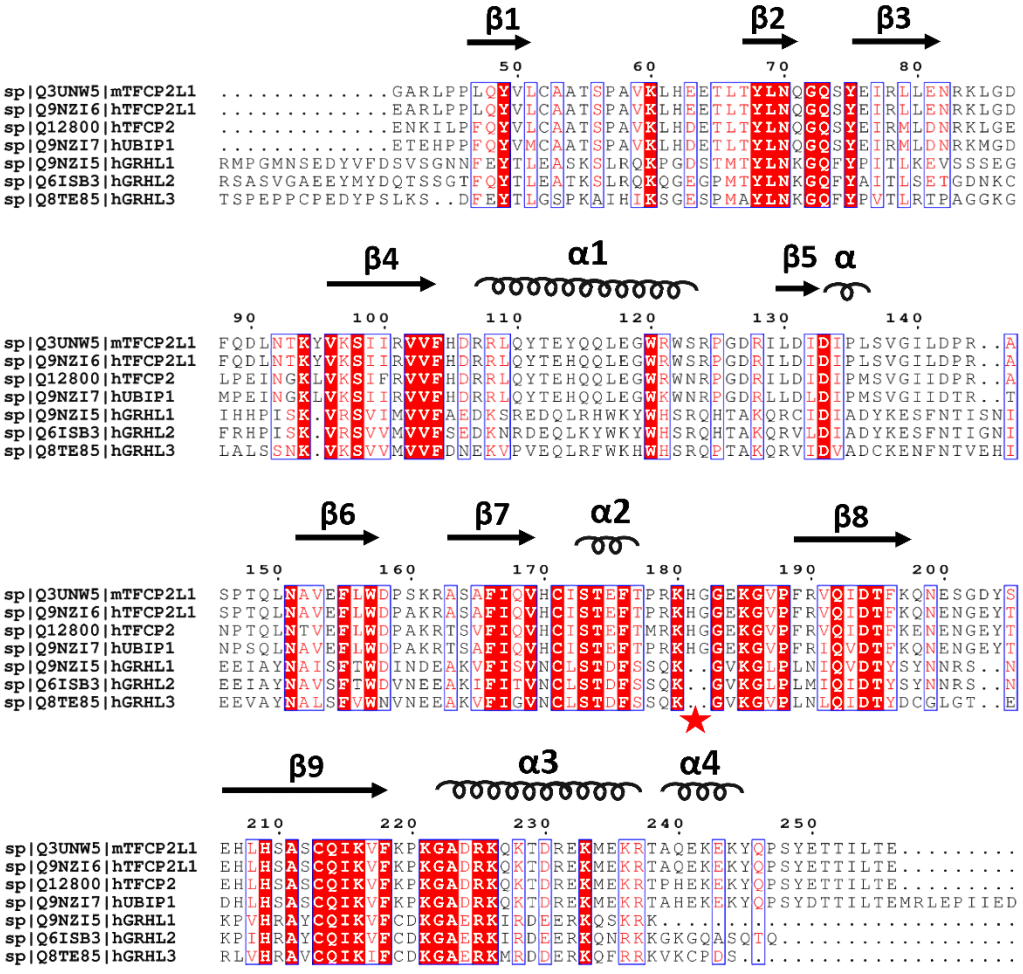

Figure 4-4. DBD sequences are conserved in the Grh/CP2 TF family. Residue counting is according to the Tfcp2l1 sequence. Perfectly conserved residues are shown on a red background, and less strictly conserved residues are boxed. The secondary structures of Tfcp2l1 and Tfcp2 DBD are shown above the sequences. The red asterisk indicates a sequence difference between Grhl and CP2 subfamily members in the loop L10 region where CP2 subfamily members contain two extra residues: H181 and G182.

4.3 The Tfcp2l1 DBD:DNA complex is similar to the Grhl1 DBD:DNA complex

4.3.1 Tfcp2l1 DBD binds to ds12bpDNA with a geometry resembling the Grhl1 DBD:DNA complex

As previously reported, Grhl1 DBD bound ds12bpDNA (AAAACCGGTTTT) in a slightly asymmetric pattern[23]. Structural alignment of the Tfcp2l1 DBD:DNA complex with the Grhl1 DBD:DNA complex (PDB entry: 5MPF) yielded a RMSD value of 2.32 Å for 2906 aligned atoms, suggesting that two complexes are structurally similar (Fig. 4-5 A). Structure alignment of the individual DBDs (chains A) resulted in a RMSD value of 0.95 Å for 1179 aligned atoms (Fig. 4-5 B). Chain B alignment from two DBDs yielded RMSD = 0.852 Å for 1204 aligned atoms (Fig. 4-5 C).

The specific contacts mediated by R225 and G183 of the Tfcp2l1 DBD:DNA complex are mediated by R427 and G387 in the Grhl1 DBD:DNA complex. Unspecific contacts involve residues T174, G183, K219, R225 and K226 of Tfcp2l1 DBD and the equivalent residues T380, G387, C421, R427 and K428 of Grhl1 DBD. In the Grhl1 DBD:DNA complex, the DNA contacting residues of the two protein chains are the same. However, in the Tfcp2l1 DBD:DNA complex, the two Tfcp2l1 DBDs do not bind the DNA with perfect symmetry, although the target-site DNA sequence is self-complementary (both strands have the same sequence). Only chain A K185 of Tfcp2l1 DBD, matching Grhl DBD K389, interacts with the G7 phosphate group but not K185 from chain B. Furthermore, the Tfcp2l1 DBD chain A residue R179 interacts with the T9 phosphate group, whereas the Tfcp2l1 DBD chain B residue K216 interacts with the G7 phosphate group, a contact only observed in the Tfcp2l1 DBD:DNA complex. Conversely, residues R430 and R434, interacting with the DNA in the Grhl1 DBD:DNA complex, have no functional equivalents in the Tfcp2l1 DBD:DNA complex.

Structure alignment of the two DNA double strands from two the complexes yielded an RMSD of 1.21 Å for 466 aligned atoms (Fig.4-5 D-E). Hence, the two DNA structures differ significantly in their conformation when bound to either Grhl1 DBD or Tfcp2l1 DBD. The central base pairs from C5 to G8 of Tfcp2l1 DBD-bound DNA superimpose well with Grhl1 DBD-bound DNA, but the A-T pairs flanking the CCGG upstream and downstream do not align well in the two DNA structures (Fig. 4-5 D and E). Compared to the 12-mer B-form DNA generated with the 3DNA program[182], the structures differ with an RMSD of 1.59 Å for 438 aligned atoms (Fig. 4-5 F). It is obvious that the major groove of Tfcp2l1 DBD-bound DNA is significantly wider than the major groove in Grhl1 DBD-bound DNA and B-form DNA (Table 4-1). Interestingly, the loop L10 inserts into DNA minor

groove which widened the DNA minor groove compared to B-form DNA (Fig. 3-24 A and B, Table 4-1). The widened minor and major groove of Tfcp2l1-bound DNA supplied the DNA structure in the center where the TF DBDs binds.

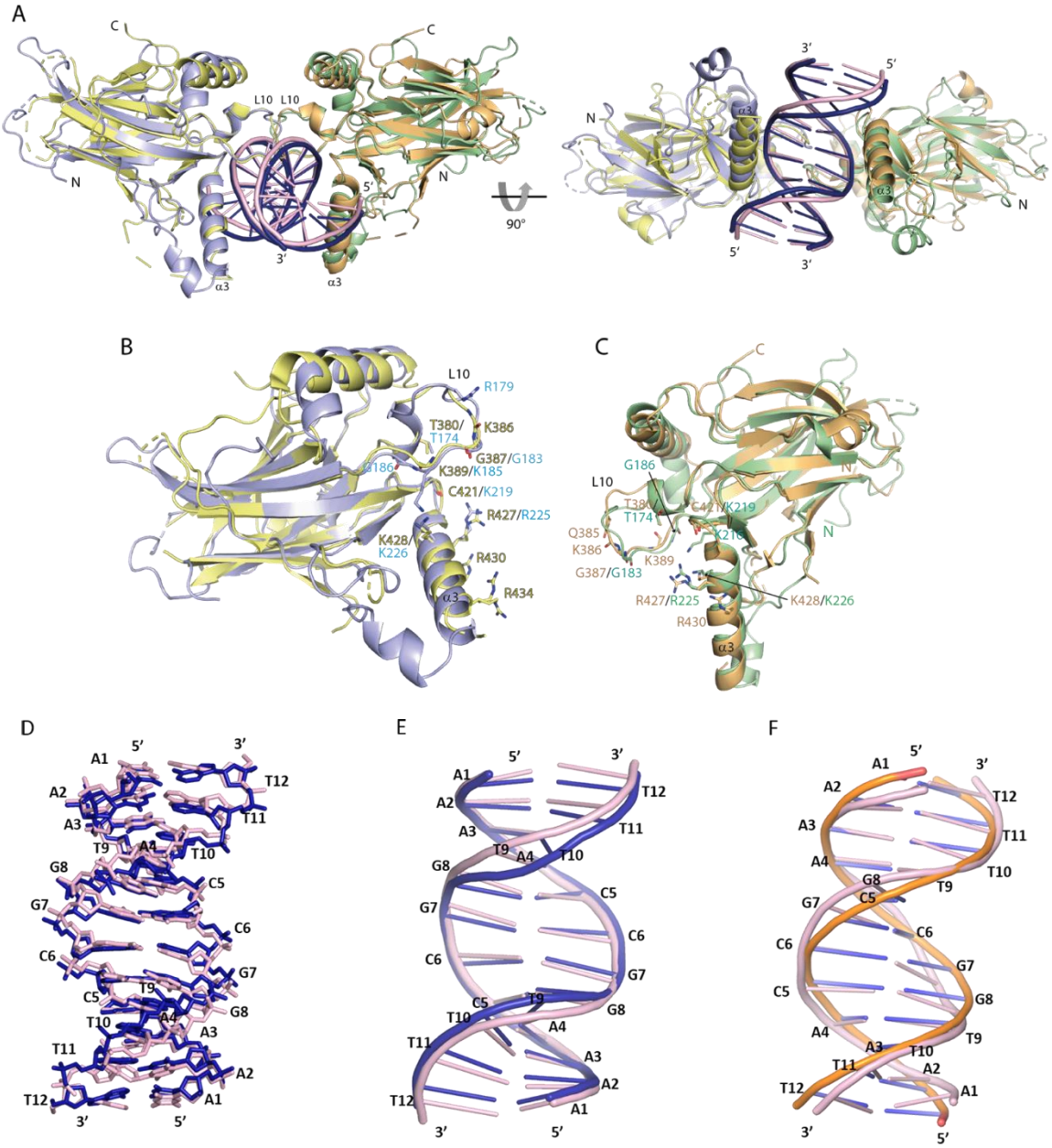

Figure 4-5. Superimposition of the Tfcp2l1 DBD:DNA complex onto the Grhl1 DBD:DNA complex. A, Orthogonal cartoon views of the superimposed Tfcp2l1 DBD:DNA and Grhl1 DBD:DNA complexes (RMSD = 2.32 Å for 2906 aligned atoms). B and C, Structural alignment of the individual DBDs with superimposed chains A and B, respectively. The residues contacting DNA are labeled and shown in stick representation. The Tfcp2l1 DBD:DNA complex is shown in light blue (DBD, chain A), lime (DBD, chain B) and pink (DNA). Corresponding features of the Grhl1 DBD:DNA complex are colored yellow, orange and dark blue. D-E, Structure alignment of DNA duplexes from the two complexes in stick (D) and cartoon (E) style. F, Structure alignment of the Tfcp2l1 DBD-bound DNA to standard B-form DNA (orange). The alignment was done with PYMOL, which performs a sequence alignment followed by a structural superposition.

Table 4-1. Minor and major groove widths: direct P-P distances

	B form ds12bp DNA		Tfcp2l1 bound ds12bp DNA		Grhl1 bound ds12bp DNA	
	Minor groove P-P (Å)	Major groove P-P (Å)	Minor groove P-P (Å)	Major groove P-P (Å)	Minor groove P-P (Å)	Major groove P-P (Å)
3 AA/TT	12.4	16.6	**9.7**	**16.4**	9.6	17.0
4 AC/GT	12.5	16.3	**10.6**	**17.5**	11.0	18.1
5 CC/GG	12.5	16.5	**12.9**	**19.1**	14.1	17.9
6 CC/GG	12.4	16.6	**14.6**	**19.2**	15.6	17.2
7 GG/CC	12.5	16.5	**13.2**	**19.7**	13.9	17.3
8 GT/AC	12.5	16.3	**10.1**	**19.2**	10.8	18.4
9 TT/AA	12.4	16.6	**8.5**	**17.3**	9.3	17.1

Based on the structure alignment with the Grhl1 DBD:DNA complex, Tfcp2l1 DBD residue K219 was matched to Grhl1 C421. The peptide NH groups of both C421 and K219 contact the phosphate of guanine G8. Therefore, there is no difference in affinity between the C421A mutant and the wild-type of Grhl1 DBD. However, in the Tfcp2l1 DBD:DNA complex, the main chain of K219 and the aromatic ring of F218 form interactions with the DNA backbone which facilitate Tfcp2l1 DBD binding to DNA. It explains why the Tfcp2l1 DBD K219A mutation increases the KD for target DNA binding threefold compared to the wild-type protein.

4.3.2 Tfcp2l1 DBD dimer formation supported by loop L10

The loop L10 region of Tfcp2l1 inserts into the DNA minor groove with residues K180, H181, G182 and G183. The distance between the residues K180 of two L10 loops of the pseudo-symmetrically bound Tfcp2l1 DBD units is 4.68 Å. There is no polar interaction between two loops. Nevertheless, there seems to be a particular hydrophobic interaction formed by K180 and H181 side chains and the protein backbones at G182 and G183. As described, H181A and G183A mutations influence the protein:DNA binding affinity (Fig. 3-26). This confirms that loop L10 is also involved in protein:DNA contacts.

In the Grhl1 DBD:DNA complex, residue K386 of one chain contacts S383 and T380 of the other via hydrogen bonds to stabilize the two Grhl1 DBD molecules forming the DNA-bound protein

dimer[23]. In the Tfcp2l1 DBD:DNA complex, the side-chain orientation of K180 is different from K386 of the Grhl DBD:DNA complex. There is no direct interaction between K180 and residues on the opposite chain. Furthermore, two more residues are located in the Tfcp2l1 L10 region, H181 and G182, which do not exist in Grhl1. These two residues, together with K180 and G183 extend the loop to a smooth conformation, allowing the L10 loops from individual Tfcp2l1 DBD molecules to form a hydrophobic area.

4.3.3 Protein:DNA interfaces are conserved in Grh/CP2 family

Structure alignment of Tfcp2l1 DBD with the domain from the Tfcp2l1 DBD:DNA complex yielded a RMSD value of 0.465 Å, demonstrating that the DBD structures are closely similar in the apo and DNA-bound states. Bound to DNA, the Tfcp2l1 DBD structure displays two helices, α1 and α2, and one additional β strand (preceding loop L10), which are not observed in the Tfcp2l1 DBD apo structure. Binding to the DNA appears to stabilize floppy loop regions in a fixed conformation.

As described, the DBD of the Grh/CP2 TFs is highly conserved. In both the Grhl1 DBD:DNA and Tfcp2l1 DBD:DNA complex structure, the DBD is interacting with duplex DNA via an interface involving helix α3 and loop L10. On the sequence level, helix α3 is most conserved, while the loop regions distant from the DNA are least conserved (Fig. 4-6). After all-atom superposition of Tfcp2l1 DBD helix α3 onto Grhl1 DBD helix α3 yielded a RMSD value of 0.425 Å. Likewise, a superposition onto Grhl2 DBD helix α3 yielded a RMSD of 0.361 Å, and superposition of Tfcp2 DBD helix α3 onto the Grhl1/2 DBDs helix α3 yielded RMSD values of 0.428 Å and 0.584 Å, respectively. It is concluded that the orientation and side chain conformations of helix α3 are conserved among the DBDs of Tfcp2l1, Tfcp2, Grhl1 and Grhl2, keeping this helix poised for contacts with the target DNA major groove.

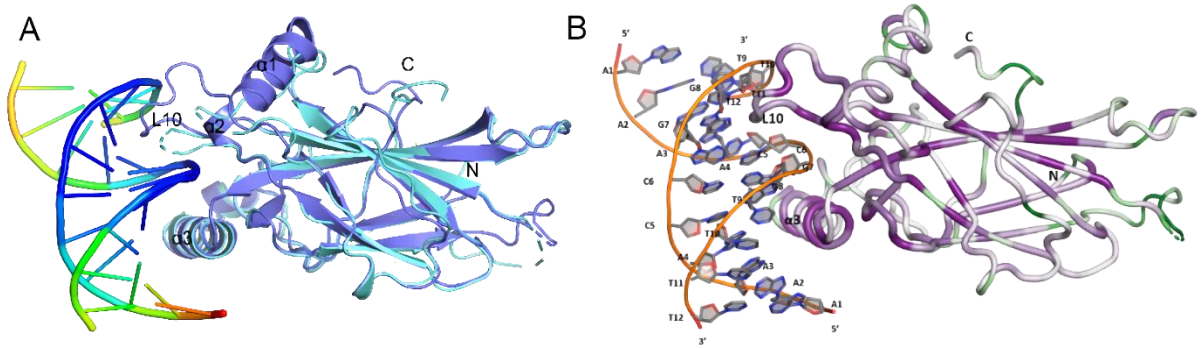

Figure 4-6. Protein:DNA interfaces are conserved in Grh/CP2 family members. A, Structure alignment Tfcp2l1 DBD (cyan) with Tfcp2l1 DBD:DNA (blue) yielding a RMSD value of 0.465 Å for 153 aligned Cα atoms. The DNA is colored based on the B-factor value (red, highest; blue, lowest). B, Crystal structure of the Tfcp2l1 DBD:DNA complex with the protein tube representation and colored according to sequence conservation with the Grh/CP2 TF family. High sequence conservation: wide tube, purple; low conservation: narrow tube, green.

The loop L10 region differs between the Grh and CP2 subgroups, and the sequences are highly conserved within a subgroup only. In the Grh subfamily, loop L10 contains the tetrapeptide sequence QKGV; in the CP2 subfamily, loop L10 contains the hexapeptide sequence RKHGGE. The longer loop L10 conserved in all CP2 subfamily members harbors residue G183, which interacts with G8 of a 12-bp target DNA motif and stabilizes protein binding to the DNA. In CP2 subfamily members, loop L10 is part of a hydrophobic area that stabilizes the DNA-bound protein dimer and contributes to the protein:DNA interface.

4.3.4 The Tfcp2l1 DBD:DNA interface resembles the TP53:DNA interface

As discussed before, the DBD structure of members of the Grh/CP2 TF family is similar to the TP53 DBD structure. However, in spite of the architectural similarity there are some differences regarding the protein:DNA interaction between Tfcp2l1 and Grhl1. These differences are the basis of recognizing different target DNA sequences. Structure alignment between the Tfcp2l1 DBD:DNA complex and the hTP53:DNA complex (PDB entry: 3TS8) yielded an RMSD value of 3.973 Å for 261 aligned α-carbon atoms. Sequence alignment between Tfcp2l1 DBD and TP53 DBD yields a very low identity score of 8.8%. hTP53 binds as a tetramer to its 26-bp target DNA (Fig. 4-7 A). At each end of the DNA double strand, two TP53 molecules bind in a dimeric arrangement stabilized by their C-terminal domains. In contrast to Grh/CP2 family members, the extended C-terminal loop region of the TP53 DBD is returning to the Ig-like domain of another subunit in the DNA bound tetramer to support the structure of the TP53-DNA complex (Fig. 4-7 A and D).

In the TP53-DNA complex, residues R280 and R248 from helix α2 contact the guanosine nucleotide G2, residues A276 and C277 interact with the adenine base T3, and residues K120 and S121 located in loop L1 interact with A4 and C5, respectively[183]. These specific interactions are the basis for target DNA recognition by TP53. In the Tfcp2l DBD:DNA complex, the specific interactions between protein and DNA differ from the TP53-DNA complex. Loop L1 following strand β1 of Tfcp2l1, corresponding to TP53 loop L1, has a different orientation, and there is no interaction between Tfcp2l1 loop L1 with DNA (Fig. 4-7 B and C). The conserved helix α3 anchors the protein to DNA via R225 by contacting guanine base G8, which assumes the primary function in DNA sequence readout by Tfcp2l1. Loop L10 inserts into the DNA minor groove to support protein binding to DNA by pulling the DNA towards the protein.

In the TP53-DNA complex, the helices α2 from two DNA-bound TP53 molecules enclose an angle of 98.2°, whereas in the Tfcp2l1 DBD:DNA complex, the angle between the two α3 helices is 65.7º (Fig. 4-7 B and C). Following the structural superimposition, helix α2 from TP53 and helix α3 from

Tfcp2l1 enclose an angle of 35.5°, demonstrating their different orientation relative to the Ig-like DBD core domain. Tfcp2l1 DBD has a similar tertiary structure as the TP53 DBD core domain, and the Tfcp2l1 DBD:DNA interface is similar to that formed by TP53 and its target DNA. The variation of specific DNA interactions between these two complexes allows each protein to recognize different DNA sequences and to control the transcription of different target genes.

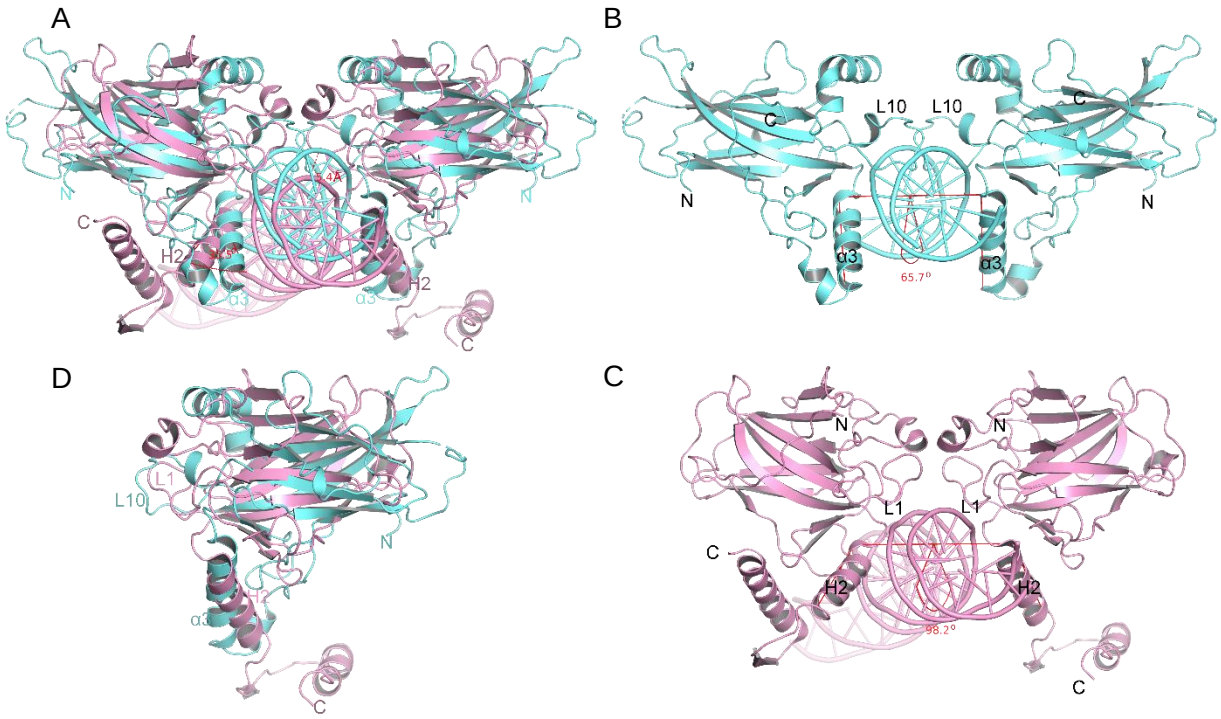

Figure 4-7. Structure alignment of the Tfcp2l1 DBD:DNA complex and the TP53:DNA complex (PDB entry: 3TS8). A, Superimposition of the Tfcp2l1 DBD:DNA complex onto the TP53:DNA complex (hiding chains C and D) yielded a RMSD 3.973 Å from 261 aligned residues based on the SSL algorithm in Coot[167]. The phosphate group distance of two DNA molecules is 5.4 Å, and helix α3 of Tfcp2l1 and helix α2 of TP53 enclose an angle 35.5°. B, Two helices α3 of Tfcp2l1:DNA complex has a dihedral angle of 65.7°, C, The two helices H2 of TP53:DNA complex enclose an angle of 98.2°. D, Least-squares superposition of chains A from the two aligned complexes. The Tfcp2l1 DBD:DNA complex is colored in aquamarine, the TP53:DNA complex in pink.

4.4 CP2 subfamily factors: DNA-binding motif

4.4.1 Tfcp2l1 binds to the ds14bpDNA sequence

It was previously reported that Tfcp2l1 binds to the 14-mer DNA motif $CC^{A}/_{G}GN_{6}CC^{A}/_{G}G$ (N, any nucleotide)[45]. This motif, here denoted ds14bpDNA, contains two copies of the $CC^{A}/_{G}G$ sequence separated by six base pairs. Assuming standard B-form DNA geometry, 10 base pairs complete one turn of the double helix, implying that Tfcp2l1 may bind to the $CC^{A}/_{G}G$ motifs at both ends of

the ds14bpDNA duplex on the same face of the helix. Another possibility is that Tfcp2l1 binding to ds14bpDNA may require an extended DNA duplex with additional 5' and 3'-terminal base pairs flanking the $CC^{A}/_{G}G$ core motif. Whether one or two Tfcp2l1 dimers bind to ds14bpDNA and if two DNA-bound dimers adopt a tetrameric arrangement is not clearly established.

In vitro assays found that Tfcp2l1 weakly binds to the ds14bpDNA sequence and that binding affinity is increased by adding flanking base pairs reaching a maximum at ds20bpDNA (Fig. 3-12 and 3-13). In model one, a 20-bp DNA sequence containing two copies of the 10-bp $N_3CC^{A}/_{G}GN_3$ sequence attracts two or four Tfcp2l1 DBDs. It is confirmed from RALS that ds20bpDNA binds four Tfcp2l1 DBD monomers, which corresponds to two Grhl1 DBD dimers binding to duplex DNA. As described above, Tfcp2l1 Δ19 forms tetramers in solution, suggesting that a Tfcp2l1 tetramer is the protein species that binds to ds20bpDNA or ds14bp DNA in 1:1 stoichiometry.

The 20-bp DNA sequence was shortened to 12bp DNA to further study the protein:DNA interaction patterns. From the ChIP-seq reports[49], two outer base pairs of the $CC^{A}/_{G}G$ motif are more conserved than the central bases. In agreement with these observations, the *in-vitro* ITC assays confirmed that C5 and G8 mutations of the ds12bpDNA abolished the protein DNA interaction (Fig. 3-28).

4.4.2. Stoichiometry of protein to DNA

ITC measurements were used to determine the stoichiometry of transcription factor to DNA. If the event of two binding events with different affinities being reflected in one titration, the MicroCal-PEAQ-ITC program evaluates the binding isotherm using a two-set binding model that returns dissociation constants and other thermodynamic parameters for each individual binding event. The isotherms from Tfcp2l1 DBD binding to DNA variants were fitted using the one-set binding model or the two-set binding model as determined by the program. The crystal structure of the Tfcp2l1 DBD:DNA complex clearly demonstrated that two Tfcp2l1 DBDs were bound to ds12bpDNA, even though the stoichiometry as determined by ITC is 1.48:1. In ITC titrations of Tfcp2l1 DBD with ds20bpDNA and ds22bpDNA, the stoichiometry values are calculated as 3.03 and 3.17, respectively. Both ds20bpDNA and ds22bpDNA accommodate two copies of the consensus core motif $CC^{A}/_{G}G$. Assuming a binding geometry and specificity as observed in the Tfcp2l1 DBD:ds12bpDNA complex, it seems possible that four Tfcp2l1 DBD molecules could be fitted onto the ds20bpDNA duplex, yielding a stoichiometry value of 4. RALS results confirmed that Tfcp2l1 DBD binds ds20bpDNA and ds22bpDNA with the stoichiometry of four protein molecules to one DNA duplex (Fig. 3-14).

Previous results showed that the 20-mer DNA motif is needed for Tfcp2l1 DBD binding to DNA (Fig. 3-13 and 3-14). From 14-bp via 16-bp, 18-bp to 19bp DNA molecules, each carrying two copies of the core binding motif, the binding affinity of Tfcp2l1 DBD is increasing. However, the binding pattern is complicated among these DNA sequences. For example, Tfcp2l1 DBD weakly binds to ds14bpDNA (5'-C_1CAGTTCAAAC$_{11}$CAG-3'), and the ITC data require fitting with the two-set binding model. Tfcp2l1 DBD binding to C_1CAGTTC is different from AAAC$_{11}$CAG binding because of the different flanking sequences. C_1CAGTTC has three nucleotides TTC at 3' of C_1CAG and lacks of nucleotides at 5', AAAC$_{11}$CAG has three nucleotides at 5' of C_{11}CAG and lacks of nucleotides at 3'.

Throughout all ITC assays with Tfcp2l1, the stoichiometry is calculated as 1.5: 1 to the 12-mer DNA and 3.0: 1 to the 20-mer DNA. However, the analytical gel filtration chromatography showed that the DNA is completely double-stranded leaving no single-strand after DNA annealing and demonstrating that the non-integral of protein dimer to DNA double-strand stoichiometry cannot be attributed to the presence of single-stranded DNA. It remains unclear why ITC yields stoichiometries that differ from 2 or 4.

4.4.3. The spacer region of Tfcp2l1 binding sites is not restricted to six base pairs

Previous ChIP-seq experiments revealed a conserved Tfcp2l1-binding DNA motif CC$^A/_G$GN$_6$CC$^A/_G$G with six random middle nucleotides separating the two consensus core motifs CC$^A/_G$G[49]. Given the helical repeat of 10.4 to 10.5 bp per turn in B-form DNA, the second 4-bp core motif following the first after 6 spacer base pairs will be placed approximately one turn away from the first on the same side of the double helix. Indeed, Tfcp2l1 Δ19 tetramers bind to ds20bpDNA with a K_D of 24.6 nM and a stoichiometry of 0.845:1. However, this exact type of Tfcp2l1 binding DNA sequence is not observed exclusively. It is well known that Tfcp2l1 binds to the *Klf4* promoter region to control the *Klf4* gene transcription. There is only one specific site fitting this strict pattern present in the -50 kb upstream region. Therefore, one possibility is that Tfcp2l1 binding to this single site is sufficient for the regulation of *Klf4* transcription by Tfcp2l1.

However, except for this strict pattern, there are multiple single variants of the 4-bp core motif, including the sequences CCAG, CCGG and CTAG. The spacing of these CCAG, CCGG and CTAG sequences is variable, ranging from 1 bp to 9 bp. Spacer lengths of 10 bp or more were not considered. It has been often observed that TFs can recognize multiple binding sites within a promoter region to achieve gene regulation[184]. It is possible that Tfcp2l1 is a monomer or forms dimers in solution which could bind to these specific DNA sequences. Interestingly, *in vitro* assays also showed that Tfcp2l1 Δ19 tetramers could bind to DNA target sequences with spacer lengths

N4, N5 and N7 (Fig. 3-30). The binding affinities are 23.5 nM, 21.0 nM and 29.1 nM, and the stoichiometries are 0.617:1, 0.87:1, and 1.06:1, respectively. This indicates that Tfcp2l1 tetramers may bind to variants of the target sequence with N5 and N7 spacers in a similar mode as to ds20bpDNA (N6). The N4 sequence variant is different due to there is not enough space for Tfcp2l1 tetramer binding. It is concluded that the spacer length of Tfcp2l1-binding DNA sequences is not strictly six base pairs. Given a degree of variability of the spacer length, the probability of two binding events to four-nucleotide core motifs leading to productive binding to a tandem sequence is significantly increased, allowing for synergistic control of transcription regulation.

Because the structure of full-length Tfcp2l1 is not known, it is difficult to assess which influence the SAM domain will have on Tfcp2l1 tetramer geometry. The protein sequence indicates the presence of an extended loop region between the DBD and the SAM domain, which is likely to provide the DBDs with some freedom in their arrangement on DNA target sites as parts of Tfcp2l1 oligomers. Therefore, Tfcp2l1 binding to DNA target sites with variable spacer lengths was modeled. The structures of DNA motifs (N4, N5, N6 and N7) were generated by 3DNA program[182]. Shorter (N4 or N5) or longer (N7) spacers in standard B-form DNA compared to the ds20bpDNA reference (N6) have two different effects on the relative position and orientation of two Tfcp2l1 dimers simultaneously bound at the two $CC^{A}/_{G}G$ core motifs: Each base pair added to or removed from the spacer region will move the Tfcp2l1 dimers away from each other or together by 3.4 Å along the helix axis. In addition, each base pair added to or removed from the spacer will change the rotational setting between the DNA-bound dimers by ~36°, the average Twist of B-DNA. The loop regions between DBD and SAM is expected to accommodate some of the necessary structural adaptations in Tfcp2l1 binding to variants of the consensus site (Fig. 4-8).

For the N7 spacer with one additional base pair relative to ds20bpDNA, there is a 44.7º dihedral angle rotation between DNA-bound Tfcp2l1 dimers. There is enough space along the DNA double strand, and the loop regions provide the required flexibility for binding. For the N5 and N4 spacers, there is a 30.8º and 59.8º dihedral angle rotation, respectively. The Tfcp2l1 dimers show serious clashes when bound to the N4 DNA due to the short spacer DNA. It was confirmed by ITC that Tfcp2l1 Δ19 tetramers bind to N4 DNA with a stoichiometry of 0.617, implying that little more than half the Tfcp2l1 Δ19 tetramers are bound to the N4 DNA sequence. Unexpectedly, the dihedral angle rotation between DNA-bound Tfcp2l1 dimers is not a well-distribution value of 36º based on one nucleotide extended or shortened in spacer of the DNA motif. Further *in vitro* and *in vivo* assays are needed to confirm these results.

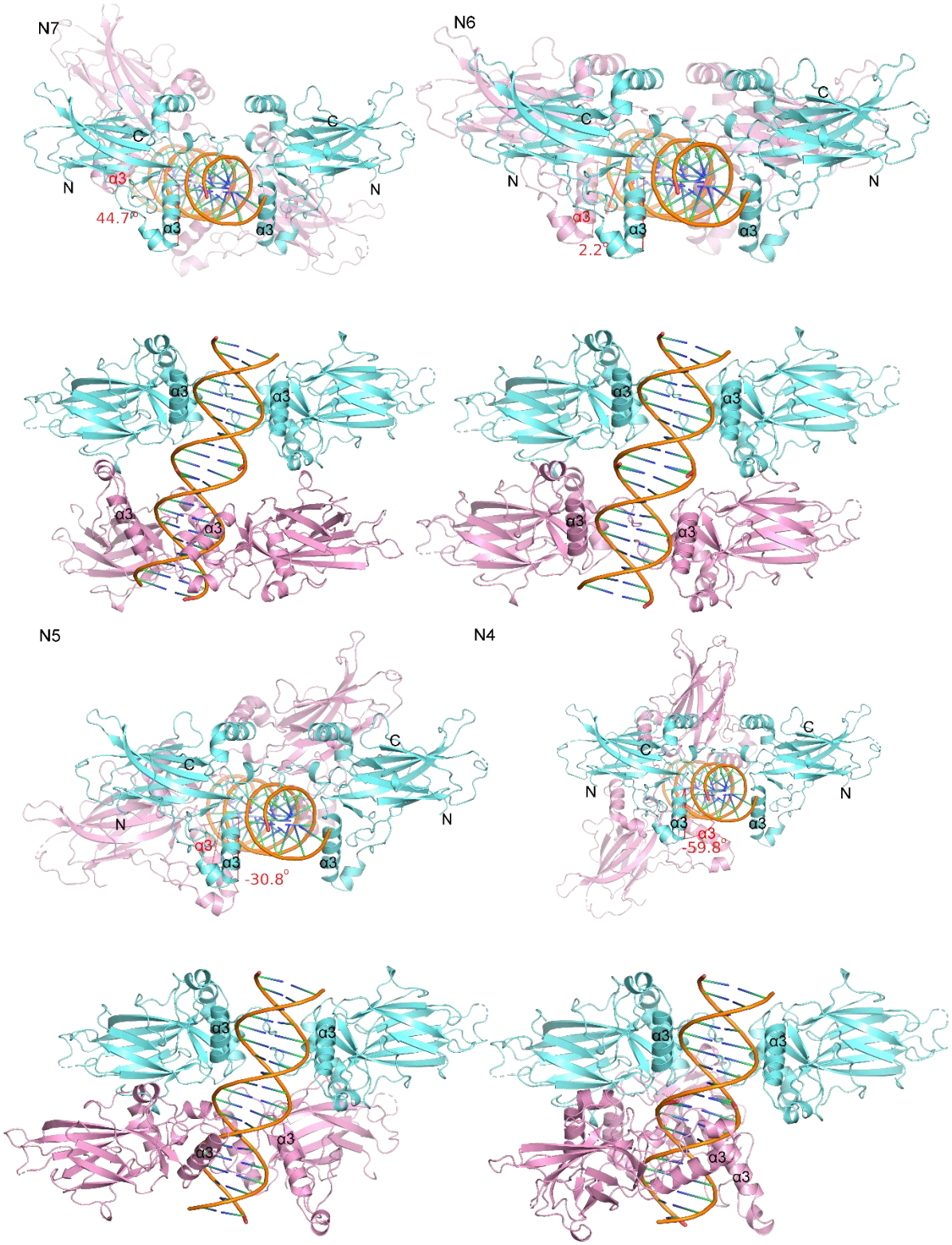

Figure 4-8. Structural model of Tfcp2l1 DBD binding to DNA variants with different spacer lengths. The DNA structures were built with the 3DNA program[182]. The Tfcp2l1 DBD dimer (colored in aquamarine) binds to the first core recognition site and a second DBD dimer (pink) binds to the second core recognition site. The dihedral angle was calculated based on two helices α3 on the same side of DNA duplex. From N7 to N4 DNA, the dihedral angles were 44.7º, 2.2º, -30.8º, and -59.8º, respectively.

4.5 Residue modifications in CP2 family members

4.5.1 Functions related to Tfcp2l1 residues

The gnomAD database lists 234 single missense mutations in human Tfcp2l1 (https://gnomad.broadinstitute.org/gene/ENSG00000115112?dataset=gnomad_r2_1). 105 of these 234 mutations are located in the Tfcp2l1 DBD region, and 125 mutations are distributed over the C-terminal region. No report about a single missense mutation relates to clinical disease. Regarding the Tfcp2l1 DBD:DNA complex, the residues which directly contact the DNA are not listed among the 234 mutations.

A recent study from Hildebrandt's laboratory showed that four mutations, P46A, Q168R, V263M and S290F, in hTfcp2l1 DBD are linked to distal tubulopathy in human[185]. The Tfcp2l1 DBD:DNA crystal structure showed that P46 is located at the N-terminal loop of the DBD, which does not stabilize the protein structure. Residue V263 is located within the extended C-terminal loop that wraps the Ig-like domain of the Tfcp2l1 DBD structure. Both P46 and V263 are distant from the DNA binding surface of Tfcp2l1 DBD. The mutation Q168R, located in the Ig-like domain, is also not in proximity to the DNA binding surface. The mutation S290F is located outside of the Tfcp2l1 DBD structure and far away from the DNA binding region. It is therefore unlikely that DNA binding of Tfcp2l1 DBD is influenced by any of these four mutations. ITC measurements of single Tfcp2l1 DBD mutant binding to ds12bpDNA did not provide any indication for a significant difference in the binding affinities between mutated and wild-type Tfcp2l1 DBD.

It has been reported that Tfcp2l1 T177 phosphorylation plays a vital role in embryonic stem-cell pluripotency and differentiation; in the absence of T177 phosphorylation the tumorigenic potency of bladder cancer cells is impaired[145]. T177 locates at loop L10 which is inserted into the DNA minor groove. In the crystal structure of the Tfcp2l1 DBD:ds12bpDNA complex, there is no direct interaction between T177 and DNA. While a T177A mutation dramatically decreases target gene expression, the T177E mutation does not influence transcription compared to T177 phosphorylation, which suggested that the negative charge introduced via Tfcp2l1 T177 phosphorylation or T177E mutation upregulates target gene expression and promotes cell-cycle progression.

Tfcp2l1 was described as a transcriptional repressor. However, K30-sumoylated Tfcp2l1 could activate target gene expression[43]. In the Tfcp2l1-DBD crystal structure, K30 locates in a polypeptide region, which does not show electronic density, presumably due to disorder. It is concluded that the K30 modification does not affect the protein:DNA contacts. The Tfcp2l1-K30

sumoylation may therefore serve as another example of a residue whose post-translational modification plays an essential role in gene expression regulation.

4.5.2 Functions related to Tfcp2 residues

The gnomAD database lists 61 single-site missense mutations in the Tfcp2 DBD region among a total of 174 mutations in Tfcp2. None of these missense mutations has been reported to link to clinical disease. It has been reported that both the Q234L or K236E mutation result in a significant reduction of protein:DNA interaction[51]. The structure of the Tfcp2 DBD is closely similar to the Tfcp2l1 DBD structure. Structure alignment of the Tfcp2 and Tfcp2l1 DBDs showed that Tfcp2 K236 is conserved and corresponds to Tfcp2l1 K216. Tfcp2l1 K216 is in contact with DNA, and the K216A mutation has been confirmed to decrease the DNA-binding activity. Therefore, it may be assumed that the Tfcp2 K236E mutation has a similar effect as a Tfcp2l1 K216A mutation. Residue Q234 of Tfcp2 is also conserved in the Grh/CP2 family, but further away from the DNA interface. The Q234L mutation reduces the DNA-binding activity of Tfcp2, but does not disturb the direct protein:DNA interaction. In the Tfcp2 DBD structure, Q234 forms polar interactions with T67 and three water molecules. The Q234L mutation will render the region around this side chain less hydrophilic which may influence the stability of the whole Ig-like domain. Further evidence is needed to support this hypothesis.

Recently, a study showed that Tfcp2 directly binds to the epidermal growth factor (EGF) and transforming growth factor-alpha (TGF-α) gene promoter regions to control their expression, which activates EGF receptor (EGFR) activity[186]. However, active EGFR reduces deficient long-term survival in breast cancer patients[187]. The Tfcp2 D153A mutation has been reported to impair Tfcp2 binding to the EGF or TGF-α promoters, which could down-regulate the EGFR activity. In the Tfcp2 DBD structure, D153 locates to an Ig-like domain loop region far away from the DNA interface, suggesting that residue D153 may have functions other than DNA binding.

Previous reports demonstrated that Tfcp2 phosphorylation by ERK and CDK2 is involved in G1 cell-cycle regulation[188]. Tfcp2 S291 phosphorylation by ERK and S309 phosphory-lation by CDK2 could inhibit the Tfcp2 transcriptional activation activity, S291 and S309 dephosphorylation could reactivate Tfcp2 activity, which is essential for cell cycle from quiescence to early G1 phase[188]. Further studies indicated that two serine-proline/ threonine-proline motifs in Tfcp2 (at residues S291 and T329) are required for association with the prolyl isomerase Pin1 to dephosphorylate Tfcp2 at these two SP/TP motifs to reactivate Tfcp2 activity[189]. All three residues are outside of the Tfcp2 DBD, and their modification did not affect the DNA-binding activity, suggesting that Tfcp2

post-translational modifications may play essential roles in cellular processes independent of transcription regulation.

Compared to Tfcp2l1, Tfcp2 contains two additional specific sequence motifs, a poly-proline sequence (AAs, 314-319) and a glutamine-rich sequence (AAs, 396-413). The poly-proline sequence might form a poly-proline helix, which is a special type of protein secondary structure. The function of the poly-proline sequence is uncertain. Q-rich sequences are sometimes called poly-glutamine tracts[190]. A polyQ tract was first found in the Notch receptor, further studies reported that the polyQ tract may result in neurodegenerative disease, such as the Huntington's disease[191]. There is no publication related to a Tfcp2 polyQ tract function so far.

4.5.3 Mutations of Ubp1 residues

Ubp1 has been reported to form heterodimers or heterooligomers together with Tfcp2 to regulate target gene expression. The gnomAD database lists 177 single-site missense mutations in Ubp1; none of them has been reported to cause clinical disease. In contrast to Tfcp2 and Tfcp2l1, there is no structural information on Ubp1, neither about the full-length protein, the DBD or the C-terminal domain. More work is needed to elucidate the UBP1 function and structure.

4.6 Tfcp2l1 binds *Esrrb* and *Klf4* gene promoter sequences

As described before, Tfcp2l1 binds to the *Esrrb* gene and *Klf4* gene to control their transcription. A search for possible Tfcp2l1 target sequences found the sequence TCG**CCAG**CCT-TGA**CTAG**TGC (from -1311 bp to -1299 bp) in the *Esrrb* promoter region and GCG**CCAG-**CGTTCG**CCGG**TGA (from -1080 bp to -1060 bp) in the *Klf4* promoter region. To avoid formation of a DNA hairpin secondary structure, k20bpDNA (ACT**CCAG**CGTTCG**CCGG**TGA) was designed by modifying the three 5'-terminal base pairs acids and synthesized for Tfcp2l1 binding studies.

ITC measurements confirmed that Tfcp2l1 binds to e20bpDNA and k20bpDNA following a two-set binding model for both. Tfcp2l1 binds to e20bpDNA with two K_D values of 364 nM and 86.8 nM and to k20bpDNA with two K_D values of 32 nM and 237 nM (Fig. 4-9). It is assumed that the sequence of the spacer nucleotides between two $CC^A/_GG$ influences the protein DNA binding affinity, although previous reports showed that and the N may be any nucleotide of $CC^A/_GGN_6CC^A/_GG$[45]. The sequence of the six base pairs influences the protein DNA binding affinity, which supports Tfcp2l1 recognizing different DNA sequence to regulate genes' expression.

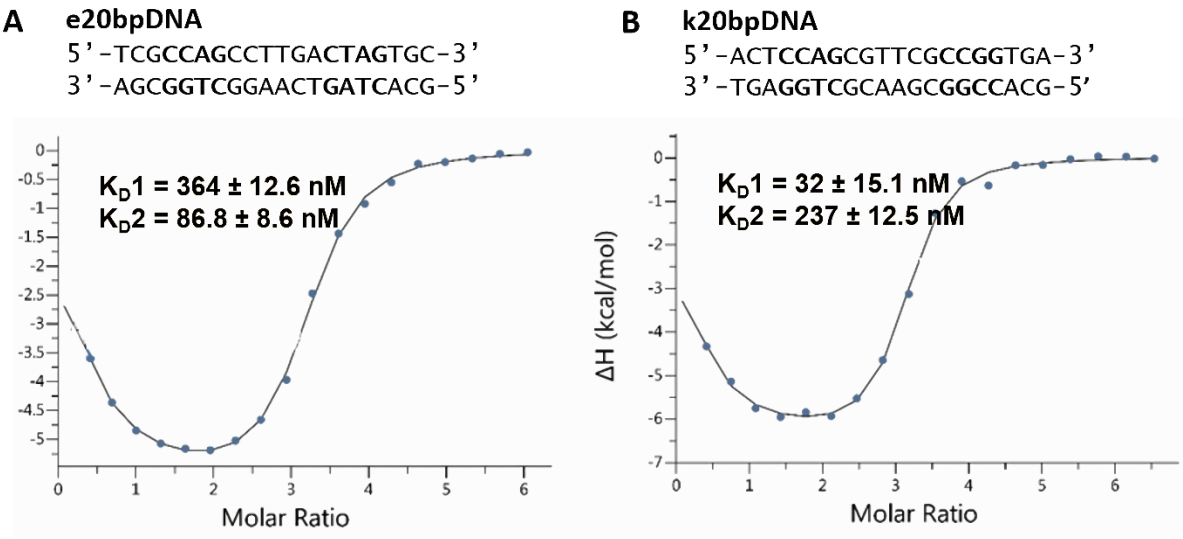

Figure 4-9. ITC measurements of Tfcp2l1 DBD binding to *Esrrb* (A) and *Klf4* promoter (B) sequences. Binding curves fitted to the two-set binding model. Experiments were done in duplicates.

4.7 Model of Tfcp2l1 Δ19:30bpDNA

4.7.1 Model of the Tfcp2l1 C-terminal domain

A model of the Tfcp2l1 SAM domain was built by the SWISS-MODEL program[176] based on the SAM domain structure from protein FLJ21935 (PDB entry 1WWU.A) (Fig. 4-11 A). Furthermore, a model of the SAM domain together with the C-terminus of Tfcp2l1 was also built based on the TelSAM domain (PDB entry 5L0P) (Fig. 4-11 B). The Tfcp2l1 SAM domain and the C-terminus are connected with a long linker, and the SAM domain model contains only helices. However, a model of the Tfcp2l1 C-terminus built with the I-TASSER ONLINE program differs from the model from SWISS-MODEL (Fig. 4-11 C)[192], which contains extensive loop regions without secondary structure. Structure alignment of the SAM domain and the CTD shows that the two SAM models are similar with a RMSD of 1.165 Å from 52 aligned Cα atoms (Fig. 4-11 D). The CTD model predicted by I-TASSER suggested that the CTD is not involved in protein oligomerization due to the lack of a defined conformation. However, it still unclear how the SAM domain and CTD orientation supports DBD to binding to target DNA at multiple sites in the absence of a defined structure of SAM and CTD.

As described previous, the Tfcp2l1 SAM domain is involved in the oligomerization function (Fig. 3-4). However, neither the SAM domain (AAs, 300-365) by itself nor Tfcp2l1 Δ301 (AAs, 301-479) were stable in solution. It has been predicted that the SAM domain may form hexamers and its

stability could be improved by the single-site mutants D345G and G355E[175]. However, *in vitro* experiments have already concluded that these two single mutants do not influence the Tfcp2l1 SAM-domain and C-terminus stability.

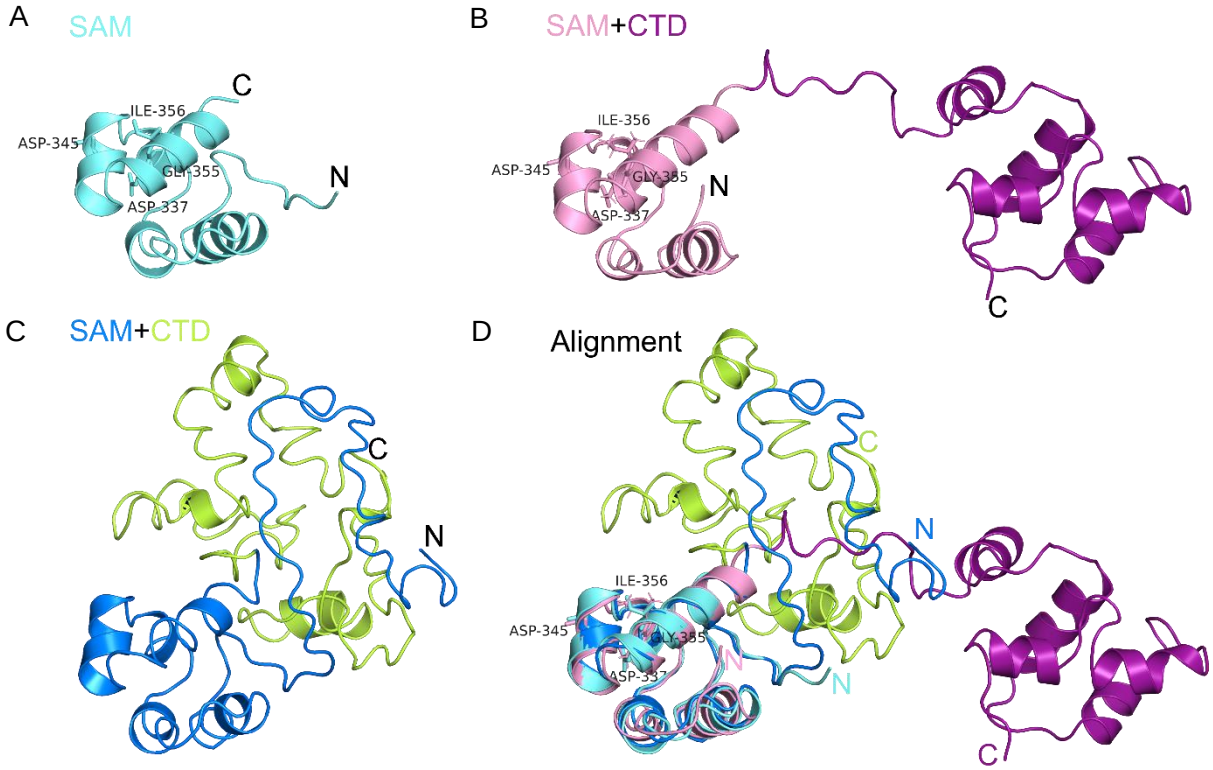

Figure 4-11. Structure models of the Tfcp2l1 SAM domain and SAM+CTD domain. A, structure of the SAM domain. B, structures of SAM and CTD built by the SWISS-MODEL program. C, structure of SAM and CTD built by the I-TASSER ONLINE program. D, superimposition of the SAM with the SAM+CTD structure based on three structure models. Four mutations are labeled. The domains are colored individually.

4.7.2 Cryo-EM model of Tfcp2l1 Δ19:ds30bpDNA

Only the longer protein constructs containing the DBD, such as Tfcp2l1 Δ19 (AAs, 19-479) and Tfcp2l1 Δ42 (AAs, 42-479), were stable during the purification. It is implied that the Tfcp2l1 DBD plays a role in preventing Tfcp2l1 C-terminal domain aggregation or polymer formation. As described before, Tfcp2l1 Δ19 is stable and forms tetramers. One unit of Tfcp2l1 Δ19 has a molecular mass of 52.9 kDa, therefore the Tfcp2l1 Δ19 tetramer has a molecular mass of ~210 kDa reaching the lower size limit of cryo-electron microscopy. For this reason, it was considered to use cryo-EM to determine the structure of a Tfcp2l1 Δ19:ds30bpDNA complex.

In vitro assays have shown that Tfcp2l1 Δ19 bind to 20-mer DNA (N6) with high affinity (Fig. 3-28). According to the Tfcp2l1 DBD:ds12bpDNA complex, it might be that the Tfcp2l1 Δ19 tetramer

binds to 20-mer DNA with 1:1 stoichiometry. The Tfcp2l1 Δ19 negative staining results showed that the particles are tiny, it remained challenging to determine the protein morphology. Therefore, the longer DNA fragment ds30bpDNA was used expecting that more copies of Tfcp2l1 DBD would associate with the duplex DNA, bringing the molecular mass to more than 300 kDa. The ideal geometry of the Tfcp2l1-DNA complex may be one where Tfcp2l1 tetramers reside on opposite faces of duplex DNA, forming a fiber. Here, I first tried ds30bpDNA, not ds40bpDNA, because ds40bpDNA with consensus DNA repeat may easily form a hairpin structure, which would effectively reduce the ds40bpDNA to a molecule resembling ds20bpDNA.

On the negative-stain electron micrographs, particles likely corresponding to Tfcp2l1 Δ19 tetramers and Tfcp2l1 Δ19:ds30bpDNA complexes could be clearly seen. For the cryo-EM tests, the composition of the particles is still under optimization. Therefore, future experiments will focus on the cryo-EM structure of the Tfcp2l1 tetramer and of DNA-bound Tfcp2l1. Based on the overall conformation, it is much clear to elucidate the protein recognize the DNA motifs.

5. APPENDICES

Appendix A: Plasmids

Table A1. Plasmids coding for Tfcp2l1 protein constructs

Name	Vector	Boundaries	Domains	Tag	Purification products
Full length	pQLinkH	1 - 479	--	N-His	Not expressed
SAM1	pQLinkH	283 – 479	SAM+CTD	N-His	High oligomer
SAM2	pQLinkH	301 – 479	SAM+CTD	N-His	High oligomer
SAM3	pQLinkH	262 – 364	SAM	N-His	High oligomer
SAM4	pQLinkH	283 – 364	SAM	N-His	Not tested
SAM11	pQLinkH	47 – 283	DBD	N-His	Soluble
M1 (Δ19)	pQLinkH	19 – 479	--	N-His	19.4 mg/ml
M2	pQLinkH	262 – 479	SAM+CTD	N-His	Not tested
M3	pQLinkH	365 – 479	CTD	N-His	High oligomer
W1	pQLinkH	19 – 283	DBD	N-His	24 mg/ml
W2 (Δ365)	pQLinkH	19 – 364	DBD+SAM	N-His	After Ni-NTA, precipitate
W3	pQLinkH	47 – 364	DBD+SAM	N-His	Not tested
SAM6	pET28a(+)	47 – 479	--	C-His	Not tested
SAM7	pET28a(+)	301 – 370	SAM	C-His	High oligomer
M5	pET28a(+)	19 – 479	--	C-His	High oligomer
M7	pET28a(+)	262 – 479	SAM+CTD	C-His	High oligomer
W4	PQLinkG	1 – 479	--	N-GST	Not expressed
W5	pQLinkG	283 – 479	SAM+CTD	N-GST	Not expressed
W6	pQLinkG	262 – 364	SAM	N-GST	Not expressed
W7	pQLinkH	19 – 260	DBD	N-His	13.6 mg/ml; structured
W8	pQLinkH	47 – 260	DBD	N-His	Not tested
W12	pQLinkH	42 – 260	DBD	N-His	Not tested
W13	pQLinkH	60 – 260	DBD	N-His	Not tested
W14	pQLinkH	396 – 479	CTD	N-His	Not tested
W15	pQLinkH	42 – 283	DBD	N-His	Not tested
W16	pQLinkH	42 – 266	DBD	N-His	12 mg/ml
W17	pQLinkH	42 – 270	DBD	N-His	20 mg/ml

Name	Vector	Boundaries	Domains	Tag	Purification products
W18	pQLinkH	19 – 266	DBD	N-His	16 mg/ml
W19	pQLinkH	19 – 270	DBD	N-His	Not tested
W20 (Δ42)	pQLinkH	42 – 479	--	N-His	8.2 mg/ml
W21	pQLinkH	308 – 479	SAM+CTD	N-His	High oligomer
W22	pQLinkH	366 – 479	CTD	N-His	Not stable after Ni-NTA
Δ266-308	pQLinkH	266 – 308	--	N-His	5 mg/ml
Δ266-366	pQLinkH	266 – 366	DBD+CTD	N-His	8.86 mg/ml

--: indicates the construct contains the whole domains

Table A2. Plasmids coding for Tfcp2 protein constructs

Name	Vector	Boundaries	Domains	Tag	Purification products
WJ1	pQLinkH	1 - 502	--	N-His	Not expressed
WJ3	pQLinkH	60-275	DBD	N-His	24.5 mg/ml
WJ4	pQLinkH	60-502	--	N-His	High oligomer
WJ11	pQLinkH	60-288	DBD	N-His	18.3 mg/ml
WJ12	pQLinkH	34-288	DBD	N-His	High oligomer

--: indicates the construct contains the whole domains. (Constructs were not tested which did not describe here.)

Appendix B: Growth media

Table B. Medium for E.coli growth.

Medium	Reagent	1 liter
LB agar plates	Agar (1.5%)	15 g
	Bacto tryptone	10 g
	NaCl	10 g
	Yeast extract	5 g
	Water	Fill to 1 l
LB medium	Bacto tryptone	10 g
	NaCl	10 g
	Yeast extract	5 g
	Water	Fill to 1 l
8*LB medium	Bacto tryptone	80 g

	NaCl	80 g
	Yeast extract	40 g
	Water	Fill to 1 l
SOB medium	Bacto tryptone	20 g
	KCl (1 M, sterile)	2.5 ml
	$MgCl_2$ (1 M, sterile)	10 ml
	NaCl	0.5 g
	Yeast extract	5 g
	Water	Fill to 1 l
SOC medium	Bacto tryptone	20 g
	glucose (1 M, sterile)	10 ml
	KCl (1 M, sterile)	2.5 ml
	$MgCl_2$ (1 M, sterile)	10 ml
	NaCl	0.5 g
	Yeast extract	5 g
	Water	Fill to 1 l

Appendix C: Primers

Table C1. PCR primers for *Tfcp2l1* gene

Name	Sequence
T7_Fw	5'-TAATACGACTCACTATAGGG-3'
T7_Rv	5'-GCTAGTTATTGCTCAGCGG-3'
pQlink_Fw	5'-TGAGCGGATAACAATTTCACACAG-3'
pQlink_Rv	5'-GGCAACCGAGCGTTCTGAAC-3'
Tfcp2l1_1- Fw	5'-CAGGGATCCATGCTGTTCTGGCACAC-3’
Tfcp2l1_19- Fw	5'-GCATGGATCCTACTTGCGTGATGTGCTGGCTCTG-3’
Tfcp2l1_42- Fw	5'-CAGGATCCGCCCGCTTGCCGCCCCTACA-3’
Tfcp2l1_47- Fw	5'-CGGGATCCCTACAGTATGTGTTGTGTGCCG-3’
Tfcp2l1_60- Fw	5'-CAGGATCCAAGCTACATGAAGAGACCTTAACATACCTC-3’
Tfcp2l1_240- Rv	5'-CTGCGGCCGCTTGAGCCGTTCTTTTTTCCATCTTTTCCC-3’
Tfcp2l1_260- Rv	5'-GCATGCGGCCGCCCATGGAGAACACTCGGTAAGGATGGT-3’
Tfcp2l1_260- Fw	5'-GCATGGATCCTGGCCTGACGTCCCCTACCAGG-3’

Name	Sequence
Tfcp2l1_266- Rv	5'-GAGCGGCCGCTTACTGGTAGGGGACGTCAGGCCA-3'
Tfcp2l1_270- Rv	5'-GAGCGGCCGCTTAGGTGTTGTTCGCCTGGTAGGG-3'
Tfcp2l1_308- Fw	5'-GAGGATCCTCAGCCTCTATCCAGGATGCACAG-3'
Tfcp2l1_365- Rv	5'-GCATGCGGCCGCCCTGCCTTTGATGGCATTGAAGAG-3'
Tfcp2l1_365- Fw	5'-GTACGGATCCAGGAATGTGAGGCCAAAGATGACCA-3'
Tfcp2l1_366- Fw	5'-GAGGATCCAATGTGAGGCCAAAGATGACCATCTAT-3'
Tfcp2l1_396- Fw	5'-GTGGATCCAACAGCCTGTGTGTATACCATGCTATC-3'
Tfcp2l1_479H- Rv	5'-GAAAGCTTCTAGAGTCCACACTTCAGGATGATGTGGTA-3'
Tfcp2l1_479- Rv	5'-GAGCGGCCGCCTAGAGTCCACACTTCAGGATGATGTG-3'
Tfcp2l1_Q435A- Fw	5'-ATCGAGCGGGTCCCGCTGGCATCCACGTGGTGG-3'
Tfcp2l1_Q435A- Rv	5'-GCGGGACCCGCTCGATAGACCCGGTGGATGTGCTG-3'
Tfcp2l1_P46A- Rv	5'-CTGTAGGGCCGGCAAGCGGGCCCCATTCTCAGGAGATAG-3'
Tfcp2l1_P46A- Fw	5'-GCCGGCCCTACAGTATGTGTTGTGTGCCGCCACCTCTC-3'
Tfcp2l1_Q168R- Fw	5'-CATTCGGGTGCACTGTATCAGCACGGAATTCACCCCC-3'
Tfcp2l1_Q168R- Rv	5'-CAGTGCACCCGAATGAATGCAGATGCTCTCTTCGATGGG-3'
Tfcp2l1_V263M- Fw	5'-CTGACATGCCCTACCAGGCGAACAACACCCCATCCCC-3'
Tfcp2l1_V263M- Rv	5'-CTGGTAGGGCATGTCAGGCCATGGAGAACACTCGGTAAGG-3'
Tfcp2l1_S290F- Fw	5'-CAGCTTCCCTAATCACCCGGTGGAGCCCTTACCCCTG-3'
Tfcp2l1_S290F- Rv	5'-GTGATTAGGGAAGCTGTTACCTTCACGGAGGCCAAAGCTGTTGG-3'
Tfcp2l1_D345G- Fw	5'-GAGATGGTTTGGTCCAGGTCTGTGGCCCTGCAGATGGG-3'
Tfcp2l1_D345G- Rv	5'-GGACCAAACCATCTCTGGACATCTTCAGGAGGTCAGCACC-3'
Tfcp2l1_G355E- Fw	5'-CAGATGAGATTCGGCTCTTCAATGCCATCAAAGGCAGGAATGTG-3'
Tfcp2l1_G355E- Rv	5'-GCCGAATCTCATCTGCAGGGCCACAGACCTGGACCAAATC-3'
Tfcp2l1_D337R- Fw	5'-GGTGCTCGCCTCCTGAAGATGTCCAGAGATGATTTGGTCC-3'
Tfcp2l1_D337R- Rv	5'-CTTCAGGAGGCGAGCACCTGAGAAGCTGGCAAAGAGCC-3'
Tfcp2l1_I356R- Fw	5'-TGGGAGACGGCTCTTCAATGCCATCAAAGGCAGGAATG-3'
Tfcp2l1_I356R- Rv	5'-GAAGAGCCGTCTCCCATCTGCAGGGCCACAAACCTGG-3'

Table C2. PCR primers for *Tfcp2* gene

Name	Sequence
Tfcp2_1- Fw	5'-CAGGGATCCATGGCCTGGGCTCTGAAGCTG-3'

Name	Sequence
Tfcp2_34- Fw	5'-CGGGATCCGGTGCTGGTGCCTATAGCATG-3'
Tfcp2_60- Fw	5'-CGGGATCCAATGAGAATAAAATCCTGCCTTTTCAATATGTG-3'
Tfcp2_275- Rv	5'-GGAGCGGCCGCTTATGTGAGTATGGTTGTCTCATAGGAAGCTG-3'
Tfcp2_288- Rv	5'-CAGCGGCCGCTTAGTTATTGACATACGTGATCTCGGGCC-3'
Tfcp2_300- Fw	5'-CGGGATCCAGTTTTTCTCTTGGGGAAGGAAATGGT-3'
Tfcp2_300- Rv	5'-CTGCGGCCGCTTAACTGCTATGGGAACTGTTGAAGCC-3'
Tfcp2_354- Fw	5'-GTGGATCCGGGGCAGATTTATTGAAATTAACTAGAGATG-3'
Tfcp2_398- Fw	5'-GTGGATCCTCACTGCAGTTGAGGGAGCAG-3'
Tfcp2_398- Rv	5'-CTGCGGCCGCTTACTGCTCCCTCAACTGCAGTGA-3'
Tfcp2_420- Fw	5'-CGGGATCCTCAAATGGTACTTTCTTCGTTTACCATGCT-3'
Tfcp2_502- Rv	5'-GGAGCGGCCGCCTACTTCAGTATGATATGATAGCTATCATTGGT-3'
Tfcp2_Q461A- Fw	5'-CAAGGCGGGGCCAACAGGAATTCATGTGCTCATC-3'
Tfcp2_Q461A- Rv	5'-GTTGGCCCGCGCTTGTAAATCTGGCTGATCTGGC-3'

Appendix D: Buffers and solutions

Table D1. Purification buffers and solutions

Buffer names	Components
Ni-NTA chromatography	
His-lysis buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 500 mM NaCl 5% glycerol 0.5 mM DTT 2.5 µg/ml DNase I Protease inhibitor (EDTA-free, cOmplete , 1 tablet / 50 ml)
His-wash buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 500 mM NaCl 5% glycerol 0.5 mM DTT 20 mM imidazole
His-elute buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8

	500 mM NaCl 5% glycerol 300 mM imidazole 0.5 mM DTT
Dialysis buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 200 mM NaCl 5% glycerol 1 mM DTT
Cation/ anion exchange chromatography	
Low salt buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 100 mM NaCl 5% glycerol 1 mM DTT
High salt buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 1 M NaCl 5% glycerol 1 mM DTT
Size exclusion chromatography	
Size Exclusion Chromatography (SEC) Buffer	25 mM HEPES-NaOH, pH 7.2 150 mM NaCl 5% glycerol 0.5 mM TCEP 2 mM $MgCl_2$* 10 mM KCl*
SEC Buffer (pH 7.8)	25 mM HEPES-NaOH, pH 7.8 200 mM NaCl 5% glycerol 0.5 mM TCEP

Table D2. Protein:DNA interaction buffers

Protein:DNA interaction buffer (pH 7.2)	25 mM HEPES-NaOH, pH 7.2 125 mM NaCl 2 mM $MgCl_2$ 0.5mM TCEP

Protein:DNA interaction buffer (pH 7.8)	25 mM HEPES-NaOH, pH 7.8 125 mM NaCl 0.5mM TCEP
DNA buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 100 mM NaCl
ITC buffer	25 mM HEPES-NaOH, pH 7.2/ pH 7.8 125 mM NaCl 0.5mM TCEP
RALS buffer (pH 7.2)	25 mM HEPES-NaOH, pH 7.2 125 mM NaCl 0.5mM TCEP
RALS buffer (pH 7.8)	25 mM HEPES-NaOH, pH 7.8 150 mM NaCl 1 mM DTT

Appendix E: SDS-PAGE related buffers and solutions

Table E. SDS-PAGE gel related buffers

Gel	**components**	**Amount for 12 gels**
12% resolving gel	Separation buffer	17.5 ml
	Acrylamide	28.0 ml
	H_2O	23.8 ml
	10% SDS	700 µl
	10% APS	600 µl
	TEMED	60 µl
15% resolving gel	Separation buffer	17.5 ml
	Acrylamide	35.0 ml
	H_2O	16.8 ml
	10% SDS	700 µl
	10% APS	600 µl
	TEMED	60 µl
5% stacking gel	Stacking buffer	7.5 ml
	Acrylamide	5.0 ml
	H_2O	17.2 ml
	10% SDS	300 µl

	10% APS TEMED Bromophenol blue (0.1%)	240 µl 30 µl 100 µl
4x SDS sample buffer	50 mM Tris/HCl 100 mM DTT 2% (w/v) SDS 0.25 (w/v) bromophenol blue 10% glycerol	SDS-PAGE sample loading
Separation buffer	1.5 M Tris/HCl, pH 8.8	SDS-PAGE gel component
Stacking buffer	0.5 M Tris/HCl, pH 6.8	SDS-PAGE gel component
10 x SDS running buffer	250 mM Tris 2 M glycine 1% (w/v) SDS	SDS-PAGE running buffer
Staining solution 1	50% v/v Ethanol 10% v/v Acetic acid	Gel staining
Staining solution 2	5% v/v Ethanol 7.5% v/v Acetic acid	Gel staining
Staining solution 3	0.25% Coomassie R250 in Ethanol	Gel staining
Wet transfer buffer	25 mM Tris/HCl, pH 8.0 192 mM glycine 20% methanol 0.04% SDS	Western blot
TBST buffer	50 mM Tris/HCl, pH 8.0 150 mM NaCl 0.1% Tween 20	Western blot
Antibody incubation buffer	50 mM Tris/HCl, pH 8.0 150 mM NaCl 0.1% Tween 20 3% skimmed milk powder	Western blot

Appendix F: Oligonucleotides

Table F. Oligonucleotides used for protein:DNA binding assays

Name	Sequence
K30bp	5'-CTTCTAGCTTACTCCAGTCTTCGCCGGTGA-3'
	3'-GAAGATCGAATGAGGTCAGAAGCGGCCACT-5'
ds22bpCG	5'-GACACCGGTTTAAACCGGTGTC-3'
	3'-CTGTGGCCAAATTTGGCCACAG-5'
ds22bpDNA	5'-GACACCAGTTCAAACCAGTGTC-3'
	3'-CTGTGGTCAAGTTTGGTCACAG-5'
Cr21bp	5'-CTCGCCAGTTCAAACCAGTGC-3'
	3'-AGCGGTCAAGTTTGGTCACGC-5'
ds21bpDNA	5'-CGAACCAGTTTGAACCAGTTC-3'
	3'-CTTGGTCAAACTTGGTCAAGC-5'
K20bp	5'-ACTCCAGCGTTCGCCGGTGA-3'
	3'-TGAGGTCGCAAGCGGCCACT-5'
K20bpt	5'-ACTCCAGTCTTCGCCGGTGA-3'
	3'-TGAGGTCAGAAGCGGCCACT-5'
K20bpc	5'-CTTCTAGCTTACTCCAGTCT-3'
	3'-GAAGATCGAATGAGGTCAGA-5'
ds20bpTTC	5'-TCGCCAGCCTTGACCAGTGC-3'
	3'-AGCGGTCGGAACTGGTCACG-5'
e20bpDNA	5'-TCGCCAGCCTTGACTAGTGC-3'
	3'-AGCGGTCGGAACTGATCACG-5'
ds20bpDNA	5'-GAACCAGTTTGAACCAGTTC-3'
	3'-CTTGGTCAAACTTGGTCAAG-5'
Cr20bp	5'-TCGCCAGTTCAAACCAGTGC-3'
	3'-AGCGGTCAAGTTTGGTCACG-5'
ds20bpTC	5'-GAACCAGTTCGAACCAGTTC-3'
	3'-CTTGGTCAAGCTTGGTCAAG-5'
ds19bpDNA	5'-GAACCAGTTCAAACCAGTT-3'
	3'-TTGGTCAAGTTTGGTCAAG-5'
ds18bpDNA	5'-AACCAGTTCAAACCAGTT-3'
	3'-TTGGTCAAGTTTGGTCAA-5'
ds16bpDNA	5'-ACCAGTTCAAACCAGT-3'
	3'-TGGTCAAGTTTGGTCA-5'
ds14bpDNA	5'-CCAGTTCAAACCAG-3'
	3'-GGTCAAGTTTGGTC-5'
ds12bpGAG	5'-GAAACCAGTTTC-3'
	3'-CTTTGGTCAAAG-5'
ds12bpAG	5'-AAAACCAGTTTT-3'
	3'-TTTTGGTCAAAA-5'
ds12DNA	5'-GAAACCGGTTTC-3'
	3'-CTTTGGCCAAAG-5'

Name	Sequence
ds12bpAC	5’-GACACCAGTGTC-3'
	3'-CTGTGGTCACAG-5’
ds12bpDNA	5’-AAAACCGGTTTT-3'
	3'-TTTTGGCCAAAA-5’
ds10bp	5’-GAACCAGTTC-3'
	3'-CTTGGTCAAG-5’
8bpCG	5’-CACCGGTG-3'
	3'-GTGGCCAC-5’
8bpCA	5’-CACCAATG-3'
	3'-GTGGTTAC-5’
8bp	5’-CACCAGTG-3'
	3'-GTGGTCAC-5’
N7	5’-AAACCAGTTCAAAACCAGTTT-3’
	3’-TTTGGTCAAGTTTTGGTCAAA-5’
N6	5’-AAACCAGTTCAAACCAGTTT-3’
	3’-TTTGGTCAAGTTTGGTCAAA-5’
N5	5’-AAACCAGTTAAACCAGTTT-3’
	3’-TTTGGTCAATTTGGTCAAA-5’
N4	5’-AAACCAGTTAACCAGTTT-3’
	3’-TTTGGTCAATTGGTCAAA-5’
N3	5’-AAACCAGTTACCAGTTT-3’
	3’-TTTGGTCAATGGTCAAA-5’

Appendix G: Abbreviations

Abbreviations	
Å	Ångström (1 Å = 0.1 nm)
AA(s)	Amino acid(s)
BESSY II	Berliner Elektronenspeicherring II
bp	Base pair(s)
C-terminus	Carboxy terminus
CV	Column volume
Da	Dalton
CTD	Carboxyl-terminal domain
DBD	DNA-binding domain

DNAse I	Deoxyribonuclease I
dNTP	Deoxyribo-nucleoside triphosphate
DTT	Dithiothreitol
E. coli	Escherichia coli
EG	Ethylene glycol
GST	Glutathione-S-transferase
H (bond)	Hydrogen (bond)
h	Hour
HF	High-fidelity
g	Gram
Ig	Immunoglobulin
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
K_D	Dissociation constant
kDa	Kilodalton
LB	Luria-Bertani medium
MALDI	Matrix-assisted laser desorption/ionization
MS	Mass spectrometry
N-terminus	Amino terminus
ng	Nanogram
ml	Milliliter
min	Minute
OD_{600}	Optical density of the sample, measured at the wavelength of 600 nm
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
PWM	Position weight matrix
RALS	Right-angle light scattering
RI	Refractive index
RT	Room temperature

RMSD	Root-mean-square deviation
rpm	Revolution per minute
s	Second
SAM	Sterile alpha motif
SEC	Size exclusion chromatography
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
TEMED	Tetramethylethylenediamine
Tris	Tris(hydroxymethyl) aminomethane

6. REFERENCES:

1 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263 (2009).

2 Furney, S. J., Higgins, D. G., Ouzounis, C. A. & Lopez-Bigas, N. Structural and functional properties of genes involved in human cancer. *BMC genomics* **7**, 3 (2006).

3 Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157 (2005).

4 Wilanowski T, T. A., Cerruti L, O'Connell S, Saint R, Parekh V, Tao J, Cunningham JM, Jane SM. A highly conserved novel family of mammalian developmental transcription factors related to Drosophila grainyhead. *Mech Dev* **114**, 14 (2002).

5 Venkatesan, K., McManus, H. R., Mello, C. C., Smith, T. F. & Hansen, U. Functional conservation between members of an ancient duplicated transcription factor family, LSF/Grainyhead. *Nucleic Acids Res.* **31(15):4304-16.** (2003).

6 Wang, S. & Samakovlis, C. Grainy head and its target genes in epithelial morphogenesis and wound healing. *Current topics in developmental biology* **98**, 35-63 (2012).

7 Kim, C. H., Heath, C., Bertuch, A. & Hansen, U. Specific stimulation of simian virus 40 late transcription in vitro by a cellular factor binding the simian virus 40 21-base-pair repeat promoter element. *Proc Natl Acad Sci* **84(17):6025-9** (1987).

8 Huang, H.-C., Sundseth, R. & Hansen, U. Transcription factor LSF binds two variant bipartite sites within the SV40 late promoter. *Genes Dev* **4(2):287-98** (1990).

9 Porta-de-la-Riva, M. *et al.* TFCP2c/LSF/LBP-1c is required for Snail1-induced fibronectin gene expression. *The Biochemical journal* **435**, 563-568 (2011).

10 Lim, L. C., Swendeman, S. L. & Sheffery, M. Molecular Cloning of the α-Globin Transcription Factor CP2. *Mol Cell Biol* **12**, 8 (1992).

11 Sato, F., Yasumoto, K., Kimura, K., Numayama-Tsuruta, K. & Sogawa, K. Heterodimerization with LBP-1b is necessary for nuclear localization of LBP-1a and LBP-1c. *Genes to cells : devoted to molecular & cellular mechanisms* **10**, 861-870 (2005).

12 Chodosh, L. A., Baldwin, A. S., Carthew, R. W. & Sharp, P. A. Human CCAAT-binding proteins have heterologous subunits. *Cell* **53** (1988).

13 Veljkovic, J. & Hansen, U. Lineage-specific and ubiquitous biological roles of the mammalian transcription factor LSF. *Gene* **343**, 23-40 (2004).

14 Bray, S. J. & Kafatos, F. C. Developmental function of Elf-1: an essential transcription factor during embryogenesis in Drosophila. *Genes Dev* **5**, 1672-1683 (1991).

15 Bray, S. J., Jackson, A., Hirsh, J., Heberlein, U. & Tjian, R. A cis-acting element and associated binding factor required for CNS expression of the Drosophila melanogaster dopa decarboxylase gene. *EMBO J.* **7(1): 177-188.** , (1988).

16 Biggin, M. D. & Tjian, R. Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. *Cell* **3**, 699-711 (1988).

17 Dynlacht, B. D. *et al.* Functinal analysis of NTF-1, a developmentally regualted Drosophila transcription factor that binds neuronal cis elements. *Genes dev* **3: 1677-1688** (1989).

18 Bray, S. J., Burke, B., Brown, N. H. & Hirsh, J. Embryonic expression pattern of a family of Drosophila proteins that interact with a central nervous system regulatory element. *Genes Dev* **3**, 1130-1145 (1989).

19 Kokoszynska, K., Ostrowski, J., Rychlewski, L. & Wyrwicz, L. S. The fold recognition of CP2 transcription factors gives new insights into the function and evolution of tumor suppressor protein p53. *Cell cycle* **7**, 2907-2915 (2008).

20 Harden, N. Of Grainy Heads and Broken Skins. *Science* **308(5720): 364-365** (2005).

21 Boivin, F. J. & Schmidt-Ott, K. M. Functional roles of Grainyhead-like transcription factors in renal development and disease. *Pediatric nephrology* **35**, 181-190 (2018).

22 Mlacki, M., Kikulska, A., Krzywinska, E., Pawlak, M. & Wilanowski, T. Recent discoveries concerning the involvement of transcription factors from the Grainyhead-like family in cancer. *Experimental biology and medicine* **240**, 1396-1401 (2015).

23 Ming, Q. *et al.* Structural basis of gene regulation by the Grainyhead/CP2 transcription factor family. *Nucleic Acids Res* **46**, 2082-2095 (2018).

24 Kotarba, G., Krzywinska, E., Grabowska, A. I., Taracha, A. & Wilanowski, T. TFCP2/TFCP2L1/UBP1 transcription factors in cancer. *Cancer letters* **420**, 72-79 (2018).

25 Taracha, A., Kotarba, G. & Wilanowski, T. Neglected Functions of TFCP2/TFCP2L1/UBP1 Transcription Factors May Offer Valuable Insights into Their Mechanisms of Action. *International journal of molecular sciences* **19** (2018).

26 Bray, S. J., Burke, B., Brown, N. H. & Hirsh, J. Embryonic expression pattern of a family of Drosophila proteins that interact with a central nervous system regulatory element. *Genes Dev* **3:1130-1145** (1989).

27 Tao, J. *et al.* BMP4-dependent expression of Xenopus Grainyhead-like 1 is essential for epidermal differentiation. *Development* **132**, 1021-1034 (2005).

28 Janicke, M., Renisch, B. & Hammerschmidt, M. Zebrafish grainyhead-like1 is a common marker of different non-keratinocyte epidermal cell lineages, which segregate from each other in a Foxi3-dependent manner. *The International journal of developmental biology* **54**, 837-850 (2010).

29 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12(6):996-1006.** (2002).

30 Rodda, S., Sharma, S., Scherer, M., Chapman, G. & Rathjen, P. CRTR-1, a developmentally regulated transcriptional repressor related to the CP2 family of transcription factors. *The Journal of biological chemistry* **276**, 3324-3332 (2001).

31 Huang, N. & Miller, W. L. Cloning of factors related to HIV-inducible LBP proteins that regulate steroidogenic factor-1-independent human placental transcription of the cholesterol side-chain cleavage enzyme P450scc. *The Journal of biological chemistry* (2000).

32 Sueyoshi, T. *et al.* A nuclear factor (NF2d9) that binds to the male-specific P450 (Cyp 2d-9) gene in mouse liver. *Mol Cell Biol* **15(8):4158-4166** (1995).

33 Barnhart, K. M., Kim, C. G., Banerji, S. S. & Sheffery, M. Identification and characterization of multiple erythroid cell proteins that interact with the promoter of the murine alpha-globin gene. *Mol Cell Biol* **8(8):3215-26** (1988).

34 Kang, H. C. *et al.* Transcription Factor CP2 Is Involved in Activating mBMP4 in Mouse Mesenchymal Stem Cells. *Mol Cells* **17(3):454-61.** (2004).

35 Yoon, J. B., Li, G. & Roeder, R. G. Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type 1 transcription in vitro. *Mol Cell Biol* **14(3): 1776 - 1785** (1994).

36 Swendeman, S. L. *et al.* Characterization of the Genomic Structure, Chromosomal Location Promoter, and Developmental Expressiono f the a-Globin Transcription Factor CP2. *The Journal of biological chemistry* **269** (1994).

37 Ramamurthy, L. *et al.* Targeted disruption of the CP2 gene, a member of the NTF family of transcription factors. *The Journal of biological chemistry* **276**, 7836-7842 (2001).

38 Huang, N. & Miller, W. L. LBP proteins modulate SF1-independent expression of P450scc in human placental JEG-3 cells. *Molecular endocrinology* **19**, 409-420 (2005).

39 Rodda, S. J., Kavanagh, S. J., Rathjen, J. & Rathjen, P. D. Embryonic stem cell differentiation and the analysis of mammalian development. *The internatonal journal of development biology* **46: 449-458** (2002).

40 Pelton, T. A., Sharma, S., Schulz, T. C., Rathjen, J. & Rathjen, P. D. Transient pluripotent cell populations during primitive ectoderm formation correlation of in vivo and in vitro pluripotent cell development. *Journal of cell science* **115: 329-339** (2002).

41 Sundseth, R. & Hansen, U. Activation of RNA polymerase II transcription by the specific DNA-binding protein LSF. Increased rate of binding of the basal promoter factor TFIIB. *The Journal of biological chemistry* **267(11):7845-55.** (1992).

42 Coull, J. J. *et al.* The human factors YY1 and LSF repress the human immunodeficiency virus type 1 long terminal repeat via recruitment of histone deacetylase 1. *J Virol* **74(15):6790-6799** (2000).

43 Sarah To, Stephen J. Rodda, Peter D. Rathjen & Keough*, R. A. Modulation of CP2 Family Transcriptional Activity by CRTR-1 and Sumoylation. *PLoS ONE 5(7):* **e11702.** (2010).

44 Kato, H., Horikoshi, M. & Roeder, R. G. Repression of HIV-1 transcription by a cellular protein. *Science* **251(5000):1476-1479** (1991).

45 Lim, L. C., Fang, L., Swendeman, S. L. & Sheffery, M. Characterization of the molecularly cloned murine alpha-globin transcription factor CP2. *The Journal of biological chemistry* **268(24):18008-18017.** (1993).

46 Frith, M. C., Hansen, U. & Weng, Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* **17(10):878-89.** (2001).

47 Lim, L. C., Fang, L., Swendeman, S. L. & Sheffery, M. Characterization of the Molecularly Cloned Murine a-Globin TF CP2. *The Journal of biological chemistry* **25**, 10 (1993).

48 Murata, T., Nitta, M. & Yasuda, K. Transcription factor CP2 is essential for lens-specific expression of the chicken αA-crystallin gene. *Genes to Cells* **3, 443-457** (1998).

49 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92 (2020).

50 Shirra, M. K. & Hansen, U. LSF and NTF-1 share a conserved DNA recognition motif yet require different oligomerization states to form a stable protein-DNA complex. *The Journal of biological chemistry* **273(30):19260-8.** (1998).

51 Shirra, M. K., Zhu, Q., Huang, H. C., Pallas, D. & Hansen, U. One exon of the human LSF gene includes conserved regions involved in novel DNA-binding and dimerization motifs. *Mol Cell Biol*, 5076-5087 (1994).

52 Kim, C. M., Jang, T. H. & Park, H. H. Functional Analysis of CP2-Like Domain and SAM-Like Domain in TFCP2L1, Novel Pluripotency Factor of Embryonic Stem Cells. *Applied biochemistry and biotechnology* **179**, 650-658 (2016).

53 Schultz, J., Ponting, C. P., Hofmann, K. & Bork, P. SAM as a protein interaction domain involved in developmental regulation. *Protein science : a publication of the Protein Society* **6(1):249-53.** (1997).

54 Kim, C. G., Swendeman, S. L., Barnhart, K. M. & Sheffery, M. Promoter Elements and Erythroid Cell Nuclear Factors That Regulate a Globin Gene Transcription in Vitro. *Mol Cell Biol* **10** (1990).

55 Chae, J. H., Lee, Y. H. & Kim, C. G. Transcription factor CP2 is crucial in hemoglobin synthesis during erythroid terminal differentiation in vitro. *Biochem Biophys Res Commun* **263(2):580-583.** (1999).

56 Solis, C., Aizencang, G. I., Astrin, K. H., Bishop, D. F. & Desnick, R. J. Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. *The Journal of clinical investigation* **107(6):753-62.** (2001).

57 Stavnezer, J. Immunoglobulin class switching. *Curr Opin Immunol* **8(2):199-205.** (1996).

58 Drouin, E. E., Schrader, C. E., Stavnezer, J. & Hansen, U. The ubiquitously expressed DNA-binding protein late SV40 factor binds Ig switch regions and represses class switching to IgA. *Journal of immunology* **168**, 2847-2856 (2002).

59 Volker, J. L., Rameh, L. E., Zhu, Q., DeCaprio, J. & Hansen, U. Mitogenic stimulation of resting T cells causes rapid phosphorylation of the transcription factor LSF and increased DNA-binding activity. *Genes Dev* **11(11):1435-46.** (1997).

60 Pagon, Z., Volker, J., Cooper, G. M. & Hansen, U. Mammalian transcription factor LSF is a target of ERK signaling. *J Cell Biochem* **89**, 733-746 (2003).

61 Casolaro, V. *et al.* Identification and characterization of a critical CP2-binding element in the human interleukin-4 promoter. *The Journal of biological chemistry* **275**, 36605-36611, doi:10.1074/jbc.M007086200 (2000).

62 El-Serag, H. B. & Rudolph, K. L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* **132**, 2557-2576 (2007).

63 Yoo, B. K. *et al.* Identification of genes conferring resistance to 5-fluorouracil. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12938-12943 (2009).

64 Yoo, B. K. *et al.* Astrocyte elevated gene-1 regulates hepatocellular carcinoma development and progression. *The Journal of clinical investigation* **119**, 465-477 (2009).

65 Yoo, B. K. *et al.* Transcription factor Late SV40 Factor (LSF) functions as an oncogene in hepatocellular carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8357-8362 (2010).

66 Robertson, C. L. *et al.* The role of AEG-1 in the development of liver cancer. *Hepat Oncol* **2(3): 303 - 312** (2015).

67 Fan, R. H., Li, J., Wu, N. & Chen, P. S. Late SV40 factor: a key mediator of Notch signaling in human hepatocarcinogenesis. *World journal of gastroenterology* **17**, 3420-3430 (2011).

68 Ning, L., Wentworth, L., Chen, H. & Weber, S. M. Down-regulation of Notch1 signaling inhibits tumor growth in human hepatocellular carcinoma. *Am J Transl Res.* **1(4):358-66.** (2009).

69 Fan, R. *et al.* Cooperation of deregulated Notch signaling and Ras pathway in human hepatocarcinogenesis. *Journal of molecular histology* **42**, 473-481 (2011).

70 Wang, R., Ferrell, L. D., Faouzi, S., Maher, J. J. & Bishop, J. M. Activation of the Met receptor by cell attachment induces and sustains hepatocellular carcinomas in transgenic mice. *J Cell Biol* **153(5):1023-34.** (2001).

71 Birchmeier, C., Birchmeier, W., Gherardi, E. & Vande Woude, G. F. Met, metastasis, motility and more. *Nature reviews. Molecular cell biology* **4**, 915-925 (2003).

72 Jeon, H. M. & Lee, J. MET: roles in epithelial-mesenchymal transition and cancer stemness. *Annals of translational medicine* **5**, 5 (2017).

73 Santhekadur, P. K. *et al.* The transcription factor LSF: a novel oncogene for hepatocellular carcinoma. *Am J Cancer Res* **2(3):269-85** (2012).

74 Yoo, B. K. *et al.* c-Met activation through a novel pathway involving osteopontin mediates oncogenesis by the transcription factor LSF. *Journal of hepatology* **55**, 1317-1324 (2011).

75 Coutessens, L. M., Fingleton, B. & Matrisian, L. M. Matrix metalloproteinase inhibitors and cancer trials and tribulations. *science* **29;295(5564)**, 2387-2392 (2002).

76 Kessenbrock, K., Plaks, V. & Werb, Z. Matrix Metalloproteinases: Regulators of the Tumor Microenvironment. *Cell* **141**, 52-67 (2010).

77 Yu, M. H. *et al.* Inhibitory effect of immature plum on PMA-induced MMP-9 expression in human hepatocellular carcinoma. *Natural Product Research* **23**, 704-718 (2009).

78 Santhekadur, P. K. *et al.* Late SV40 factor (LSF) enhances angiogenesis by transcriptionally up-regulating matrix metalloproteinase-9 (MMP-9). *The Journal of biological chemistry* **287**, 3425-3432 (2012).

79 Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial–mesenchymal transition. *Nature Reviews Molecular Cell Biology* **15**, 178-196 (2014).

80 Xu, X. *et al.* Characterization of genome-wide TFCP2 targets in hepatocellular carcinoma: implication of targets FN1 and TJP1 in metastasis. *Journal of experimental & clinical cancer research : CR* **34**, 6 (2015).

81 Powell, C. M., Rudge, T. L., Zhu, Q., Johnson, L. F. & Hansen, U. Inhibition of the mammalian transcription factor LSF induces S‐phase‐dependent apoptosis by downregulating thymidylate synthase expression. *The EMBO journal* **19**, 4665-4675 (2000).

82 Grant, T. J. *et al.* Antiproliferative small-molecule inhibitors of transcription factor LSF reveal oncogene addiction to LSF in hepatocellular carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 4503-4508 (2012).

83 Koehler, A. N. A complex task? Direct modulation of transcription factors with small molecules. *Current opinion in chemical biology* **14**, 331-340 (2010).

84 LaCroix, A. Z. *et al.* Health Outcomes After Stopping Conjugated Equine Estrogens Among Postmenopausal Women With Prior Hysterectomy. *JAMA* **305(13):1305-1314** (2011).

85 Qin, C., Samudio, I., Ngwenya, S. & Safe, S. Estrogen-dependent regulation of ornithine decarboxylase in breast cancer cells through activation of nongenomic cAMP-dependent pathways. *Molecular carcinogenesis* **40**, 160-170 (2004).

86 Ling, H., Fabbri, M. & Calin, G. A. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature Reviews Drug Discovery* **12**, 847-865 (2013).

87 He, T. *et al.* miR‐660‐5p is associated with cell migration, invasion, proliferation and apoptosis in renal cell carcinoma. *Molecular Medicine Reports* (2017).

88 Shen, Y. *et al.* Inhibition of miR-660-5p expression suppresses tumor development and metastasis in human breast cancer. *Genetics and molecular research : GMR* **16** (2017).

89 Zhao, Y. *et al.* A Feedback loop comprising EGF/TGF-α Sustains TFCP2-mediated Breast Cancer Progression. *Cancer Research*, canres.2908.2019 (2020).

90 Győrffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Research and Treatment* **123**, 725-731 (2009).

91 Labidi-Galy, S. I. *et al.* Elafin drives poor outcome in high-grade serous ovarian cancers and basal-like breast tumors. *Oncogene* **34**, 373-383 (2014).

92 Warnakulasuriya, S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncology* **45**, 309-316 (2009).

93 Fu, J., Bian, M., Jiang, Q. & Zhang, C. Roles of Aurora Kinases in Mitosis and Tumorigenesis. *Molecular Cancer Research* **5**, 1-10 (2007).

94 Tanaka, H. *et al.* Targeting Aurora kinase A suppresses the growth of human oral squamous cell carcinoma cells in vitro and in vivo. *Oral Oncology* **49**, 551-559 (2013).

95 Jiang, J., Guo, Z., Xu, J., Sun, T. & Zheng, X. Identification of Aurora Kinase A as a Biomarker for Prognosis in Obesity Patients with Early Breast Cancer. *OncoTargets and Therapy* **Volume 13**, 4971-4985 (2020).

96 Chen, C. H. *et al.* Metformin disrupts malignant behavior of oral squamous cell carcinoma via a novel signaling involving Late SV40 factor/Aurora-A. *Scientific reports* **7**, 1358 (2017).

97 Guarneri, C. *et al.* NFkappaB inhibition is associated with OPN/MMP9 downregulation in cutaneous melanoma. *Oncology reports* **37**, 737-746 (2017).

98 Xu, Y. *et al.* The co-expression of MMP-9 and Tenascin-C is significantly associated with the progression and prognosis of pancreatic cancer. *Diagnostic pathology* **10**, 211 (2015).

99 Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Gastroenterology Review* **14**, 89-103 (2019).

100 Jiang, H. *et al.* LSF expression and its prognostic implication in colorectal cancer. *Int J Clin Exp Pathol* **7(9):6024-6031** (2014).

101 de Martel, C., Plummer, M., Vignat, J. & Franceschi, S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *International journal of cancer* **141**, 664-670 (2017).

102 Wagner, K.-U. *et al.* Tsg101 Is Essential for Cell Growth, Proliferation, and Cell Survival of Embryonic and Adult Tissues. *Molecular and Cellular Biology* **23**, 150-162 (2003).

103 Broniarczyk. Analysis of expression and structure of the TSG101 gene in cervical cancer cells. *International Journal of Molecular Medicine* **25** (2010).

104 Broniarczyk, J. K. *et al.* Expression of TSG101 protein and LSF transcription factor in HPV-positive cervical cancer cells. *Oncology letters* **7**, 1409-1413 (2014).

105 Brett M, R. *et al.* Epidemiology of ovarian cancer: a review. *Cancer Biology & Medicine* **14**, 9-32 (2017).

106 Tcherkassova, J. *et al.* Combination of CA125 and RECAF biomarkers for early detection of ovarian cancer. *Tumor Biology* **32**, 831-838 (2011).

107 Kaur, M. *et al.* In Silico discovery of transcription factors as potential diagnostic biomarkers of ovarian cancer. *BMC Syst Biol* **5:144** (2011).

108 Ryan, N. S., Rossor, M. N. & Fox, N. C. Alzheimer's disease in the 100 years since Alzheimer's death. *Brain* **138**, 3816-3821 (2015).

109 Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimers disease progress and problems on the road to therapeutics. *Science* **297(5580):353-356.** (2002).

110 Telese, F. *et al.* Transcription regulation by the adaptor protein Fe65 and the nucleosome assembly factor SET. *EMBO reports* **6**, 77-82 (2005).

111 Nakaya, T., Kawai, T. & Suzuki, T. Regulation of FE65 Nuclear Translocation and Function by Amyloid β-Protein Precursor in Osmotically Stressed Cells. *Journal of Biological Chemistry* **283**, 19119-19131 (2008).

112 Kashour, T., Burton, T., Dibrov, A. & Amara, F. M. Late Simian virus 40 transcription factor is a target of the phosphoinositide 3-kinase Akt pathway in anti-apoptotic Alzheimer's amyloid precursor protein signalling. *The Biochemical journal* **370(3): 1063 - 1075** (2003).

113 Bruni, P. *et al.* Fe65, a ligand of the Alzheimer's beta-amyloid precursor protein, blocks cell cycle progression by down-regulating thymidylate synthase expression. *The Journal of biological chemistry* **277**, 35481-35488 (2002).

114 Lichtenthaler, S. F. Alpha-secretase cleavage of the amyloid precursor protein proteolysis regulated by signaling pathways and protein trafficking. *Curr Alzheimer Res* **9(2):165-77.** (2012).

115 Panza, F. *et al.* LBP-1c/CP2/LSF gene polymorphism and risk of sporadic Alzheimer' disease. *J Neurol Neurosurg Psychiatry* **75(1): 166 - 168** (2004).

116 Lambert, J. C. *et al.* The transcriptional factor LBP-1cCP2LSF gene on chromosome 12 is a genetic determinant of Alzheimers disease. *Hum Mol Genet* **9(15):2275-2280** (2000).

117 Parada, C. A., Yoon, J. B. & Roeder, R. G. A Novel LBP-1-mediated Restriction of HIV-1 Transcription at the Level of Elongation in Vitro. *The Journal of biological chemistry* **270: 2274 - 2283** (1995).

118 Romerio, F., Gabriel, M. N. & Margolis, D. M. Repression of Human Immunodeficiency Virus Type 1 through the Novel Cooperation of Human Factors YY1 and LSF. *J Virol* **71(12): 9375 - 9382** (1997).

119 Pomerantz, R. J. Reservoirs, sanctuaries, and residual disease: the hiding spots of HIV-1. *HIV clinical trials* **4**, 137-143 (2003).

120 Smith, A. G. Embryo-derived stem cells of mice and men. *Annu Rev Cell Dev Biol.* **17:435-62** (2001).

121 Matsuda, K. *et al.* ChIP-seq analysis of genomic binding regions of five major transcription factors highlights a central role for ZIC2 in the mouse epiblast stem cell gene regulatory network. *Development* **144**, 1948-1958 (2017).

122 Ouyang, Z., Zhou, Q. & Wong, W. H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 21521-21526 (2009).

123 Liu, K., Zhang, Y., Liu, D., Ying, Q. L. & Ye, S. TFCP2L1 represses multiple lineage commitment of mouse embryonic stem cells through MTA1 and LEF1. *Journal of cell science* **130**, 3809-3817 (2017).

124 Villarejo, A., Cortés-Cabrera, Á., Molina-Ortíz, P., Portillo, F. & Cano, A. Differential Role of Snail1 and Snail2 Zinc Fingers in E-cadherin Repression and Epithelial to Mesenchymal Transition. *Journal of Biological Chemistry* **289**, 930-941 (2014).

125 Galvagni, F. *et al.* Snai1 promotes ESC exit from the pluripotency by direct repression of self-renewal genes. *Stem cells* **33**, 742-750 (2015).

126 Ying, Q.-L., Nichols, J., Chambers, I. & Smith, A. BMP Induction of Id Proteins Suppresses Differentiation and Sustains Embryonic Stem Cell Self-Renewal in Collaboration with STAT3. *Cell* **115**, 281-292 (2003).

127 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117 (2008).

128 Martello, G., Bertone, P. & Smith, A. Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. *The EMBO journal* (2013).

129 Ye, S., Li, P., Tong, C. & Ying, Q. L. Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. *EMBO J* **32**, 2548-2560 (2013).

130 Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol.* **7(7):540-6** (2006).

131 Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell* **10**, 105-116 (2006).

132 Loh, Y. H., Zhang, W., Chen, X., George, J. & Ng, H. H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes Dev* **21**, 2545-2557 (2007).

133 Short, K. M. & Smyth, I. M. The contribution of branching morphogenesis to kidney development and disease. *Nature reviews. Nephrology* **12**, 754-767 (2016).

134 Pei, D., Shu, X., Gassama-Diagne, A. & Thiery, J. P. Mesenchymal-epithelial transition in development and reprogramming. *Nature cell biology* **21**, 44-53 (2019).

135 Tun, H. W. *et al.* Pathway signature and cellular differentiation in clear cell renal cell carcinoma. *PloS one* **5**, e10696 (2010).

136 Zaravinos, A. *et al.* New miRNA profiles accurately distinguish renal cell carcinomas and upper tract urothelial carcinomas from the normal kidney. *PloS one* **9**, e91646 (2014).

137 Yamaguchi, Y., Yonemura, S. & Takada, S. Grainyhead-related transcription factor is required for duct maturation in the salivary gland and the kidney of the mouse. *Development* **133**, 4737-4748 (2006).

138 Max Werth, K. M. S.-O., Thomas Leete, Andong Qiu, Christian Hinze, Melanie Viltard, Neal Paragas, Carrie J Shawber, Wenqiang Yu, Peter Lee, Xia Chen, Abby Sarkar, Weiyi Mu, Alexander Rittenberg, Chyuan-Sheng Lin, Jan Kitajewski, Qais Al-Awqati, Jonathan Barasch. Transcription factor TFCP2L1 patterns cells in the mouse kidney collecting ducts. *Elife* **6** (2017).

139 Aguilar, A. Development: Tfcp2l1 drives Notch signalling and epithelial diversity in the collecting duct. *Nature reviews. Nephrology* **13**, 445 (2017).

140 Landemaine, T. *et al.* A six-gene signature predicting breast cancer lung metastasis. *Cancer Res* **68**, 6092-6099 (2008).

141 Culhane, A. C. & Quackenbush, J. Confounding effects in "A six-gene signature predicting breast cancer lung metastasis". *Cancer Res* **69**, 7480-7485 (2009).

142 Otto, B. *et al.* Transcription factors link mouse WAP-T mammary tumors with human breast cancer. *International journal of cancer* **132**, 1311-1322 (2013).

143 Cai, S. Y. *et al.* Gene expression profiling of ovarian carcinomas and prognostic analysis of outcome. *Journal of ovarian research* **8**, 50 (2015).

144 Kim, H. S. *et al.* Microarray Analysis of Papillary Thyroid Cancers in Korean. *The Korean Journal of Internal Medicine* **25**, 399 (2010).

145 Heo, J. *et al.* Phosphorylation of TFCP2L1 by CDK1 is required for stem cell pluripotency and bladder carcinogenesis. *EMBO molecular medicine* **12**, e10880 (2020).

146 Parekh, V. *et al.* Defective extraembryonic angiogenesis in mice lacking LBP-1a, a member of the grainyhead family of transcription factors. *Mol Cell Biol* **24**, 7113-7129, doi:10.1128/MCB.24.16.7113-7129.2004 (2004).

147 Kang, H. C. *et al.* Erythroid cell-specific alpha-globin gene regulation by the CP2 transcription factor family. *Mol Cell Biol* **25**, 6005-6020 (2005).

148 Wang, X. *et al.* The transcription factor TFCP2L1 induces expression of distinct target genes and promotes self-renewal of mouse and human embryonic stem cells. *The Journal of biological chemistry* **294**, 6007-6016 (2019).

149 Sambrook, J. & Russell, D. W. *The condensed protocols from molecular cloning: a laboratory manual*. (2006).

150 Hilgarth, R. S. & Lanigan, T. M. Optimization of overlap extension PCR for efficient transgene construction. *MethodsX* **7**, 100759 (2020).

151 Edelheit, O., Hanukoglu, A. & Hanukoglu, I. Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure-function studies. *BMC biotechnology* **9**, 1-8 (2009).

152 Zheng, L., Baumann, U. & Reymond, J.-L. An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic acids research* **32**, e115-e115 (2004).

153 Cox, M. M. & Nelson, D. L. *Lehninger principles of biochemistry*. (Wh Freeman, 2008).

154 Ehresmann, B., Imbault, P. & Well, J. Spectrophotometric determination of protein concentration in cell extracts containing tRNA's and rRNA's. *Analytical biochemistry* **54**, 454-463 (1973).

155 Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein science* **4**, 2411-2423 (1995).

156 Wilfinger, W., Mackey, K. & Chomczynski, P. in *Assessing the Quantity, Purity and Integrity of RNA and DNA Following Nucleic Acid Purification* 291-312 (MA: Jones and Bartlett Publishers, 2006).

157 Liu, Z.-Q., Mahmood, T. & Yang, P.-C. Western blot: technique, theory and trouble shooting. *North American journal of medical sciences* **6**, 160 (2014).

158 Huynh, K. & Partch, C. L. Analysis of protein stability and ligand interactions by thermal shift assay. *Current protocols in protein science* **79**, 28.29. 21-28.29. 14 (2015).

159 Webster, J. & Oxley, D. in *Chemical Genomics and Proteomics* 227-240 (Springer, 2012).

160 Schmitz, K. S. *Introduction to dynamic light scattering by macromolecules*. (Elsevier, 2012).

161 Pierce, M. M., Raman, C. & Nall, B. T. Isothermal titration calorimetry of protein–protein interactions. *Methods* **19**, 213-221 (1999).

162 Heinemann, U., Büssow, K., Mueller, U. & Umbach, P. Facilities and methods for the high-throughput crystal structural analysis of human proteins. *Accounts of Chemical Research* **36**, 157-163 (2003).

163 Leslie, A. G. The integration of macromolecular diffraction data. *Acta Crystallographica Section D: Biological Crystallography* **62**, 48-57 (2006).

164 Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallographica Section D: Biological Crystallography* **67**, 271-281 (2011).

165 Krug, M., Weiss, M. S., Heinemann, U. & Mueller, U. XDSAPP: a graphical user interface for the convenient processing of diffraction data using XDS. *Journal of Applied Crystallography* **45**, 568-572 (2012).

166 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography* **66**, 213-221 (2010).

167 Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta crystallographica section D: biological crystallography* **60**, 2126-2132 (2004).

168 Project, C. C. The CCP4 suite: programs for protein crystallography. *Acta crystallographica. Section D, Biological crystallography* **50**, 760-763 (1994).

169 Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica Section D: Biological Crystallography* **58**, 1948-1954 (2002).

170 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66**, 12-21 (2010).

171 Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research* **35**, W375-W383 (2007).

172 Smirnova, E. *et al.* A new mode of SAM domain mediated oligomerization observed in the CASKIN2 neuronal scaffolding protein. *Cell Communication and Signaling* **14**, 1-14 (2016).

173 Knight, M. J., Leettola, C., Gingery, M., Li, H. & Bowie, J. U. A human sterile alpha motif domain polymerizome. *Protein Science* **20**, 1697-1706 (2011).

174 DaRosa, P. A., Ovchinnikov, S., Xu, W. & Klevit, R. E. Structural insights into SAM domain-mediated tankyrase oligomerization. *Protein Science* **25**, 1744-1752 (2016).

175 Kim, C. M., Jang, T.-h. & Park, H. H. Functional analysis of CP2-like domain and SAM-like domain in Tfcp2l1, novel pluripotency factor of embryonic stem cells. *Applied biochemistry and biotechnology* **179**, 650-658 (2016).

176 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* **46**, W296-W303 (2018).

177 Becker, M., Kissick, D. J. & Ogata, C. M. Locating and Visualizing Crystals for X-Ray Diffraction Experiments. *Protein Crystallography*, 143-164 (2017).

178 Broecker, J. *et al.* High-throughput in situ X-ray screening of and data collection from protein crystals at room temperature and under cryogenic conditions. *Nature protocols* **13**, 260-292 (2018).

179 Martiel, I., Olieric, V., Caffrey, M. & Wang, M. Practical Approaches forIn SituX-ray Crystallography: from High-throughput Screening to Serial Data Collection. (2018).

180 Maity, A., Singh, A. & Singh, N. Denaturation of DNA at high salt concentrations. (2015).

181 Butcher, S., Hainaut, P. & Milner, J. Increased salt concentration reversibly destabilizes p53 quaternary structure and sequence-specific DNA binding. *Biochemical Journal* **298**, 513-516 (1994).

182 Li, S., Olson, W. K. & Lu, X.-J. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic acids research* **47**, W26-W34 (2019).

183 Emamzadah, S., Tropia, L. & Halazonetis, T. D. Crystal structure of a multidomain human p53 tetramer bound to the natural CDKN1A (p21) p53-response element. *Molecular Cancer Research* **9**, 1493-1499 (2011).

184 Siggers, T. & Gordaˆn, R. Protein–DNA binding: complexities and multi-protein codes. *Nucleic acids research* **42**, 2099-2111 (2014).

185 Klämbt, V. *et al.* Mutations in transcription factor CP2-like 1 may cause a novel syndrome with distal renal tubulopathy in humans. *Nephrology Dialysis Transplantation* **36**, 237-246 (2021).

186 Zhao, Y. *et al.* A Feedback Loop Comprising EGF/TGFα Sustains TFCP2-Mediated Breast Cancer Progression. *Cancer research* **80**, 2217-2229 (2020).

187 Ali, R. & Wendt, M. K. The paradoxical functions of EGFR during breast cancer progression. *Signal transduction and targeted therapy* **2**, 1-7 (2017).

188 Saxena, U. H. *et al.* Phosphorylation by cyclin C/cyclin-dependent kinase 2 following mitogenic stimulation of murine fibroblasts inhibits transcriptional activity of LSF during G1 progression. *Molecular and Cellular Biology* **29**, 2335-2345 (2009).

189 Saxena, U. H., Owens, L., Graham, J. R., Cooper, G. M. & Hansen, U. Prolyl isomerase Pin1 regulates transcription factor LSF (TFCP2) by facilitating dephosphorylation at two serine-proline motifs. *Journal of Biological Chemistry* **285**, 31139-31147 (2010).

190 Fiumara, F., Fioriti, L., Kandel, E. R. & Hendrickson, W. A. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell* **143**, 1121-1135 (2010).

191 Lee, J. K. *et al.* Sex-specific effects of the Huntington gene on normal neurodevelopment. *Journal of neuroscience research* **95**, 398-408 (2017).

192 Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research* **43**, W174-W181 (2015).

7. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all the people who helped and supported me during the four precious years that passed in Berlin.

First of all, I want to thank my Ph.D. supervisor Prof. Dr. Udo Heinemann. I appreciate the opportunity as the last student to study in the lab. With your supervision, I learned a lot and made enormous progress. You always give me enough space to think about the project and encourage me to try different ways to figure out the problem and don't give up. Meanwhile, I could quickly reach to you to discuss the project when I had a problem. Besides, you put a lot of efforts into correcting my thesis and guide me to make it better. It's so lucky to have a supervisor of you.

Furthermore, I want to thank my secondary supervisor Prof. Dr. Oliver Daumke. You are so kind and patient with me. Thanks for your advice in my research project and for sharing the facilities in your lab.

Additional, I want to thank all lab mates of the Heinemann group and the Daumke group. Foremost, I would like to express my special thanks to Dr. Yvette Roske. You guided me in crystallography, from setting up the crystallization plate to crystal diffraction to the structure determination. Thanks to Maria, you made my Ph.D. life not too lonely at the last moment. Thanks to Sasa and Martin, you are so kind and warmhearted to care about my life and study even though you left the lab. Thanks to Anja, Ankur, and Qianqian for all discussions and fun we had. Thanks to Birgit, you made my life in the lab much easier. Thanks to my students Yanming, Georgia, and Viktória for working with me. Thanks to Nancy and Vivi for the technical assistance. Thanks to Kathrin, it's really a good experience playing Ping-pong with you. Thanks to Katja, Stephen, Elena, Saif, Thiemo, and Ferdinand from Daumke group for your help in my experiments.

Last but not least, I want to express my great appreciation to my family. Thanks to my parents for your support and understanding. Thanks to my two elder sisters for taking care of our family. Thanks to my wife Xiaoyan, it is you who accompany me for three years in Berlin. It is you who made my life in Berlin more colourful.