Characterization of transcription factor networks upon FOXA1 overexpression in prostate cancer

Mohamed Abdelhalim

Master thesis at the Department of Biosciences

Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

April 21st, 2020

Characterization of transcription factor networks upon FOXA1 overexpression in prostate

cancer

Mohamed Abdelhalim

Acknowledgements

My ultimate thanks go to the almighty Allah (God) for His great mercies and for giving me the opportunity to work in this project and to meet a lot of wonderful people. I express my deepest gratitude to my main supervisors Prof. Ragnhild Eskeland and my co-supervisor Dr. Marie Rogne for their excellent supervision throughout my master thesis. They have always been ready to support, encourage, and guide me during this work.

I would like to thank Trung Tran and Ignacio Cuervo for their contribution to this project. Thanks to my college at Chromatin Biology group, thank you all for encouragement, good social atmosphere, and moral support. I am also thankful to Prof. Philippe Collas for giving me the opportunity to work in my thesis, while I'm a member in his group

Finally and most importantly, I thank my wife Eman for her patience, kindness, care and love. My beautiful twins (Ziad and Lara), they are really great gift from God. I would like to extend my deepest gratitude to my family in Egypt especially my parents and my parents-in-low for their priceless support and encouragements. Without this support my work will not finish in such away. So thank you again and God bless you all.

Mohamed Abdelhalim

April - 2020

Summary

Transcription factors (TFs) are an essential component of biological pathways in all living cell. The gene expression regulation process is controlled by a complex network of interactions between different TFs and cis-regulatory regions in the genome. A key to understand the transcriptional regulation by TFs is to identify their DNA binding features and how post translational modifications can regulate the TFs affinity to chromatin. Post-translational modifications such as SUMOylation can have an impact on the transcription function of TFs and their target binding sites. However, how K6 SUMOylation of the pioneer TF Forkhead Box Protein A1 (FOXA1) could affect its TF function or affinity for specific gene target sites remains unknown.

In the present study we used an enrichment method of concatenated tandem array of consensus TF response elements (catTFREs) for over 390 TFs to enrich for TF proteins from prostate cancer cells. To assess the role of FOXA1 K6 SUMOylation on TF function and substrate specificity stable DU145 cell lines overexpressing 3xTy FOXA1, the SUMOylation deficient 3xTy FOXA1-K6R mutant or control 3xTy were established. We show that 3xTy FOXA1 and 3xTy FOXA1-K6R bind to TFRE constructs. Intriguingly, the SUMOylation deficient 3xTY FOXA1-K6R mutant displayed higher affinity to the DNA under high salt conditions. The specificity of the catTFRE system were assessed using the insulator protein CCCTC-binding factor (CTCF) that has binding sites uniquely present in one of the TFRE constructs. Specific DNA binding of CTCF was observed at high salt concentrations and interestingly we observed higher CTCF binding affinity to the catTFRE system in cell lines overexpressing 3xTy FOXA1 than cells overexpressing the SUMOylation deficient 3xTy FOXA1-K6R. Global RNA-seq analysis showed upregulation of important cancer related genes in 3xTy FOXA1 and 3xTy FOXA1-K6R compared to the 3xTy control cells. The analysis of differentially expressed genes showed a higher number of differentially expressed genes in the SUMOylation mutant cells compared to 3xTy FOXA1 overexpressed cells. Collectively, the findings indicate that that K6R SUMOylation of FOXA1 could have a role its TF substrate specificity in prostate cancer cells.

Table of Contents

1. Introduction

1.1. Regulation of eukaryotic gene expression

All cells in multicellular organisms share the same genetic information, and a specific cell type arises as cell lineage specific signalling and transcriptional programs are initiated. Gene expression in eukaryote organisms are tightly controlled at the level of transcription.

The gene expression in eukaryotic can be divided into two types, constitutive and inducible. The constitutive genes, often referred to as housekeeping genes, are expressed in all cells. The inducible genes, are cell or tissue type specific and becomes induced under particular conditions and are often under tight spatiotemporal control (Thomas et al. 2006). Transcription initiation is controlled by a group of proteins called transcription factors (TFs) that have the ability to recognize and bind to specific DNA regulatory sequences in order to regulate the activity of RNA polymerase at the binding site (Pan et al. 2010). The gene expression in various cells of multicellular organisms is regulated by the collective equilibrium of numerous different TFs (Pan et al. 2010). Moreover, additional levels of complexity are introduced trough the packaging of DNA into higher order chromatin and its post translational modification as illustrated in figure 1.1 (Nicolas et al. 2017).
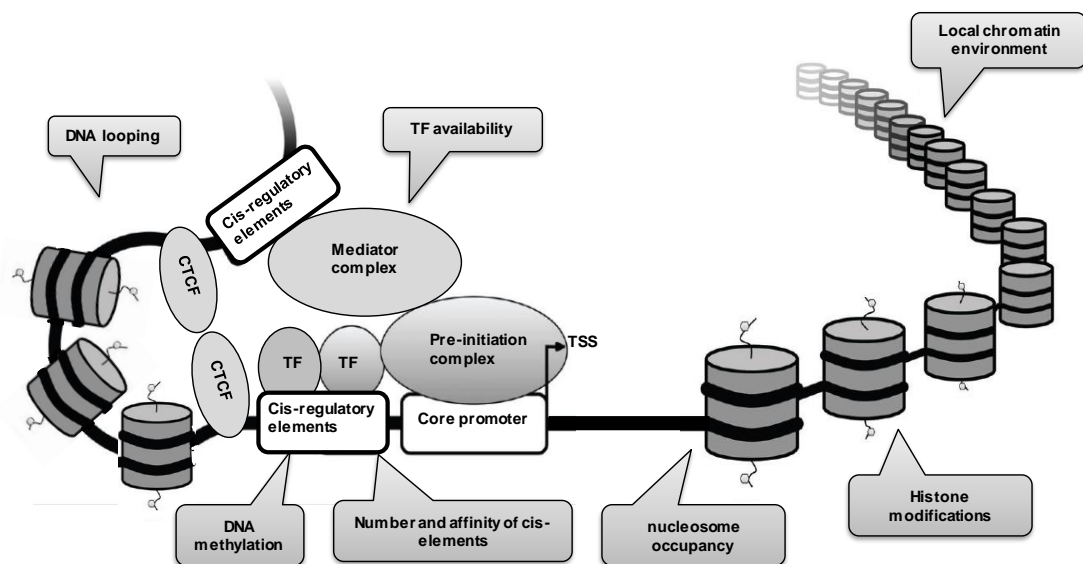
Figure-1.1. Molecular mechanisms regulating gene expression. Illustration of the cis-regulatory elements and different proteins associated with the control of gene transcription, including DNA (black line), transcription start site (TSS), transcription factor (TF) and nucleosomes. Text boxes indicate the specific molecular mechanisms that contribute to transcriptional regulation. (Adapted from Nicolas et al. 2017).

The human genome consists of three billion base pairs (bp) of DNA is separated into 22 different chromosomes (Venter et al. 2001) and measures about 1.5 cm if stretched out (Cooper et al. 2000). DNA compaction into the nucleus is achieved by a set of proteins called histones that consecutively coil and fold the DNA into higher levels of organization. The nucleosome is the first level of chromatin compaction. Each nucleosome consists of an octamer with two histone molecules each of histone 2A (H2A), histone 2B (H2B), histone 3 (H3), and histone 4 (H4) wrapped in 147 bp of double-stranded DNA (Mariño-Ramírez et al. 2005). Consequently, a typical diploid human cell contains nearly 30 million nucleosomes, which converts it's DNA into a chromatin thread about one-third of its initial length (Gao et al. 2020). Each of the core histones has N-terminal and C- terminal amino acid "tail", which are subject to post translation modifications such as methylation, phosphorylation and acetylation, that is associated with chromatin states such as gene expression or heterochromatin (Creyghton et al. 2010; Jenuwein et al. 2001). Regulation of gene expression involves processes to enhance or reduce the production of RNA from a specific gene or set of genes. Complex DNA-protein interactions are important to activate developmental pathways or to respond to environmental conditions (Jaenisch et al. 2003).

1.1.1 Dysregulation of gene expression

The human body contains hundreds of well-differentiated cell types with unique transcription profile. Many of these differences in the transcription profile happen through cell differentiation and remain through mitosis (Jaenisch et al. 2003). The stable transcription alterations to ensure the somatic inheritance of differentiated states are defined as epigenetic alterations, which consists of specific changes to chromatin components including DNA, RNA and proteins (such as histones) but do not involve underlying mutations of the DNA (Jones et al. 2016). These modifications are written by set of enzymes (writers) and are recognized by another set of enzymes (readers) that can modify specific genomic regions to moderate gene expression (Chen et al. 2018b). These active and repressive marks can be removed by other enzymes (erasers), which play a major role in cell plasticity (Chen et al. 2018b).

Transcription is regulated by a multi-level hierarchy network that includes transcription factor and promoter or enhancer interaction, DNA methylation, microRNA-modification and post-translational modification (Jones et al. 2016). Moreover, DNA accessibility controlled by

nucleosome positioning plays an important role in how regulatory elements function (Klemm et al. 2019) (Figure 1.1). Promoters are generally accessibility regions, while enhancer generally has lower DNA accessibility (Andersson et al. 2019). Therefore, the chromatin landscape can control enhancer activity by working as a guide system for gene transcription in a cell type-specific manner (Criscione et al. 2016). Specific histone modifications mark enhancer region such as histone 3 lysine 4 mono- and dimethylation (H3K4me1/me2), while histone 3 lysine 9 demethylation (H3K9me2) is found in inactive regions (Benayoun et al. 2011). Several studies have confirmed the deleterious effect once transcription factors become deactivated or activated in a dysfunctional way, which cause several cellular malfunctions, instability and can trigger tumorigenesis (Darnell 2002).

1.2. Cancer development

Cancer is a disease caused by the abnormal division of particular cells in the tissue independent of the presence of growth factors or other signals. These cells can invade tissues nearby or move to other parts of the body through the circulation or the lymph system. The term cancer spans over more than 100 different diseases (Hassanpour et al. 2018). Alterations in two main types of genes are responsible for cancer development, the activation of oncogenes and/or deactivation of tumour suppressor genes. These alterations can lead to unrestricted cell cycle progression and cell proliferation (Vogelstein et al. 2004). The difference between benign tumours and malignant cancers is that the second are able to metastasize (Papaccio et al. 2017). The occurrence of metastatic cancer requires the suppression of the cell adhesion receptors, which is necessary for cell to cell attachment in the tissue, and the activation of receptors that increase cell motility (Sarkar et al. 2013). These genetic and cellular changes occur thought different mechanisms such as mutations, dysregulation of signalling pathways by epigenetic changes, chromosomal translocations or deletions (Mansoori et al. 2017). Cancer is a complicated disease that involves different cell types with a heterotypic interaction (Hanahan et al. 2011). Therefore, cancer development have been divided into six hallmarks that describe the biological transformations that develop during the multistep development and growth of cancer tumours (Sarkar et al. 2013). These hallmarks are sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis as illustrated in figure. 1.2 (Hanahan et al. 2011)
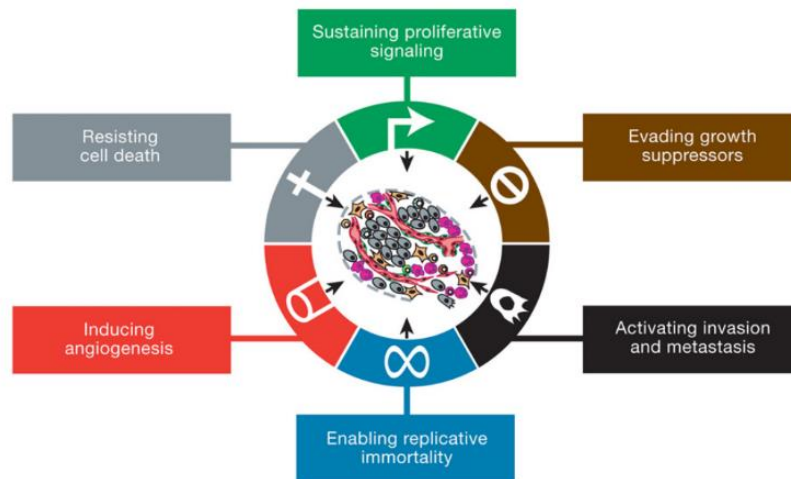
Figure 1.2. The Hallmarks of Cancer. This illustration of the six hallmark of cancer development. (from Hanahan et al. 2011).

1.2.1 Prostate cancer

Prostate cancer remains one of the major causes of cancer-related mortality and morbidity in men at present, with 1,600,000 estimated cases and 366,000 deaths annually (Wang et al. 2018a). The incidence of prostate cancer is the highest in developed countries in North America, Western and Northern Europe (Figure. 1.3) (Rebbeck 2017). However, the mortality of prostate cancer is highest in Africa particularly in sub-Saharan countries with rates ranging from 18.7 to 29.3 deaths per 100,000 populations (Rebbeck 2017). In Norway death rates due to prostate cancer is one of the highest in the world (Chen et al. 2018a). Each year about 28 % of all new diagnosed male cancers is prostate cancer and since 1950 the incidence rate has tripled (Aksnessæther et al. 2019).
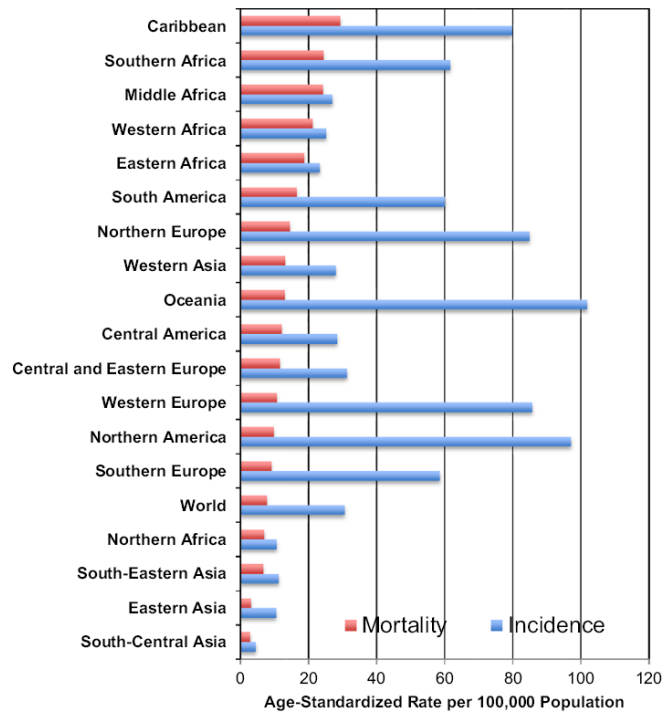
Figure 1.3. Global prostate cancer incidence and mortality by world region. (Adapted from Rebbeck 2017).

Adenocarcinoma is the most common type of prostate cancer and caused by malignancies of epithelial tissue similar to breast and colon cancer (Shen et al. 2010). The peripheral zone of the prostate is the source of approximately 60 to 75 % of prostate cancer (Wang et al. 2018a). Prostate cancer can be distinguished by its association with age since the clinically detectable prostate cancer is not usually revealed until the age of 60 or 70 (Shen et al. 2010). Hereditary factors are only found for about ten present of diagnosed prostate cancers and is usually linked to the early disease development (Shen et al. 2010). Genome wide analysis from over a thousand prostate cancer patients have shown that mutations in genes encoding epigenetic machinery components are found in approximate 15 – 20 % of prostate cancers patients (Yegnasubramanian et al. 2019).

The development of malignant prostate cancer follows a multistep process, starting as prostatic intraepithelial neoplasia (PIN), then localized prostate cancer followed by increasing local invasion of adenocarcinoma and with more progression the disease reach the metastatic stage (Figure 1.4) (Wang et al. 2018a).
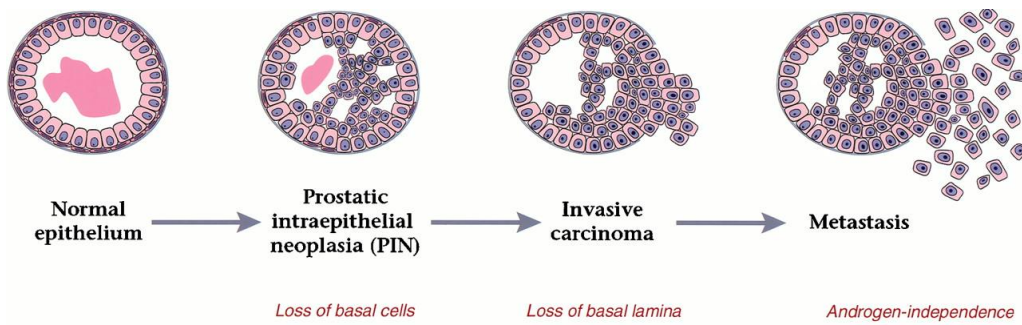
Figure 1.4. Stages of prostate cancer progression. Adapted from Shen et al. (2010).

The level prostate-specific antigen (PSA) in the blood is considered the main diagnosis of prostate cancer over the past few decades (Sumanasuriya et al. 2018). PSA is produced in normal prostate secretions, but becomes released into the blood as an effect of disrupted prostate function. Patients with elevated PSA levels usually take a prostate biopsy to evaluate the possible occurrence of prostate cancer (Sumanasuriya et al. 2018). The histopathological grading of prostate cancer was originally defined by Gleason scoring based on histological patterns of prostate adenocarcinoma (Wang et al. 2018a).

Hormone responsiveness is an important feature of prostate cancer. In prostate cancer the androgen receptor (AR) and its coregulators play very important role, with 80–90% of prostate cancers being dependent on androgen signal (Davey et al. 2016).In normal prostate, testosterone and 5α-dihydrotestosterone binds to AR to induce a conformational change that dissociate the chaperone proteins and expose its nuclear localisation signal (NLS), which indorse the association of AR with its coregulators (Davey et al. 2016). These Coregulators bind to the activated AR in a ligand-dependent manner to either enhance (co-activator) or repress (corepressor) its activity (Fujita et al. 2019). Following, AR becomes translocated to the nucleus and binds to androgen response elements (AREs), which represent the promoter regions of target genes to induce cell proliferation and apoptosis (Davey et al. 2016). The androgen/AR complex can also signal through non-DNA binding-dependent pathways, such as those involving transforming growth factor beta (TGF-β), interleukin 6 (IL-6), and insulin-like growth factor 1 (IGF-1) (Berger et al. 2018). The use of agents that block the androgen pathway (androgen deprivation therapy ADT) is considered the standard treatment for prostate cancer (Wang et al. 2018a). ADT resistance can cause the development of castration-resistant prostate cancer (CRPC) or metastatic CRPC (mCRPC) (Wang et al. 2018a).

1.3. Transcription factors

Transcription factors TFs control gene expression by binding to cis-acting regulatory DNA sequences. The transcription specificity is based on the interaction between different regulatory elements such as enhancers, promoters and insulators. An enhancer is a regulatory element that enhance the transcription its target gene(s) over a distance in an orientation-independent manner (Thomas et al. 2006), while promoter is a regulatory element found in close proximity to the Transcription Start Site (TSS) and play important role in recruiting the transcription machinery (Chronis et al. 2017). An insulator is a regulatory element that insulates chromatin domains by assisting the formation of chromatin looping and prevent the spread of epigenetic modification (Rowley et al. 2017).

TFs are a large family of proteins with approximately 2000 to 3000 members (Wingender et al. 2015) that recognize and bind to specific DNA sequences. TFs are characterized by two main functions: first, is to recognize specific DNA regulatory sequences, and second is to regulate gene expression by recruiting the transcription machinery together with co-activators or co-repressors that can alter the chromatin states (Venters et al. 2009). Identification of TFs binding site is important to define the function of these TFs. TFs DNA-binding site are often represented as A position weight matrix (PWM), which is usually short sequence (6 to 12 bases) preferred by a given TF and can be used to scan longer sequences such as promoters to recognize possible binding sites (Lambert et al. 2018).

The three dimensional crystallography of protein–DNA complex structures have enabled a better understanding of the mechanisms that governs the specific interaction (Harteis et al. 2014). The specific protein-DNA recognition known as 'base readout' is established by the physical contact between the side chains of the TF and the nearby ends of the DNA base pairs (Figure 1.5A). These interactions involve direct hydrogen bonds, hydrophobic contacts and water-mediated hydrogen bonds (Slattery et al. 2014). Another form of protein-DNA recognition is known as 'shape readout', which depend on the dynamic properties of the DNA structure in the major and minor groves such as the negative electrostatic capacity between DNA and arginine or histidine residues (Figure 1.2B) (Harteis et al. 2014).
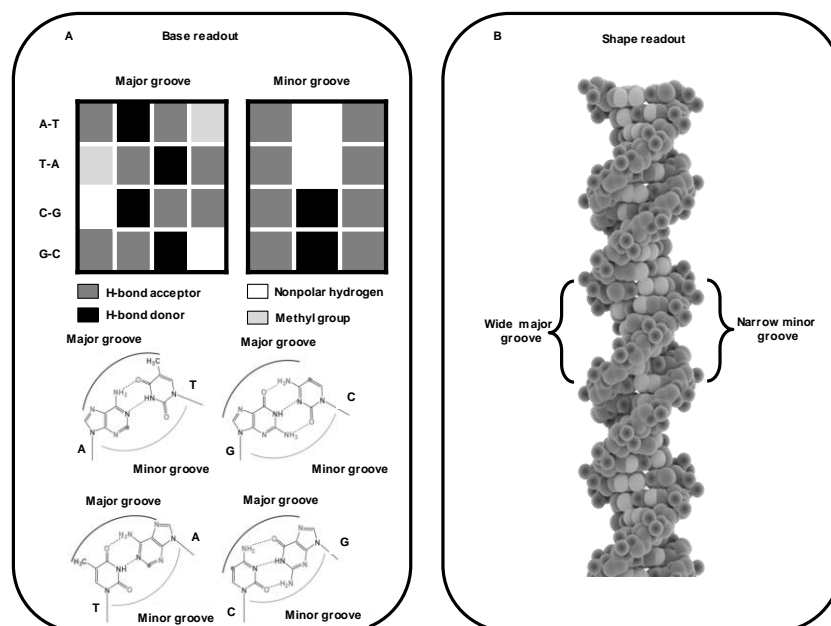
Figure 1.5. Base and shape readout contribute to TF–DNA binding specificity. (A) Base readout defines the direct connections between amino acids of a TF and the functional groups of the DNA nucleotides. The hydrogen bond acceptors (grey) and donors (black), heterocyclic hydrogen atoms (white) and the hydrophobic methyl group (light grey) is base pair-specific in the major groove, the pattern is degenerate in the minor groove. (B) Shape readout includes any form of structural readout based on global and local DNA topographies, including conformational flexibility and shape-dependent electrostatic potential. Adapted from Harteis et al. (2014); Slattery et al. (2014).

Closely related TFs binds to distinct transcription factor binding sites to perform different functions. The mechanisms by which paralogous TFs select very similar, but not identical, target sites are not fully understood (Shen et al. 2018). This specific binding generates transcriptional regulatory signals that regulate the transcription of DNA by RNA polymerase II to fine-tune spatiotemporal gene expression (Vaquerizas et al. 2009). TFs can be grouped into two broad categories; "general" TFs that recruit the basal transcriptional machinery around RNA polymerases and "specific" TFs that regulate target genes by binding to their regulatory cis-element(s) to activate or repress their transcription in response to different biological signals (Benayoun et al. 2011). The specific TFs share common features, such as the presence of a DNA binding domain (DBD) and a trans-activation domain (TAD) (Arnold et al. 2018). The recruitment of TFs to their binding sites is also regulated by the chromatin state, which includes DNA methylation, nucleosomes distribution, histone modifications, chromatin folding and three-dimensional chromatin organization (Klemm et al. 2019). Modification of TFs on the protein level can work as "molecular switch" that can carry the effect of upstream cellular signals, in response to different environmental conditions or other stimulus, to the downstream target genes (Everett et al. 2009).

1.3.1 Posttranslational modifications regulate TF binding activity

Posttranslational modifications (PTMs) are modifications of a protein side chains or backbone generated by catalytic enzymes. The catalytic enzyme responsible for the PTMs can be classified to two general categories based on their activity. The enzymes that introduce new chemical group (usually electron-rich) to the protein side chain and the enzymes that promote a break of the protein backbone by proteolytic enzymes (Walsh et al. 2005).

TFs can be modified by ATP-dependent phosphorylation, acetylation, methylation, ubiquitination and SUMOylation (Csizmok et al. 2018; Müller 2017). These PTMs can work as a "molecular barcode" to define a specific function of the TFs and play an important role in TFs ability to bind to DNA, be active and to interact with other regulatory elements. PTMs of TFs can control the spatiotemporal gene expression to safeguard the gene activation patterns of specific tissues at specific time points during development (Benayoun et al. 2009). The intricate combinations of various PTMs of the same TF and their cross-talk collectively govern the resulting TFs activity. Adding to the complexity, sequential PTMs of TFs can have antagonistic effect, where newly added PTMs can cancel the effect of the previous PTMs (Filtz et al. 2014).

The various PTMs with diverse properties have different effect on the TFs. In general phosphorylation is dynamic and reversible, and as such can work as a temporal regulatory modification. Phosphorylation of TFs can affect their cellular localization, DNA interaction and stability which can affect their target gene(s) regulation (Filtz et al. 2014). Methylation of TFs usually occurs at arginine residues and can affect DNA-binding affinity, protein–protein interactions, and crosstalk with other PTMs (Han et al. 2019). Polyubiquitination is usually linked to protein degradation. However, monoubiquitination be linked to protein activation and intracellular transport (Filtz et al. 2014). Ubiquitination is an important modification to control the levels of essential TFs such as tumor Protein P53 (TP53), MYC proto-oncogene and E2F in a proteolysis dependent manner (Muratani et al. 2003). This process enables a rapid response to specific signals and to keep the appropriate cell type specific gene expression profile (Muratani et al. 2003).

1.3.2. Roles for SUMOylation in transcriptional regulation

Small Ubiquitin-like Modifier proteins (SUMO) is a family of conserved proteins in eukaryotic organisms and share resemblance with ubiquitin. Conjugation of SUMO to lysine(s) residues of target proteins is carried out by SUMO E1, E2, and E3 enzymes (Johnson 2004). The E1 enzyme attaches mature SUMO to its active-site cysteine using ATP hydrolysis. Then SUMO is transferred to an E2 (UBC9) that has the ability to transfer SUMO onto a Lys residue of a target protein by the help of the E3 SUMO-ligase (Figure 1.6). E2 SUMO recognises a consensus sequence for SUMO conjugation "cKxE/D" where c is a hydrophobic residue; x is any amino acid and K is the sumoylation site (Zhao 2018).
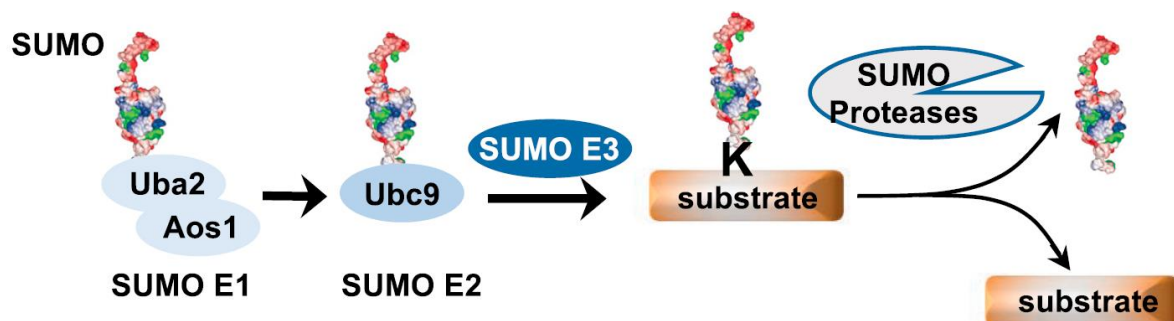
Figure 1.6. The outline of SUMO conjugation and deconjugation. SUMO is conjugated with E1 heterodimer (Aos1/Uba2) by a thioester bond in an ATP-dependent manner. Then SUMO is moved to the E2 (Ubc9) by thioester bond. SUMO is then transferred to the lysine residue (K) on the target protein via the help of SUMO E3. SUMOylation can reversed by SUMO removal by SUMO-specific proteases (SENPs). Adapted from Zhao (2018).

Most organisms have a single SUMO E1 and E2 enzyme and multiple isoforms of SUMO E3. Moreover, higher eukaryotes possess at least three SUMO isoforms known as SUMO1–3, whereas plants express eight different SUMO isoforms (Geiss-Friedlander et al. 2007). In humans, four SUMO isoforms are expressed, but SUMO2 and SUMO3 share higher sequence similarity and are often described together (Pichler et al. 2017). These isoforms vary in their expression levels, response to stress, SUMO E3 preferences and the capability to create a poly-SUMO chain by the conjugation of one SUMO molecule to another via different lysine residues (Pichler et al. 2017).

The effect of SUMOylation varies from altered protein function, interactions with DNA or RNA and subcellular localization (Zhao 2018). One of the most described effects is the interactions between SUMO and SUMO interaction motifs (SIMs) (Yang et al. 2017). SUMO-

SIM interaction can be enriched by the multiple binding of SIMs to SUMO chains. This interaction plays an important role in the effect of SUMO on gene expression and chromatin structure (Geiss-Friedlander et al. 2007). Protein SUMOylation and ubiquitination both affect lysine residue, though both can work together to regulate biochemical function or compete for the same target. However, in most cases SUMOylation and ubiquitination are competing for common lysine residues (Zhao 2018).

SUMOylation can affect a large number of TFs and other gene expression regulators (Gill 2005). SUMOylation frequently increases protein stability such as SUMOylation of Oct4 increase its stability and DNA binding during embryonic cell development (Yang et al. 2017). The SUMOylation of TFs, cofactors or chromatin remodelling factors, represent nearly half of SUMOylation targets proteins. These modifications can alter the transcriptional activity of these portions and control its signalling pathways, for example the steroid hormone receptor pathways which is very important to cancer progression (Lee et al. 2017). Knockdown of SUMO activating enzyme E1 or SUMO conjugating enzyme (E2) inhibits maintenance and self-renewal of colorectal cancer stem cells (Pichler et al. 2017). SUMOylation of tumour suppressors and oncogenes such as TP53, c-Jun and c-Myc have been reported (Bettermann et al. 2012). Most of these modifications are linked to suppressive effects and partial activation. However, SUMOylation can also be associated with activation of TFs activity such as T-cell factor- 4 (TCF-4), heat shock factor (HSF2) and TP53 (Hong et al. 2001; Yamamoto et al. 2003). SUMOylation can suppress TFs activity by several mechanisms such as increasing the interaction with repressors or disturbing the TFs acetylation or phosphorylation, which promote TFs activity (Rosonina et al. 2017). SUMO deconjugation can also cause a major change in the TFs gene-regulatory activity. The balance between SUMO conjugation and deconjugation of the TFs can control their activity and work as a molecular switch for the gene expression fine-tuning (Rosonina et al. 2017). Moreover, SUMOylation can interrupt the cooperative synergetic effect by interaction between TFs, which is important for the assembly of the transcription machinery as well as productive mRNA elongation. This effect has been described for transcriptional activator c-Myb, steroidogenic factor 1 (SF-1), zinc finger transcription factor ZBP-89 and Melanocyte Inducing Transcription Factor (MITF) (Chupreta et al. 2007; Molværsmyr et al. 2010).

Several SUMO pathways have been shown to be dysregulated in human cancers (Seeler et al. 2017). For example, the overexpression of SUMO E2 enzyme UBC9 is associated with

accelerated ovarian cancer progression, and the mutations in UBC9 is linked to breast cancer incidences (Mo et al. 2005). High expression of UBC9 is correlated with highly metastasizing and poor differentiated breast cancer subtypes, with poorer prognosis and lower response to chemotherapy (Rabellino et al. 2020). Mutations of the SUMO deconjugation enzymes SENP1 and SENP2, are also shown to associate with breast cancer occurrence (Mirecka et al. 2016). In many cancer types, SUMOylation is significantly upregulated, therefore SUMOylation may contribute to cancer cell survival and proliferation (Seeler et al. 2017).

1.4. Pioneer transcription factors

Pioneer transcription factors were first coined when it was discovered that they can bind to condensed chromatin and alter gene activity during liver development (Cirillo et al. 2002). Pioneer TFs have the important role of open closed chromatin domains during development to allow the implementation of new cellular programs. This can initiate the complete rewiring of a cell's gene-expression program and reprogram it into another cell type. Pioneer TFs therefore have positive and negative effects on gene expression (Mayran et al. 2018). Enabling this is their unique abilities to recognize, bind and open their target DNA sequences in compact or "closed" chromatin regions independently of other factors. Pioneer transcription factors trigger remodelling of the surrounding chromatin landscape to provide accessibility to non-pioneer transcription factors. Pioneer TFs help to facilitate the deposition of active epigenetic marks and prevent the repressive modifications at regulatory sites (Swinstead et al. 2016). Epigenetic modifications at active enhancers have been shown to be remodelled and maintained by pioneer TFs and in this way contribute to the establishment of specific transcriptional programs (Magnani et al. 2011; Sekiya et al. 2009). Pioneer TFs are essential for cell development and differentiation and often dysregulated in diseases such as cancer (Jiang et al. 2009; Magnani et al. 2011). Pioneer TFs may also work as 'readers' of epigenetic marks, and this can explain their interaction with the insulator protein CTCF (Iwafuchi-Doi et al. 2014). It was recently shown that CTCF influences the binding activity of the pioneer TF FOXA1 either by competing for the same binding sites or modulate the surrounding positions of the binding sites (Jung et al. 2019).

Cancer is usually associated with the alteration of gene expression and it is possible to distinguish cancer subtypes based on their transcriptional profiles (Hutter et al. 2018).

Alteration of the chromatin landscape plays an active role in the establishment of this dysregulated transcriptional program (Weinberg 2013). Pioneer TFs are the leading candidates to control these changes due to their ability unmasking closed chromatin domains during development to allow the implementation of new cellular programs (Magnani et al. 2011). Genes encoding pioneer TFs and their genomic activities are altered in several types of cancer (Magnani et al. 2011). These functional alterations can be caused by mutations, translocations or overexpression. Pioneer TFs have been shown to be overexpressed in different cancers such as breast, ovarian and prostate cancers (Bhagwat et al. 2015; Swinstead et al. 2016). The most studied pioneer TFs are octamer-binding transcription factor 4 (OCT4), SRY-Box Transcription Factor 2 (SOX2), Krüppel-like factor 4 (KLF4), and MYC. These factors are able to reprogram differentiated cells into induced pluripotent stem cells (Dobersch et al. 2019). The expression of these factors is linked with cancer development in intestine, skin, pancreas, stomach, gallbladder and kidney (Chronis et al. 2017).

1.4.1. The FOX TFs family

Forkhead box (FOX) proteins are members of an evolutionarily conserved family of transcription factors, that play an essential role during cell development, differentiation and proliferation processes such as cell cycle control, tissue homeostasis, ageing, stress tolerance and metabolism regulation (Sutinen et al. 2014). Forkhead proteins are part of the TFs superfamily "winged helix" according to the structural classification of proteins (SCOP) (Laissue 2019). The name forkhead was coined after the first discovery of these proteins in a mutant of *Drosophila melanogaster* called "*fork head*" by Weigel et al. (1989). At present, more than 2000 proteins have been identified as forkhead proteins in 108 species of fungi and animals (Benayoun et al. 2011; Laissue 2019). The number of genes in each species are diverse, with 16 genes in *Caenorhabditis elegans*, 18 in *D. melanogaster*, 49 in the zebrafish and 50 in humans (Kaestner et al. 2000). There are 19 human subfamilies (A to S) of FOX TFs (Laissue 2019). The distribution of FOX genes in the human genome is not random, 26 of the 50 FOX genes are arranged into nine genomic clusters (Wotton et al. 2006). All FOX members share a highly conserved ~ 100 residue DBD (FOX-DBD) that binds to a target core sequence (5′-G/A)(T/C)(A/C)AA(C/T)A-3′) (Benayoun et al. 2011). Sequences next to the core sequence are similarly important for TF differential functions and DNA affinity (Laissue 2019). So far, the Protein Data Bank-DBD (https://www.rcsb.org/) contains several FOX-DBD structures

such as FOXA1, FOXA2, FOXM1, FOXN1, FOXO1, FOXO2, FOXO4, FOXK1, FOXK2, FOXP1, FOXP2 and FOXP3 (Laissue 2019). Forkhead domain (FHD) contains three N-terminal α-helices (labelled as H1to H3), three β-strands (labelled as S1to S3) and two loops resembling butterfly wings or a "winged helix" (labelled as W1–2) towards its C-terminal region, in some FHD additional α-helix (H4) is found between H2 and H3 (Figure 1.7) (Benayoun et al. 2011; Laissue 2019; Li et al. 2017; Obsil et al. 2008).
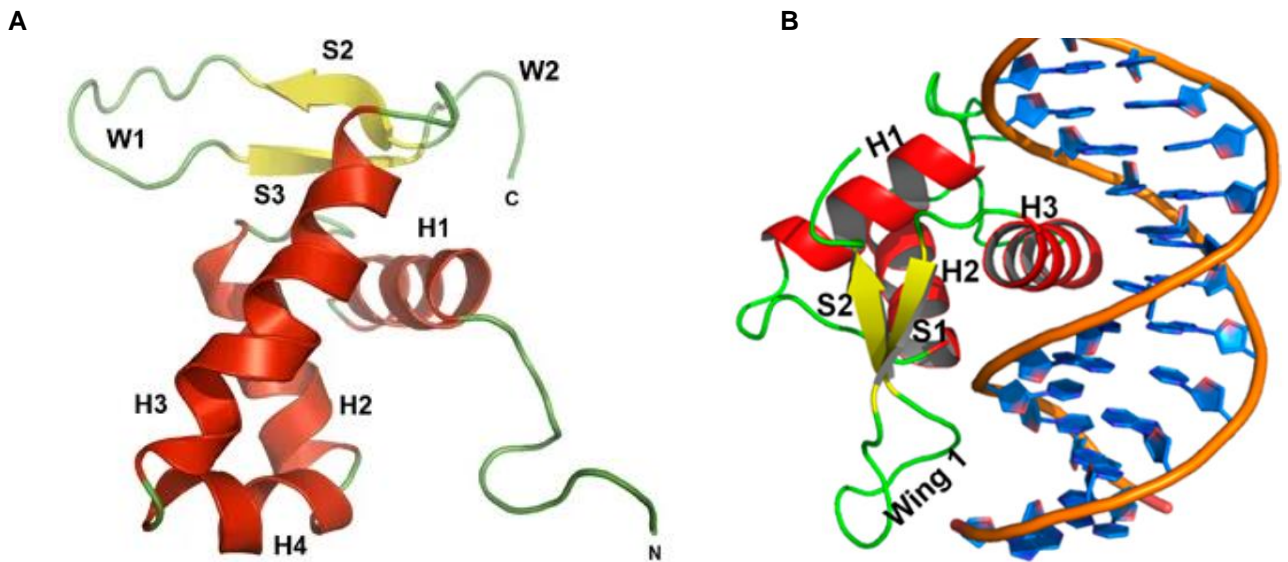
Figure 1.7. Ribbon representation of the solution structure of the FOXO4 (A) and FOXA2 bound to DNA (B). Secondary structure elements are labelled. Adapted from Li et al. (2017); Obsil et al. (2008).

The α-helix (H3) is considered the recognition helix that interacts with specific sequences and the major groove of DNA (Laissue 2019). Moreover, the junction of helices H2 and H3 and wings W1 and W2 play import role in DNA-binding specificity, which interacts with the minor groove of DNA (Obsil et al. 2008). According to Cirillo et al. (2007), the wing domains can regulate the DNA-binding affinity and specificity of FOXA1. FOX TFs are subject to several post-translational modifications, such as acetylation, phosphorylation, ubiquitination and SUMOylation, which can modulate DNA-binding affinity different and their molecular functions (Cirillo et al. 2007).

The presence of two NLS at both ends of the FHD support its function as nuclear receptor as documented for FOXF2, FOXA2, FOXE1 and FOXP3 (Benayoun et al. 2011; Carlsson et al. 2002). The highly conserved C-terminal NLS is involved in the shared mechanisms of the nuclear signal between different FOX TFs. However, the flanking regions that contain effector

domains such as transactivation and/or transrepression domain are poorly conserved (Laissue 2019). In most cases, FOX TFs tend to bind to DNA as monomers (Golson et al. 2016). Nevertheless, in some cases, FOXP2 and FOXK1 can bind to their target sequence as homodimers or heterodimers (Golson et al. 2016). FOX TFs usually work in collaboration with other TFs from other families such as SMAD3, STAT3 and HOXA5 (Zaret et al. 2016).

Dysregulation or mutation in the FOX TFs can cause genetic diseases, cancer progression or deregulated ageing (Golson et al. 2016). Germinal mutations in FOX genes can promote several diseases such as speech disorders and immunological defects (Benayoun et al. 2011). Several of the FOX TFs play an important role during embryonic development by controlling morphogenesis and differentiation and other functions such as stress response or carbohydrate and lipid metabolism in well-differentiated cells (Dobersch et al. 2019; Iwafuchi-Doi et al. 2016).

1.4.2. Pioneer transcription factor FOXA1

FOXA1 is a pioneer transcription factor with the ability to mediate transcription through chromatin rearrangements (Iwafuchi-Doi et al. 2016). These chromatin rearrangements leads to a more accessible chromatin, allowing other TFs to bind enhancers and modulate gene expression (Iwafuchi-Doi et al. 2016). FOXA1 plays an important role in androgen-depended prostate cancer by maintaining the proliferation of prostate cancer cells in absence of androgen through G2/M checkpoint activation via the activation of *UBE2C* gene expression (Dobersch et al. 2019). FOXA1 has been shown to decrease chromatin condensation in an ATP-independent manner and can promote *in vitro* nucleosomes reposition (Cirillo et al. 2002; Magnani et al. 2011). These abilities can in part be explained by the presence of the winged-helix motif that is homologous to histone H1, which allows FOXA1 to replace histone H1 and promote chromatin opening (Cirillo et al. 1998; Magnani et al. 2011). Several members of the FOXA family of TFs share the same biological function, which can be defined as 'gene sharing', which occur when a single protein performs separate functions according to the biological situations (Benayoun et al. 2011). For example, FOXA2 can control liver and pancreas development during embryogenesis, however, in later differentiated stages this protein controls insulin secretion and gluconeogenesis (Le Lay et al. 2010). The interaction between nuclear receptors such as estrogen (ER/Nr3a1), or androgen (AR/Nr3c4) and FOXA TFs facilitate the recruitment of these nuclear receptors to enhancers regions (Fournier et al. 2016). Moreover, FOXA TFs initiate the chromatin opening at these enhancers (Mayran et al.

2018). Posttranslational modifications can have important impact on the interaction of FOXA TFs and nuclear receptors (Tomasi et al. 2018) for example: nuclear exclusion of FOXA2 is promoted by insulin through AKT-mediated phosphorylation of FOXA2 threonine 156 and cells expressing the phosphorylation-deficient FOXA2-T156A are unresponsive to insulin (Choi et al. 2020). Also, acetylation of FOXA2 results in increased stability and transcriptional activity due to retention in the nucleus (Zhang et al. 2018b). In general, PTMs can control both DNA-protein and protein–protein interaction of FOXA TF by the phosphorylation of serine and tyrosine residues (Calnan et al. 2008) or by acetylation and deacetylation of lysines residues (Benayoun et al. 2009; Li et al. 2007). Moreover, serval members of FOXA TFs can undergo arginine methylation, serine/threonine O-GlcNAcylation and ubiquitination (Ho et al. 2010; van der Horst et al. 2006; Yamagata et al. 2008). All these modifications can affect FOX TFs cellular localization and Transcriptional activity. Phosphorylation of Tyr429 and Tyr464 in the FOXA1 C-terminal region promotes the activation of estrogen signalling by inducing its binding to histones (Yamaguchi et al. 2017). O-linked N-acetylgalactosamine (O-GalNAc) modification at the C-terminal of FOXA1 has been shown to reduce the protein stability, and thereby have an impact on the estrogen signalling transcription network (Liu et al. 2019; Zhang et al. 2019a). Moreover, PTMs can instruct the FOXA TFs response to a particular environmental signal such as for example, acetylation, which controls apoptosis or cell survival (Calnan et al. 2008). Acetylation of FOXA1 at K295 can also change its binding preference and genomic distribution (Li et al. 2019b). However, PTMs of FOX TFs can also facilitate the fine-tuning of its DNA binding efficiency and specificity to different DNA binging sites (Benayoun et al. 2011).

1.5. Methods to study transcriptional regulation

1.5.1. In vivo methods

The development of methods to study TFs action *in vivo* has been important to understand the role of TFs in the transcriptional regulation (Lambert et al. 2018). One of the challenges of such methods is that TFs from the same or different families can recognize a similar or overlapping DNA sequence and it is likely that multiple TFs can bind to a distinct regulatory element *in vitro* (Viola et al. 2016). Therefore, revealing the genome "transcriptional code" can be more difficult than its "genetic code" (Harbison et al. 2004). Transcription profiling can be applied to stable cell lines that overexpresses or has downregulated the level of specific TFs to

identify the global expression changes (Viola et al. 2016). One of the methods used to study global expressional changes is high throughput sequencing of RNA (RNA-seq).

The use of RNA-seq technology provides novel insight, not only by the quantification of gene expression, but also for the identification of the transcripts at single-nucleotide level without depending on prior annotation or prior knowledge regarding transcribed regions (Marguerat et al. 2010). The major challenges working with RNA-seq data is the sequence assembly and statistical model used for identifying differentially expressed genes. Due to the high usage demands of this technology, the data analysis methods develop continually (Egan et al. 2012). Therefore, several methods for detecting and analysing differentially expressed genes are available and constantly evaluated (Marguerat et al. 2010). The analysis method can be different according to the study goal if the interest in only messenger RNA, different splice isoforms, non-coding RNA or microRNA (miRNA) levels (Khang et al. 2015). Well-annotated human RNA-seq data is analysed based on the existing annotated reference genome, also novel transcripts can be identified using reference transcriptome. However, there is no optimized pipeline for different applications of RNA-seq analysis and the analysis strategies depending on the research goals and the organism being studied (Conesa et al. 2016).

In order to understand transcriptional regulation by TFs it is important to identify their target genes. Usually the gene expression analysis in cell lines that overexpresses or has downregulated the level of specific TFs presents the first clues to TFs function (Viola et al. 2016). Nevertheless, additional investigations are required to clarify the signal transduction cascades moderated by the TF (Viola et al. 2016). Several methods are used to identify the specific DNA binding sites by different TFs *in vivo*, such as chromatin immunoprecipitation (ChIP), cleavage under targets and release using nuclease (CUT&RUN) and DNA adenine methyltransferase identification (DamID) (Lambert et al. 2018). ChIP is dependent on specific antibodies for enrichment by immunoprecipitation to reveal the DNA sequences that the specific TF is associated with (Collas 2010). This is identified by either quantitative PCRs for specific loci of interest (ChIP-qPCRs) or next-generation sequencing to identify genome-wide binding sites (ChIP-Seq) (Wagner et al. 2016). This method has been extensively used to map the genomic locations of histone variants, histone modifications and chromatin associated proteins such as TFs (Collas 2010). The main limitations for ChIP-Seq method are the following: (1) the requirement of vast starting material (millions of cells), (2) sonication for fragmentation can destroy the epitope used for immunoprecipitation, (3) the ligation of sequencing adaptors can result in the loss of DNA material and low library complexity

(Gutiérrez et al. 2017; Mieczkowski et al. 2016). However, extensive work has been done to develop ChIP methods the use a fewer number of cells (Akhtar et al. 2019), also MNase digestion-based methods has been used to reduce the effect of sonication on studied materials (Lion et al. 2019).

RNA-seq can be coupled with protein's binding by ChIP-seq (Wade 2015). This integration increases our understanding of the interactions between the genome localization of TFs or histone modification and downstream target gene expression (Zhang et al. 2018a). Using a Bayesian network analysis histone modifications and gene expression were correlated in human CD4+ T-cells and only a small number of histone modifications were necessary to predict gene expression (Zhang et al. 2018a). The main challenges for the integration of RNA-seq and ChIP-seq output are: the effect of chromatin 3D structure, the influence of DNA methylation on transcription and the presence of unidentified TFs or histone modifications on distally encoded genes (Wade 2015).

1.5.2 In vitro methods (proteomic methods)

1.5.2.1 Labelling based methods

Many of the labelling-based methods use Stable Isotope Labelling of Amino Acids in Cell Culture (SILAC) to establish an expected mass shift among peptides from different experimental conditions using mass spectrometry (MS) (Simicevic et al. 2017). SILAC method use a chemically synthesized isotope-labelled peptides as standards to identify the amount of the endogenous protein existing within the biological sample (Ong 2012). By growing cells in two different conditions, one in a medium that contains a normal amino acids (light) and the other in a medium that contains a labelled amino acid (he avy) (Mann 2006). A limitation of this method is the cost of the isotope-labelled peptides and its short period of storage (Simicevic et al. 2017). These modifications allow all tryptic peptides generated from the protein to be monitored. SILAC can be used to identify TFs by incubating synthesized double-stranded DNA with nuclear extract (light vs. heavy) and TFs that binds to the DNA can be detected with MS as sequence-specific DNA-binding protein (Snider et al. 2019). SILAC can be also used to profile the chromatin dynamics and histone marks (Völker-Albert et al. 2018).

1.5.2.2 Label-free based methods

The main challenge for quantitation of TFs on a proteome level is the low abundance of these proteins (Wang et al. 2019). In general, the use of RNA-seq for mRNA profiling cannot be directly translated into TFs protein levels (Simicevic et al. 2017). Recently, significant effort has been dedicated to combine the benefits of label-free analyses with the sensitivity of targeted mass spectrometry (MS) approaches to quantify TFs protein levels (Simicevic et al. 2017). In this chapter some of these methods will be discussed.

Systematic evolution of ligands by exponential enrichment (SELEX)

SELEX is important method to define TFs DNA-binding sequence specificities *in vitro* (Darmostuk et al. 2015). The method was first described more than twenty years ago (Tuerk et al. 1990). Since then the method has undergone several alterations and improvements (Ohuchi 2012). In this method, a chemically synthesized double-stranded oligonucleotide library (20–30 bp flanked by non-random sequences) incubated with target proteins and the unbound proteins are removed and the protein-DNA complex go through consecutive steps of binding and amplification (Darmostuk et al. 2015). SELEX is also a powerful method to determine whether a particular TF can bind to a specific DNA sequence. However, because this method is very labour intensive, more alternative methods has been developed for oligonucleotide selection (Viola et al. 2016). This method is extensively used to identify TF Protein-binding microarrays (PBMs) and around 913 of the known TF motifs were obtained by using high-throughput HT-SELEX assays (Lambert et al. 2018).

Active TF Identification (ATI)

Active TF Identification (ATI) was first described by (Wei et al. 2017) as *in vitro* parallel assay to Identify TFs in cells. In this method proteins extracted from cells or tissues are incubated with double-stranded DNA oligonucleotides (40 bp of random sequences). DNA bound to the proteins are separated from unbound DNA by native PAGE gel purification and amplified by PCR. The amplification step can be repeated at least three times to enrich DNA library that represent the activate TF binding sites. The method can be divided to two parts: motif analysis from amplified library sequence and MS identification of the active TFs. The method was used

to characterise TFs in mouse embryonic stem cells and mapped around 70 TFs from several families such as Nuclear Factor I (NFI), regulatory factor X (RFX), KLF and octamer-binding protein (POU).

Concatenated tandem array of consensus TF response elements

Ding and colleagues developed an affinity reagent method for the enrichment of TFs at the proteome scale using a synthetic DNA containing TF response elements for different TF families (Ding et al. 2013). Using this method Ding's team was able to identify 400 TFs from a single cell line and a total 878 TFs from 11 different cell types, covering more than 50% of the gene products that code for the DNA-binding TFs in the genome (Ding et al. 2013). The use of multiple TFREs on catTFRE system allows for the enrichment and identification of many TFs in high throughput way (Zhou et al. 2017). However, the number of TFs found in this study significantly exceeded the original design of the multiple TFREs n catTFRE system (Zhou et al. 2017). CatTFRE method has been applied in several studies recently, a summary of some of these studies are represented in table 1.1.

Table 1.1 Studies where CatTFRE method has been applied

Author/s	Tissue/organism	Finding
Zhang et al. (2017)	Cardiomyocytes	Use integrating approach with catTFREs and RNA-seq to analyse the critical TFs in the protection against H2O2 injury (378 TFs are quantitatively identified)
Wang et al. (2018b)	Mouse liver and tissues	Profile the dynamic of the TFs (297 TFs are quantitatively identified) during the circadian cycle
Torres et al. (2018)	Colorectal cancer metastasis	TF enrichment to investigate the alterations in transcription and splicing regulators. define splicing factor SRSF3 as biomarker for colorectal cancer metastasis
Wei et al. (2019)	Rat models	Inspect the important TFs involved in the response to anti-steatosis drugs
Li et al. (2019a)	Human adipose-derived stem cells	Quantitatively identified 472 TFs during adipogenesis and identify three novel TFs (BATF3, MAFF and MXD4) that regulate adipogenesis.

1.6 Thesis aims

The transcriptional regulation system orchestrates almost all biological processes in the cell. TFs control transcription by binding to specific DNA motifs in regulatory elements to regulate the amount and timing of gene expression. CatTFREs was recently developed as a method to quantitatively identify TFs profile of a cell type at the proteome scale (Ding et al. 2013).

The aims of this study are:

- To investigate the pioneer TF FOXA1 and the role of a mutation K6R, which abrogates a SUMOylation site, on the FOXA DNA binding activity and global gene expression in prostate cancer cell line DU145.
- To assess the advantages and the limitations of catTFRE system by exploring FOXA1 pull-down by western blotting

In order to do so, stable DU145 FOXA1 cell lines will be established and nuclear extracts prepared. Enrichment of the TF FOXA1 will be performed using a catTFRE system previously designed in the lab. Different experimental pull-down conditions will be tested to reach the optimal conditions for TF enrichment. Unique binding site present in only one of the TFRE construct will be used to assess the specificity of the catTFRE system. The effect of FOXA1 K6R mutation on gene expression levels will be investigated using global RNA-seq analysis. A gene network analysis will be performed to identify the significant differentially expressed TFs that have binding sites in our catTFRE system.

This study will provide new knowledge about the catTFRE system to set up a global TF enrichment analysis. Global RNA-seq analysis will help us to understand the transcriptional regulation network regulated by FOXA1 and give new insight into how this SUMOylation deficient mutant could modify its transcriptional regulation.

2. Methods

The appendix chapter 6.4 contain the list of materials and recipes of buffers to perform the experiments described in this thesis. A list of computer software are also included in the appendix 6.5.

2.1 Bacterial cell preparation, storage and transformation

Autoclaved and sterile materials were used in all culturing protocols to avoid bacterial contamination.

2.1.1 Preparation of competent *E. coli* DH5α cells

Escherichia coli (*E. coli*) bacterial strain DH5α cells were treated with calcium chloride buffer to enable DNA plasmids to be attached to the cell membrane and make the membrane more permeable.

Procedure:

1. Competent DH5α cells were thawed on ice.

2. 5 μl of the cells were plated onto a 10 cm lysogeny broth (LB) agar plate and incubated for 18 hours at 37 °C.

3. Several colonies (10 to 12) from the LB agar plate were sub cultured in 100 ml super optimal broth (SOB) medium and incubated for 4 hours with shaking at 37 °C.

4. A calculated amount (0.05 x 250 / OD600) of the preculture was added to 250 ml of fresh SOB medium to obtain a final optical density at 600 nm (OD600) of 0.05.

5. The bacterial culture was incubated for 16-18 hours with shaking at 18 °C then chilled on ice for 10 minutes.

6. The culture was split to five 50 ml sterile falcon tubes and centrifugated at 2500 relative centrifugal force (RCF) for 10 minutes at 4 °C.

7. The bacterial cell pellets were pooled and resuspended in 80 ml transformation buffer (appendix 6.5) and kept on ice for 10 minutes and then centrifugated at 2500 RCF at 4 °C for 10 minutes.

8. The bacteria were resuspended in 20 ml transformation buffer and 7 % (v/v) final concentration of Dimethyl sulfoxide (DMSO) (Sigma-Aldrich, USA) and incubated for 15 minutes on ice.

9. Bacterial cell suspension was in divided in 200 µl aliquots (in sterile and cold tubes) that were snap-frozen in liquid nitrogen and stored at -80 °C for future use.

10. Transformation efficiency was tested by transforming the bacteria with 0.1, 1 and 10 pg of pUC19 plasmid DNA and plated onto ampicillin containing LB agar plates. Colonies were counted after 16-18 hours to predict the expected number of colonies from 1 µg of plasmid DNA.

2.1.2 Transformation of competent cells

Bacterial transformation enables the introduction of a DNA plasmid in competent bacterial cells for amplification. A successful transformation requires that the plasmid DNA contain a bacterial origin of replication and the proper antibiotic resistance gene.

Procedure:

1. 50 µl of competent cells were thawed on ice.

2. 100 ng DNA (1-5 µl) was added to the cells.

3. The cells and the DNA mixture were incubated for 30 minutes on ice.

4. The cells at were heat shock incubated for exactly 90 seconds at 42 °C.

5. The cells were incubated for 2 minutes on ice.

6. Transformed cells were spread on a LB agar plate containing the appropriate selection antibiotics and incubate for 16-18 hours at 37 °C.

7. Colonies (4 to 6) were picked and inoculated in 5 ml LB medium containing the appropriate antibiotics and incubate for 16-18 hours with shaking (200 RPM) at 37 °C. Following purification of the plasmids, successful clones were verified by restriction.

2.1.3 Culturing and storage of competent cells

The transformed DH5α cells were cultured in LB medium or on LB agar plates complemented with ampicillin (100 µg/ml) used as selectable marker and incubated overnight with shaking at 37 °C (200 RPM). The bacterial clones after a successful transformation were resistant to ampicillin and used to prepare bacterial stocks for plasmid amplification prior to purification using mini-, maxi-, or gigaprep kits. Positive clones were stored at -80 °C. The cryopreserved stocks were prepared by adding 17 % (vol/vol) final concentration of glycerol and in 1ml aliquots. The competent cells could be sub-cultured directly from the glycerol stocks in LB medium with proper antibiotics for downstream experiments.

2.2 DNA related methods

2.2.1 Plasmid DNA isolation and purification from recombinant bacteria

Plasmid DNA purification following transformation differed based on the amount and quality of DNA required for downstream experiments. To verify the plasmids the clones where miniprepped following the Macherey-Nagel manual, protocol 5.1, 2017b using the NucleoSpin Plasmid kit (Macherey-Nagel). To obtain a larger amounts of high quality plasmid DNA, as used for the purification of the FOXA1-3TY clone used to establish stable overexpressing DU145 prostate cancer cell lines, the maxiprep Macherey-Nagel manual, protocol 7.1, 2017b from the NucleoSpin Plasmid kit (Macherey-Nagel) was used. For catTFRE construct purification the Qiagen Giga Prep ® plasmid isolation kit was used (QIAGEN-Plasmid-Purification-Handbook April-2012). The gigaprep plasmids were purified from 2.5 liter of overnight culture according to the manufacturer's instructions. Following elution the DNA was precipitated with 35 % isopropanol followed by wash with 70 % (vol/vol) ethanol and reconstituted in 2 ml of TE buffer.

2.2.2 Biotinylation of catTFRE constructs

Three synthesized TFRE-plasmids and control pGL4.26PvuII- plasmid (Appendix 6.3, figure 6.3.1) were transformed into DH5☐ cells, amplified and purified using Qiagen Giga Prep ® plasmid isolation kit. Purified plasmids were restriction enzymes digested as indicated:

Table 2.1 Restriction digestion 1

Construct 1	Construct 2	Construct 3	pGL4.26PvuII-
500 µg DNA (53 µl)	500 µg DNA (416 µl)	500 µg DNA (166 µl)	500 µg DNA (277 µl)
75 µl BSA (1 µg/µl)	75 µl BSA (1 µg/µl)	75 µl BSA (1 µg/µl)	75 µl BSA (1 µg/µl)
75 µl Buffer 2.1 (10x NEB)	75 µl Buffer 1.1 (10x NEB)	75 µl Buffer 2.1 (10x NEB)	75 µl Buffer 3.1 (10x NEB)
10 µl Eco53Ki (NEB)	10 µl KpnI (NEB)	10 µl Eco53Ki (NEB)	10 µl PvuII(NEB)
537 µl ddH$_2$O	174 µl ddH$_2$O	424 µl ddH$_2$O	312 µl ddH2O

The restriction enzymes were added into preheated mixtures (10 minutes at 37 °C) and incubated for 3 hours at 37 °C. To, confirm efficient digestion, 5 µl from each reaction were separated on a 0.8 % (w/vol) agarose gel alongside 1 µl of undigested plasmids.

The digested constructs from above were mixed as indicated:

Table 2.2 Restriction digestion 2

Construct 1	Construct 2	Construct 3	pGL4.26PvuII-
750 µl D1	750 µl D1	750 µl D1	750 µl D1
5 µl BSA (1 µg/µl)	5 µl BSA (1 µg/µl)	5 µl BSA (1 µg/µl)	5 µl BSA (1 µg/µl)
42.5 µl 1 M NaCl (50 mM)	85 µl 1 M NaCl (100 mM)	37.5 µl 1 M NaCl (50 mM)	5 µl Buffer 3.1 (10x NEB)
10 µl BamHI (NEB)	10 µl BamHI (NEB)	10 µl BamHI (NEB)	10 µl BamHI (NEB)
42.5 µl ddH$_2$O	6 µl ddH$_2$O	42.5 µl ddH$_2$O	30 µl ddH2O

Vortex and incubated for 3 hours at 37 °C. 5 µl from each reaction were checked on a 0.8 % (w/vol) agarose gel as above.

DNA precipitation

Following the second digestion, the DNA was precipitated by adding 85 µl NaOAc (3M) pH 5.2 and 600 µl Isopropanol to each of the reactions above and incubated for 30 min on ice. The reactions were centrifuged at 16000 RCF for 30 minutes at 4 ℃. The sedimented DNA pellet

was washed once with 1.5 ml 70 % (vol/vol) ethanol. Dried pellets were dissolved in 252.9 µl ddH2O by carful pipetting and incubation for 30 min at 37 ℃.

DNA Biotinylation

Klenow Fragment (3′→ 5′ exo-) was used to add biotinylated nucleotides to the linearized construct using the following protocol in table 2.3.

Table 2.3 Biotinylation mix

DNA	252.9 µl
Buffer 2 (NEB)	30 µl
dCTP (100 mM)	0.8 µl
dGTP (100 mM)	0.8 µl
Biotinylated dUTP (1 mM)	1.5 µl
Biotinylated dATP (0.4 mM)	4 µl
Klenow (NEB 5U/µl)	2 µl

The reaction from above was incubated for 2 hours at 37 °C followed with a 20 minutes incubation at 70 °C to inactivate the Klenow enzyme. Excess of biotinylated nucleotides were removed by Sepharose G50 columns according to manufactures instructions (Roche).

Biotinylated DNA constructs were digested with PstI to separate biotinylated-TFRE from plasmid backbone.

Table 2.4 Restriction digestion 3

Biotinylated DNA	300 µl
BSA (1 µg/µl)	75 µl
Buffer 3.1 (10x NEB)	45 µl
PstI (NEB)	5 µl
ddH2O	25 µl

The reaction was incubated for 3 hours at 37 °C. To verify effective digestion, 1 µl of digested DNA construct alongside known concentrations of each of the undigested constructs were separated on a 0.8 % (w/vol) agarose gel.

Immobilization of biotinylated DNA

Biotinylated DNA (1 ug) were immobilized onto Dynabeads (5 µl) using the kilobaseBINDER™ Kit (Catalog no. 60101, 2011 revised manual) according to the manufacturer's instructions. The DNA binding mix was rotated overnight at room temperature. Immobilized DNA was resuspended in 10 µl TE buffer and kept at 4 ℃ for later use. To test

the DNA binding efficiency, immobilized DNA was boiled in 0.1 % SDS at 95 ℃ for 5 min and 10 µl was checked on a 0.8 % (w/vol) agarose gel. To verify specific binding of the biotinylated-TFRE construct to the Dynabeads 10 µl of the supernatant and the wash preceding DNA immobilization were also kept and separated on an a 0.8 % (w/vol) agarose gel.

2.2.3 DNA quantification

DNA concentration was measured by a NanoDrop spectrophotometer that calculates the concentration in ng/µl using ultraviolet (UV) light absorption at 260 nm wavelength. The DNA purity of a sample with a values around 1.8 and 2.0 respectively was also analyzed

2.2.4 Agarose gel electrophoresis

Agarose gel electrophoresis was used to separate DNA fragments according to size as smaller fragments migrate faster in the gel than the larger molecules. The pore sizes in the agarose matrix is important factor and can be adjusted by agarose concentration, with lower percentages (w/vol) used for separation of larger fragments and higher percentages (w/vol) used for shorter fragments. For the agarose gel electrophoresis Tris-acetate-EDTA (TAE) buffer was used as running buffer. The TAE buffer help to stabilize the pH and the electrical current, and the EDTA protects DNA from enzymatic degradation by chelating magnesium ions. The DNA fragments were visualized by ethidium bromide (EtBr) that intercalates between the base pairings of dsDNA and omit fluorescence when exposed to UV light.

Procedure:

1. To prepare a 0.8 % (w/vol) agarose gel, 0.8 g agarose were added to 100 ml of 1x TAE buffer in an Erlenmeyer colbe.

2. The solution was heated until the agarose completely dissolved and chilled briefly.

3. One drop of EtBr (approximately 0.5 µg/ml) was added.

4. The gel was casted into a gel tray and left 20 minutes to solidify at room temperature.

5. DNA samples and a DNA size marker (1kb+) (Thermo Fisher Scientific, Waltham,

MA) were loaded into the gel wells.

6. The gel was run for 40-60 minutes at 100 V.

7. The DNA fragments were visualized with UV-light using Gel Documentation Systems (Bio-Rad).

2.2.5 DNA extraction and purification from agarose gels

DNA fragments were extracted from the gel and purified using the 'DNA extraction from agarose gels' (protocol 5.2) protocol from the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer's instructions.

2.2.6 DNA sequencing

To verify if that cloned DNA constructs contained the correct sequence purified plasmids were sent for sequencing to Eurofins GATC Biotech following the company's instructions. The DNA sequence was analyzed by SnapGene software (from Insightful Science; available at snapgene.com)

2.3 Mammalian cell culture

Human metastatic prostate cancer cell line DU145 was obtained from the American Type Culture Collection (ATCC, Rockville, Maryland, USA). According to the provider's description, the cell line was isolated from a lesion in the brain of a patient with metastatic carcinoma of the prostate and previous lymphocytic leukaemia (Source: https://www.lgcstandards-atcc.org/) . Cell culturing was performed in a sterile laminar flow hood. The cells were grown in a humified atmosphere of 5 % CO_2 at 37 °C.

2.3.1 Cell culturing

DU145 cells were cultured in 1X Dulbecco's Modified Eagle Medium (DMEM) high glucose (Gibco, Cat. 41965039) with 10 % fetal bovine serum (FBS) and 1 % (vol/vol) Penicillin-Streptomycin mixture (P/S) (Gibco, Cat. 15140-122) and 400 µg/ml G418 (Geneticin) (ThermoFisher). In order to keep the cell growth under optimal condition cells were sub-cultured when they reached confluence of 80-90 %. Since DU145 are adherent cells and grow attached to the culturing vessel. An enzyme with a protease activity that can break down the adhesion proteins, trypsin, were used to detach the cells. To avoid cell degradation upon trypsination, fresh culture medium containing FBS was added to deactivate the enzyme.

Procedure:

1. Cell culture medium and trypsin-EDTA were pre-warm to 37 °C.
2. The medium were aspirated of the cells and rinsed once with PBS.

3. 1x trypsin-EDTA were added to cover the cells and incubated for 2 minutes at 37 °C to detach and verified in the microscope.

4. Trypsin-EDTA activity were quenched by diluting 5-fold in fresh FBS containing medium.

5. The cell suspension was centrifuged at 200 RCF for 4 minutes at room temperature.

6. The cell containing pellet were resuspend in medium, counted using the Countess II FL automated cell counter (Invitrogen), before appropriate cell numbers were passed into new flasks/trays for downstream experiments.

The volumes Trypsin-EDTA, FBS containing medium to quench the enzyme and the number of DU145 cells to sustain culture for the different culturing vessels

2.3.2 Cryopreservation and thawing of cells.

Cell lines can be stored for long storage in liquid nitrogen at -196 °C. Standard gradient freezing with slow rate (-1 °C/minute) to prevent formation of ice crystals were used to freeze all cell lines in this study. A cell freezing container, placed in a -80 °C freezer were used to achieve the slow rate freezing. **Procedure for freezing cells:**

1. 2x concentrated freezing medium was prepared and cooled down.

Table 2.5 2X freezing medium

Reagent	Final concentration/volume
FBS	20 %
DMSO	10 %
Cell culture medium used for passaging	7 ml
Total volume	10 ml

2. The sub-culturing procedures in section 2.3.2 were followed from step 2-6.

3. The cell containing pellet in step 6 were resuspended with 4×10^6 cells/ml in culturing medium.

4. 2×10^6 cells (0.5 ml) of the cell suspension were mixed gently with 0.5 ml of the 2x freezing medium and transferred into a cryopreservation vial, placed in a cryopreservation box and transferred to a -80 °C freezer. Frozen cells were transferred for long term storage in liquid nitrogen the following day.

Procedure for thawing cells:

1. The cryopreservation vial with frozen cells was thawed at 37 °C until a small clump of ice was left.

2. The cells were slowly resuspended in 10 ml of medium and pelleted by centrifugation at 200 RCF at for 4 minutes at room temperature.

3. The supernatant was discarded and the cell containing pellet carefully resuspended in cell culture medium and seeded into an appropriate cell culturing vessel.

4. Cell survival was verified by microscopy the following day.

2.3.3 Transfection

To generate stable DU145 cell lines overexpressing FOXA1-3xTY or 3xTY transfection by electroporation were applied to introduce the plasmids into the cells. The cells are subjected to an electrical pulses that temporarily create small pores in the cell membrane that allow the linearized foreign DNA to enter the cells. The constructs contain an antibiotic resistance gene to enable selection of successfully transfected cells by the selected antibiotics. The DNA constructs were linearized prior electroporation by the Neon™ Transfection System (Invitrogen) according to the manufacturer's instructions. The FOXA1-K6R-3xTY stable cell line was generated by Ignacio Cuervo and was recloned to assure individual clean clones. Illustration of the DNA constructs are shown at appendix 6.4, figure 6.4.2

Procedure:

1. 2×10^6 cells were washed once in PBS and pelleted by centrifugation at 200x RCF for 4 minutes at room temperature.

2. The cell pellet was resuspended in 100 μl of Resuspension Buffer R.

3. A Neon® Tube with 3 ml Electrolytic Buffer E2 was set into the Neon® pipette station.

4. 2 μg of linearized DNA in 10 μl were transferred into an eppendorf tube.

5. The cells from step 2 were added to the tube containing the plasmid DNA and gently mixed.

6. The cell/DNA mix was pipetted into the Neon® pipette using the 100 μl Neon® Tip.

7. The Neon® Pipette was placed into the Neon® Tube electroporation performed with DU145 cell type specific settings: Pulse voltage (v) to 1.26, pulse width (m/s) to 20 and pulse number to 2.

8. Electroporated cells were carefully resuspended in medium with a pasteur pipette and the cells transferred into a culturing dish.

9. 24 hours post-transfection medium were replenished on the transfected cells with fresh media containing 400 µg/ml G418 antibiotics to select successfully transfected cells. Most cells should be successfully transfected under selection conditions where all the cells in the negative control (cells without added DNA) had died.

2.3.4 Generation of stable single colony cell lines

Successfully selected stable cell lines were seeded into 15 cm petri dishes for single clone selection ($1x10^3$ and $5x10^3$ cell/plate). Single colonies were picked at day 7 using 0.5 cm cloning disks (Sigma-Aldrich, USA) soaked with Trypsin-EDTA and cells transferred to 12 wells plate for further expansion. Confluent cell clones were tested for ectopic protein expression using western blot analysis and frozen down for future experiments. Three clones of each construct with similar ectopic protein expression were selected for downstream experiments.

2.3.5 Mycoplasma testing

All cell lines used in this study were tested for mycoplasma contamination before any further experiments were conducted. Prior to testing cells were cultured in antibiotic free medium for one week prior to DNA isolation. DNA from 1 million cells was sent to Eurofins GATC Biotech for testing. All included cell lines were mycoplasma negative.

2.4 Protein related methods

2.4.1 Nuclear protein extraction

Nuclear protein extracts are extensively used to explore TFs DNA binding or to analyze TFs trafficking between the nucleus and the cytosol. The method is also important to reduces protein mixture complexity (background from cytosolic proteins) and facilitate the detection of the low abundant TFs. To prepare DU145 cell nuclear extracts the following experimental setup were performed:

Cell harvest

1. Three large (T175) cell culture flasks of DU145 cells were grown to 80 % confluency and harvested as described in section 2.3.2

2. Cells were washed twice with ice cold 1xPBS and centrifugated at 200 RCF for 8 minutes at room temperature.

3. The cell pellet were resuspended in ice cold 30 ml PBS and the cells were counted as described in section 2.3.3

Nuclear Isolation

1. Cell suspension was centrifuged at 200 RCF for 10 minutes at 4 ºC.

2. The cell containing pellet was resuspended in 20 ml of ice cold hypotonic buffer N (Appendix 6.4).

3. Cells were resedimented by centrifugation at 200 RCF for 10 minutes at 4 ºC.

4. The cell pellet was resuspend in10-20 vol of ice cold hypotonic buffer N containing 1 mM DTT and protease inhibitors and incubated for 30 minutes on ice.

5. Nuclei were released with a glass Dounce thigh pestle (cooled before use) on ice. Cell lysis were assessed by phase contrast microscope to make sure that all cells were lysed but that the nuclei were remained intact. DU145 cells required 10 to 12 strokes with a 7 ml Dounce pestle to release the nuclei.

6. 125 μl of 2M sucrose was added per ml of released nuclear solution and mixed well by inversion.

7. The lysate were centrifuged 200 RCF in 15 ml falcon tubes for 10 min at 4 °C in swinging bucket rotor.

8. The nuclei containing pellet were resuspended in 10 ml of ice-cold buffer N and resedimented by centrifugation at 200 RCF for 10 min at 4 °C.

9. The clean nuclear pellet was resuspended in appropriate volumes of nuclear extract buffer C with 0.5 % NP40.

10. The nuclei were snap frozen and kept at – 80 °C for downstream experiments.

Preparation of nuclear extracts

Frozen nuclei were thawed on ice and sonicated using the Biorupter Pico with the following settings (30sec on/off x 3 cycles).

2.4.2 Bradford protein assay

In this study Pierce™ Coomassie (Bradford) Protein Assay Kit (Thermo Fisher, Cat. 23200) were used to estimate the total protein concentration in the nuclear extract. The assay is based on the binding between the acidic residues, arginine, lysine and histidine in the proteins and the coomassie dye which results in a color change.

Procedure:

1. Enough amount of the Bradford Solution was prepared by adding coomassie (Bradford) solution to Milli-Q water with 1: 4 ratios

2. A Bovine serum albumin BSA standard curve (0, 0.5, 1, 2, 3 and 5 µg/µl) diluted in buffer C was used to calculate the nuclear extract concentrations.

3. Three technical replicates and several dilutions were performed for each sample to have accurate measurement were pipetted into a Greimer labortechnik PS micro plate flat bottom 96 plate.

4. 200 µl Bradford user solution were added to each well and incubated for 30 minutes at room temperature in the dark following absorbance measurements at wavelength 595 nm using a FLUOstar OPTIMA plate reader.

2.4.3 Transcription factor pulldown using the catTFRE system

Biotinylated Tandem-TFRE DNA was incubated with DU145 cell nuclear extracts prepared as described in section 2.4.1. Transcription factors bound to mobilize biotinylated DNA (referred to as DNA-bound beads) were pulled down using magnetic force and visualized by western blot.

Procedure:

1. Frozen nuclear extracts were thawed on ice and pre-cleared by centrifugation 11000 RCF for 10 min at 4 ℃.

2. 20 µg of nuclear extracts were adjusted to a final volume of 50 µl in buffer C.

3. 1 µg of mobilized biotinylated TFRE construct from each clone and condition were washed twice in 200 µl of 1 mg/ml BSA in 1x PBS and resuspended in 200 µl in buffer BC.

Note: Different BC buffers were used to optimize the conditions for TFs enrichment (appendix 6.5).

4. Nuclear extracts were incubated with the DNA-bound beads rotating for 2 hours at 4 ℃.

5. DNA-bound beads were washed 3 times rotating for 10 min at 4 ℃ in buffer BC.

6. DNA-bound beads were washed twice in 200 µl 1x PBS

7. The supernatant was removed and TF containing DNA-bound beads boiled for 5 min in 20 µl 2X SDS loading dye (Appendix 6.4). Transcription factors were separated by SDS PAGE Gel 4-20 % and visualized by coomassie staining and western blotting.

2.4.4 Western blotting

The main purpose of this method is the recognition of a specific protein using antibodies. All released proteins were separated by size in a gel matrix that is subjected to an electrical current. The presence of the SDS in the sample buffer facilities the protein unfolding and coat these proteins with a negative charge allowing them to migrate towards an anode. The acrylamide gel matrix form pores through which the proteins can move and larger proteins move slower than smaller proteins. After the proteins were separated by size through the gel, these are transferred to a Polyvinylidene difluoride (PVDF) solid support matrix by an electrical current. A blocking step was required to stop any unspecific binding by the antibodies during the next steps. First the primary antibodies were added for 1 hour at room temperature or overnight at 4 °C to recognizing proteins of interest. Then the secondary antibodies were added to recognize the primary antibodies. Secondary antibodies are coupled to an IR800 or IR680 (LI-COR) fluorescent tag to enable visualization of the protein of interest.

SDS-PAGE:

1. Protein samples in SDS running buffer were run on a 4-20 % TGX (BioRad) gradient gel for +/- 1 hour at 160 V and 400 mA until the front ran out.

2. Transfer of the proteins to PVDF membranes were performed using The Trans-Blot Turbo System (BioRad) according to the manufacturers instructing using the BioRad Standard SD 30 min program.

Western blotting:

3. The protein containing PVDF membrane was transferred to a container with 1:1 1x PBS and blocking buffer (LI-COR), the blot was blocked for 1 hour at RT with constant shaking.

4. The blot was incubated with primary antibodies in 1:1 1x PBS and blocking buffer (LI-COR) for 1 hour with constant shaking at RT or overnight at 4 °C.

5. The blot was washed 5 times for 5 minutes with 1x TBS-T with constant shaking at room temperature.

6. The blot was incubated with secondary antibodies in 1:1 1x PBS and blocking buffer (LI-COR) for 1 hour with constant shaking at RT in a light-protective box.

7. The blot was washed 5 times for 5 minutes with 1x TBS-T with constant shaking at room temperature.

8. The blot was developed with the Odyssey CLX imager (LI-COR).

2.4.5 Immunolocalisation analysis (performed by M. Rogne)

DU145 cells were fixed for 30 min in 4% paraformaldehyde, permeabilized for 15 min with 0.1% Triton-X 100 (Sigma),blocked for 2 hours in PBS with 0.01% Tween-20 (PBST) and 5% (fatty acid free) BSA. Primary antibody solution containing the indicated antibody were prepared in PBST-BSA and incubated over night at room temperature. Coverslips were washed five times in PBST-BSA before incubation in secondary Alexa fluor 488 donkey anti-mouse (1:500)(Invitrogen), Alexa fluor 555 donkey anti-rabbit (1:500) (Invitrogen) antibodies for 1 h. Coverslips were washed five times for 5 min in PBS with 0.1% tween, incubated 5 min with 4',6-diamidino-2-phenylindole (DAPI) (Sigma), rinsed in dH2O, and mounted with vectashield T1000. Images were acquired with an LSM880 Airyscan confocal microscope(Zeiss, Oberkochen, Germany) fitted with a $100 \times$ NA 1.45 oil plan Apochromat objective and using the Zen Black 2014 software. Identical settings and display were used between replicates and clones.

2.6 RNA-seq analysis

Total RNA was isolated from DU145 cell line overexpressing FOXA1 3xTy and FOXA1 K6R 3xTy using an RNA isolation kit (Qiagen - RNeasy mini kit) by Roza B. Lemma. The quality and quantity of total RNA was determined using NanoDrop (Thermo Fisher Scientific) and agarose gel electrophoresis, respectively. RNA samples were delivered for sequencing to the Norwegian sequencing centre, Oslo, Norway. Sequencing libraries were prepared using strand-specific TruSeq library preparation kit. Four biological replicates were generated for each cell line. Transcriptomic data was produced using Illumina HiSeq 4000 sequencer and 150 bp paired-end reads were obtained. Quality of raw reads were investigated using fastqc (Version 0.11.5) (Andrews 2010). Adaptors and low quality reads were removed using Trim Galore (Version 0.6.4) (Krueger 2015). Trimmed reads were mapped to the Ensembl GRCh38.p12 human genome using STAR (Version 2.7.1) (Dobin et al. 2013). Gene expression levels were quantified as raw read counts by HTSeq (Version 0.11.2) (Anders et al. 2015). Normalized FPKM values (reads per kilobase per million reads) was calculated for each transcript and the Differentially expressed genes (DEGs) were identified using DESeq2 (Version 1.26.0) (Love et al. 2014). Volcano plots showing differentially expressed genes were generated using R package ggplot2 and cluster heatmap for the shared significantly differentially regulated genes were generated using R package pheatmap. The network analysis of the significant differentially expressed transcription factor included was performed in the Genemania App of Cytoscape (Franz et al. 2018; Warde-Farley et al. 2010). Genemania utilizes different cancer related datasets to obtain coexpression and high confidence experimentally obtained protein-protein physical or predicted interactions, shared protein domains, co-localization and genetic interaction.

3 Results

3.1 Design and characterization of the catTFRE system

The catTFRE clones were designed by selecting 309 TFs binding sites in duplicate (covering an estimate of 387 TFs) spanning 25 different TF families from JASPAR by Trung Tran. The TF binding sites were randomly subdivided into three TFRE-DNA binding constructs where the two tandem repeats of each DNA binding site were separated by a spacer of three nucleotides (ATC), synthesized and cloned into pUC57 (Genscript) (Appendix Table 6.2.1 and Figure 6.3.1). As control plasmid, a pGL4.26[luc2/minP/Hygro] luciferase reporter vector (Promega) was selected, as it has reduced number of consensus transcription factor binding sites in the vector backbone and *Amp* and *Luc2* gene. This vector had been modified to remove a PvuII site (pGL4.26PvuII-) for the biotinylation design in house. A Schematic illustration of the experimental setup for biotinylation of TFRE DNAs is shown in Figure 3.1.

1. Amplification of the TFRE-plasmids
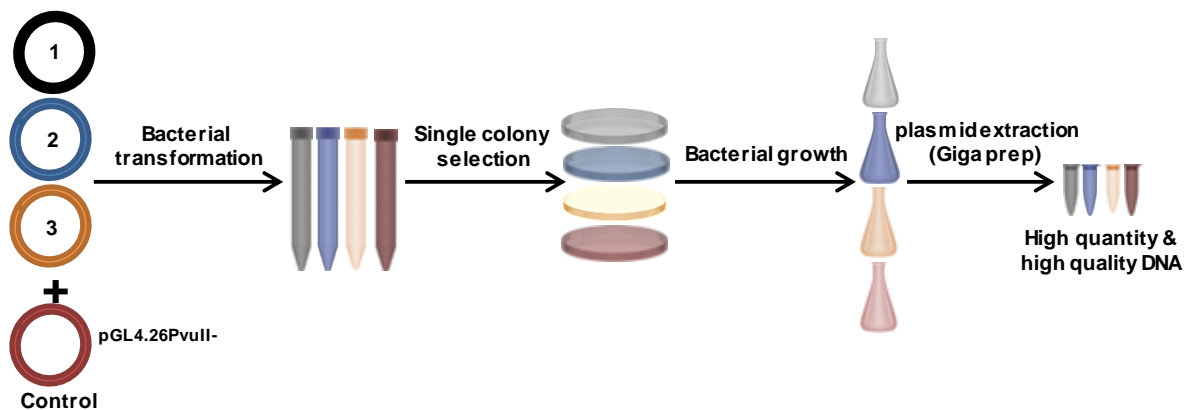
2. Linearization & biotinylation
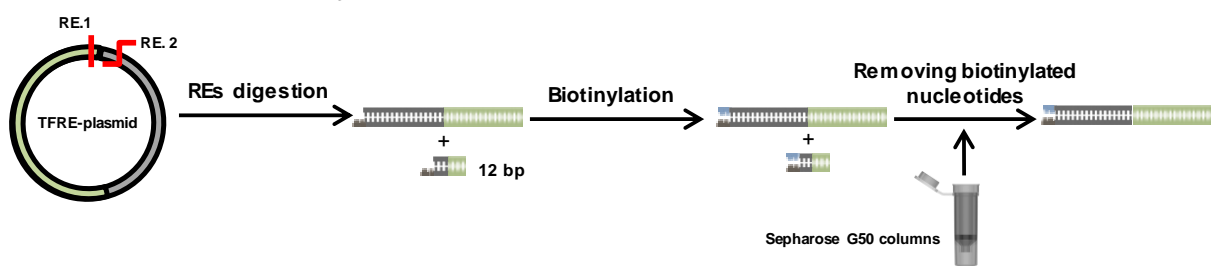
3. REs digestion & Immobilization
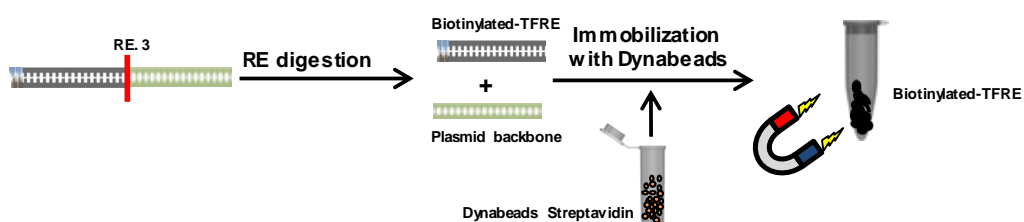
Figure 3.1. Schematic illustration of the experimental setup for biotinylation of TFRE DNAs.

Purified plasmids were digested with restriction enzymes 1 and 2 for linearization followed by biotinylated by Klenow Fragment (3′→ 5′ exo-) and biotinylated nucleotides. Excessive biotinylated nucleotides were removed by sepharose G50 columns. Finally, DNA was digested with restriction enzyme 3 to separate biotinylated-TFRE from the plasmid backbone. Biotinylated-TFRE was immobilized onto Streptavidin Dynabeads and pulled down using a magnetic force.

The three synthesized TFRE-plasmids and control pGL4.26PvuII- plasmid were amplified and purified using QIAGEN® Plasmid Plus Giga Kit. A schematic illustration of the synthesized TFRE-plasmids including the restriction enzyme cutting sites that were chosen for the downstream biotinylation are showed in Figure 3.2A. To examine the restriction enzyme cutting of the TFRE clones prior to biotinylation catTFRE clones were digested with Eco53I or Eco53I + BamHI (catTFRE clone 1 and 3) and KpnI or KpnI + BamHI (catTFRE clone 2) and visualized on an agarose gel (Figure 3.2B). The expected size of the constructs after the digestion were approximately 5937 bp for constructs 1 and 3 and 5903 bp for construct 2. The digested catTFRE clones were subjected to biotinylation using the Klenow (3′→ 5′ exo-) enzyme, before DNA was digested with PstI to separate biotinylated-TFRE from plasmid backbone (Figure 3.2C). The constructs expected size after the digestions were 3264 bp for the biotinylated-TFRE constructs and 2710 bp for plasmid backbone. Next, we immobilized the biotinylated clones onto Dynabeads™ and observed specific catTFRE DNA binding to the beads with only the plasmid backbone remaining in the supernatant (Figure 3.2D). The control pGL4.26PvuII- plasmid was first digested with PvuII or PvuII + BamHI and visualized on an agarose gel. The digested control plasmid was subjected to similar biotinylation setup as the catTFRE clones (Appendix Figure 6.3.3)
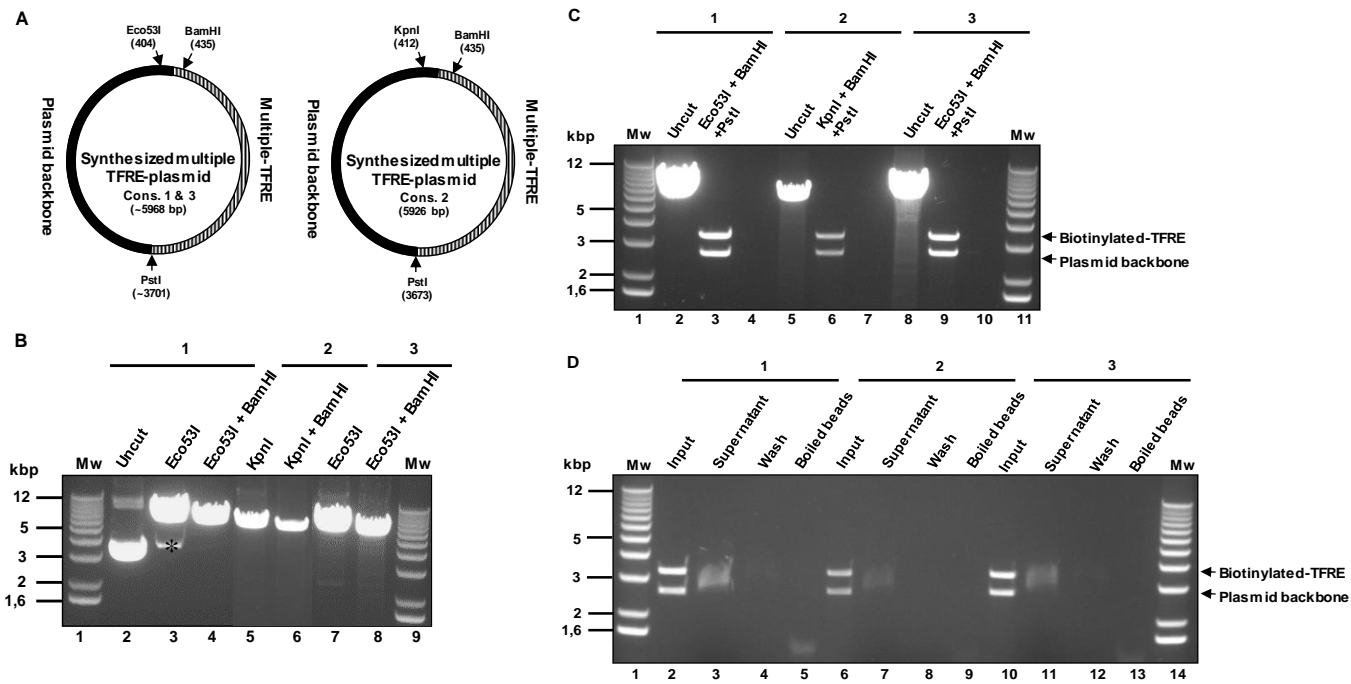
Figure 3.2. Binding of biotinylated TFRE DNA to Dynabeads.

(A) Schematic illustration of the synthesized TFRE-plasmids (construct 1 and 3 left) and the TFRE-plasmid (construct 2 right). The restriction enzyme cutting sites are indicated in black arrows. (B) First, each construct was linearized by digestion with a single cutter RE to produce blunt ends (Eco53I for construct 1 and 3, KpnI for construct 2). Next, each construct was digested with a single cutter BamHI to produce sticky ends enabling addition of biotinylated nucleotides by Klenow. (C) After biotinylation, DNA was digested with PstI to separate biotinylated-TFRE from plasmid backbone. Lane 2, 3 for uncut and cut construct 1. Lane 5, 6 for uncut and cut construct 2. Lane 8, 9 for uncut and cut construct 3. (D) Biotinylated DNA (1 ug) from construct 1, 2 and 3 was immobilized onto 4 µl of Dynabeads™ using kilobaseBINDER™ Kit to ensure specific binding of the biotinylated-TFRE construct to the beads. 10 % of the supernatant and the washes after DNA immobilization was loaded to the gel. Immobilized DNA was boiled in 0.1 % SDS at 95 ºC for 5 min and 50 % was loaded to the gel together with 1 µl from the input DNA was also loaded to the gel.

To estimate the optimal binding ratio between biotinylated DNA and the Dynabeads, 1 µg of the DNA from each of the catTFRE constructs and a pGL4.26PvuII- control plasmid were immobilized onto increasing amounts of Dynabeads (1, 2, 3 and 4 µl) (Figure 3.2A). We show an increase in specific binding of the catTFRE constructs to increased volumes of Dynabeads with 100 % binding capacity observed for1 µg of DNA with 4 µl of Dynabeads. We observed partial release of the biotinylated DNA to the supernatant upon boiling the Dynabeads (Figure 3.3A, well 11-14). The optimal binding ratio of 1µg biotinylated DNA to 4 µl of Dynabeads was verified (Figure 3.3B) and used for the downstream catTFRE pulldown experiments.
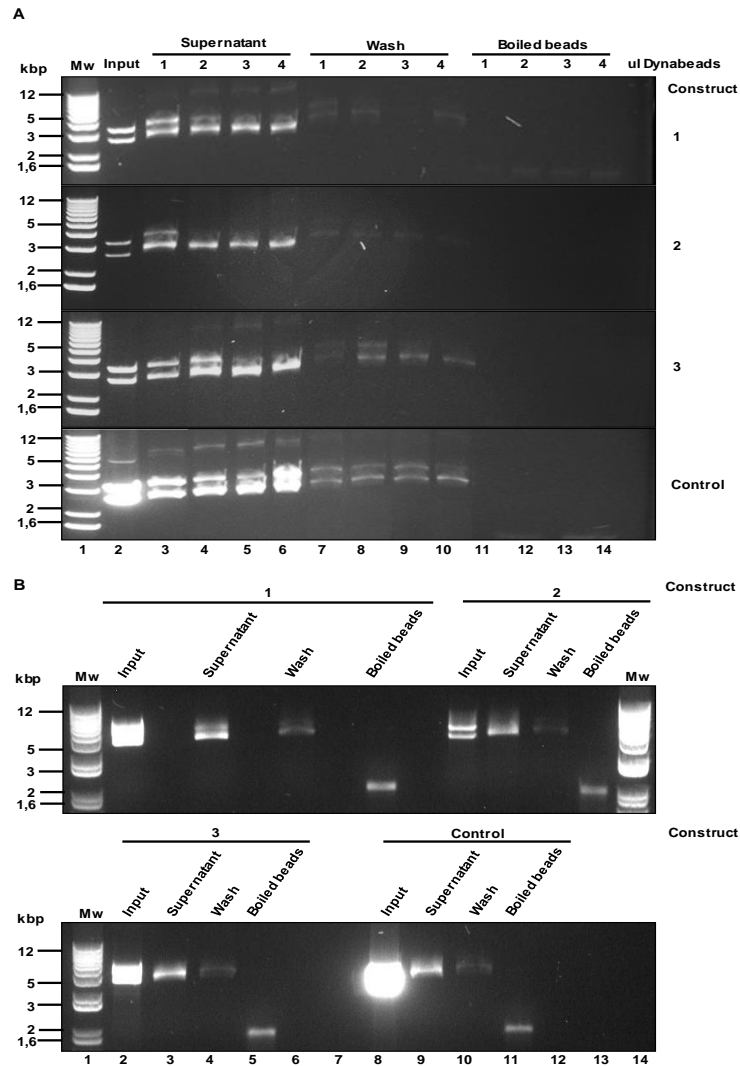
Figure 3.3. Optimizing the biotinylated TFRE and Dynabeads™ ratio.

(A) Biotinylated DNA (1 μg) was immobilized onto increasing amount of Dynabeads (1, 2, 3 and 4 µl) to assess binding efficiency. 2 µl of the supernatant and the washes was loaded on the agarose gel as a positive control for Dynabeads binding (lane 3 to 10). Biotinylated DNA was boiled in 0.1 % SDS at 95 °C for 5 min and 10 µl was loaded to the gel (lane 11 to 14). Input DNA (1 μg) was also loaded to the gel (lane 2). (B) Immobilization of the biotinylated TFRE constructs for downstream catTFRE experiments. Biotinylated DNA (1 μg) was immobilized using 4 µl of Dynabeads™ and kilobaseBINDER™ Kit. 2 µl of the supernatant and the wash after DNA immobilization loaded to confirm proper binding conditions. As positive control for DNA binding to Dynabeads, the beads was boiled formamide at 90 ºC for 2 min in 10 mM EDTA and 90 % formamide and loaded on the gel. Input DNA (1 μg) was also loaded to the gel (lane 2).

40

3.2 Generation of stable FOXA1 overexpressing DU145 cell lines

To investigate the role of SUMOylation in FOXA1 binding activity, SUMOylation-deficient FOXA1 mutant was generated by introducing a mutation at lysine K6 to mutate into arginine K6R. The position of this mutation was chosen based on the scanning for conserved FOXA1 SUMOylation sites in different species. Also based on protein intrinsic disorder prediction model using DISOPRED3 software (Jones et al. 2015) that allow us to detect the intrinsically disordered regions and protein-binding sites of FOXA1. Consequently, K6 was in highly disordered region and the only SUMOylation sites with high confidence score for protein binding as shown in Figure 3.4.
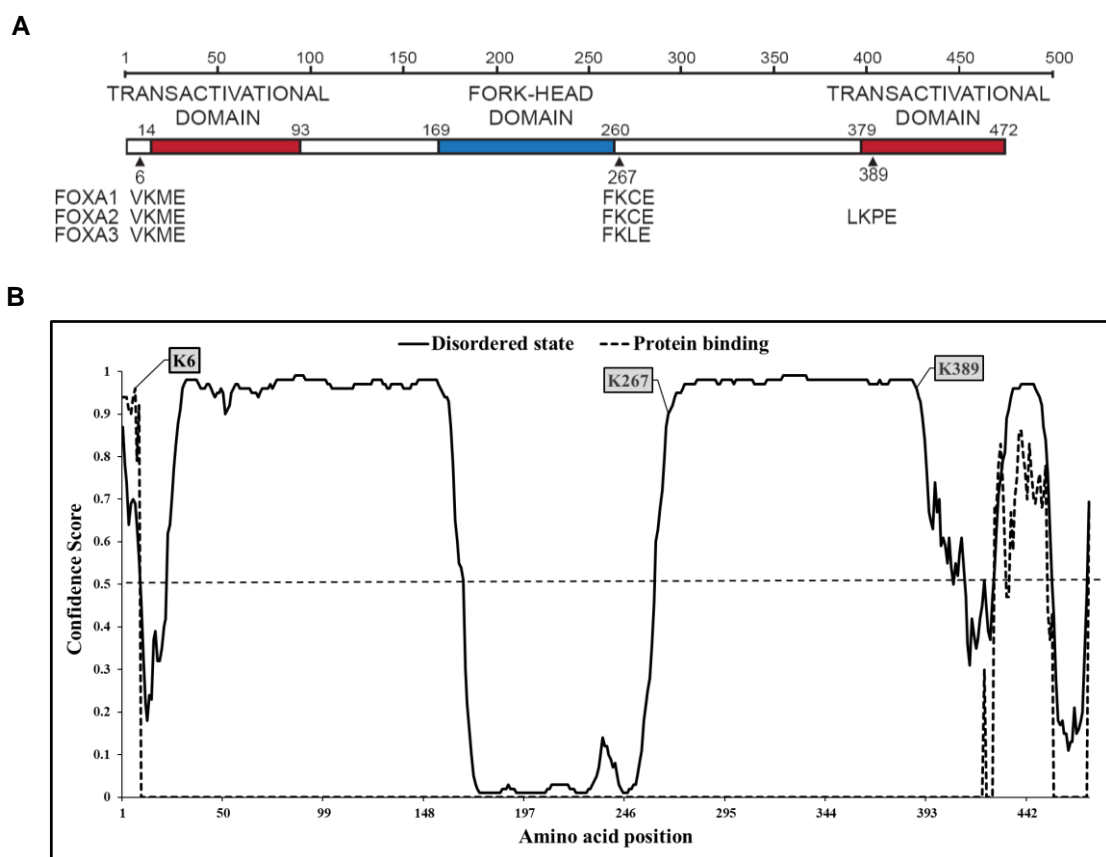
Figure 3.3 FOXA1 SUMOylation sites and intrinsically disordered regions.

(A) Schematic structure of FOXA1 showing the major functional domains and the lysines, K6, K267, and K389 that resemble SUMOylation sites as adapted from Sutinen et al. (2014).(B) A graph represents the DISOPRED3 disorder confidence level against the sequence positions as a solid black line. The black dashed line shows the confidence of disordered residues being involved in protein-protein interactions.

To assess the sensitivity and specificity of the catTFRE pull-down system, we generated stable clones of FOXA1-3xTY, FOXA1-K6R-3xTy and 3xTy overexpressing DU145 cells. The goal was to examine the DNA binding of the pioneer TF FOXA1 and its SUMOylation deficient K6R mutant. In order to achieve stable DU145 cell lines we transfected cells with FOXA1-3xTY, or the empty 3xTy control construct. DU145 cell lines we transfected cells with FOXA1-K6R-3xTy and the single clone cell lines were generated from a bulk culture. After antibiotic selection that efficiently selected for FOXA1-3xTy, FOXA1-K6R-3xTy and 3xTy expressing cells. A Schematic illustration for generating stable single clones of DU145 cell lines is showed in Figure 3.5.
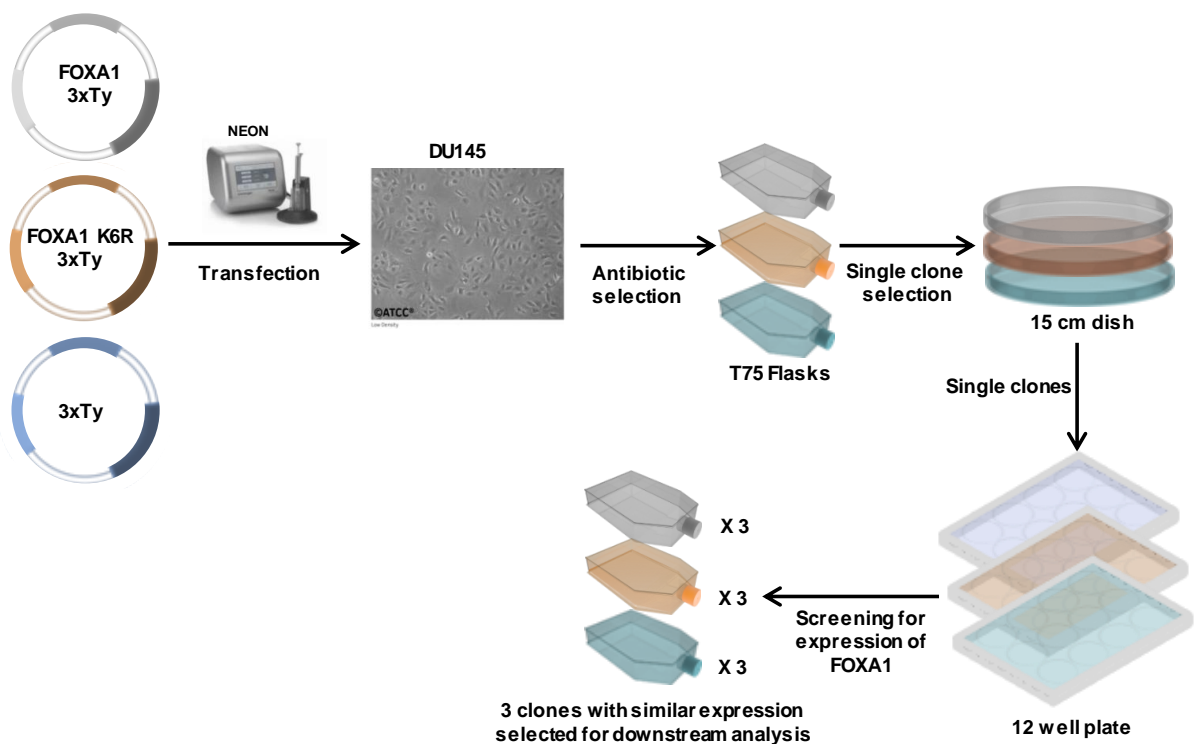
Figure 3.5. Experimental setup for generating stable single clones of DU145 cell lines overexpressing FOXA1 3xTy and 3xTy.

Stable cell lines were generated by DU145 transfection with constructs towards FOXA1-3xTy and 3xTy. FOXA1 K6R 3xTy was cloned from stable DU145 cell line generated by Ignacio Cuervo. Whole cell extract was made of $1x10^6$ cells to examine ectopic protein expression by western blotting. Three clones with similar ectopic protein expression were selected for each construct for downstream analysis.

Three single clones from each of the constructs, with comparable protein expression levels was visualized by immunoblotting (Figure 3.6A) and included for downstream experiments. Immunolocalization studies, performed by Marie Rogne, demonstrated the presence of a specific nuclear staining of FOXA1 3xTy and FOXA1 K6R 3xTy while the 3xTy overexpressing cell lines displayed a diffuse cytoplasmic localization (Figure 3.6B).

To prepare nuclear extracts for validation of the catTFRE system, purified nuclei from FOXA1-3xTy and 3xTy expressing cells were lysed in 500 µl RIPA buffer. Increasing concentrations (5, 10, 20 and 30 µg) of nuclear extracts were immunoblotted for the presence of FOXA1-3xTy (Figure 3.6C). The nuclear extract displayed robust FOXA1-3xTy levels. Moreover, nuclear extracts separated by SDS PAGE and stained with Colloidal Blue showed a protein rich lysate with numerous proteins at similar levels (Figure 3.6D).
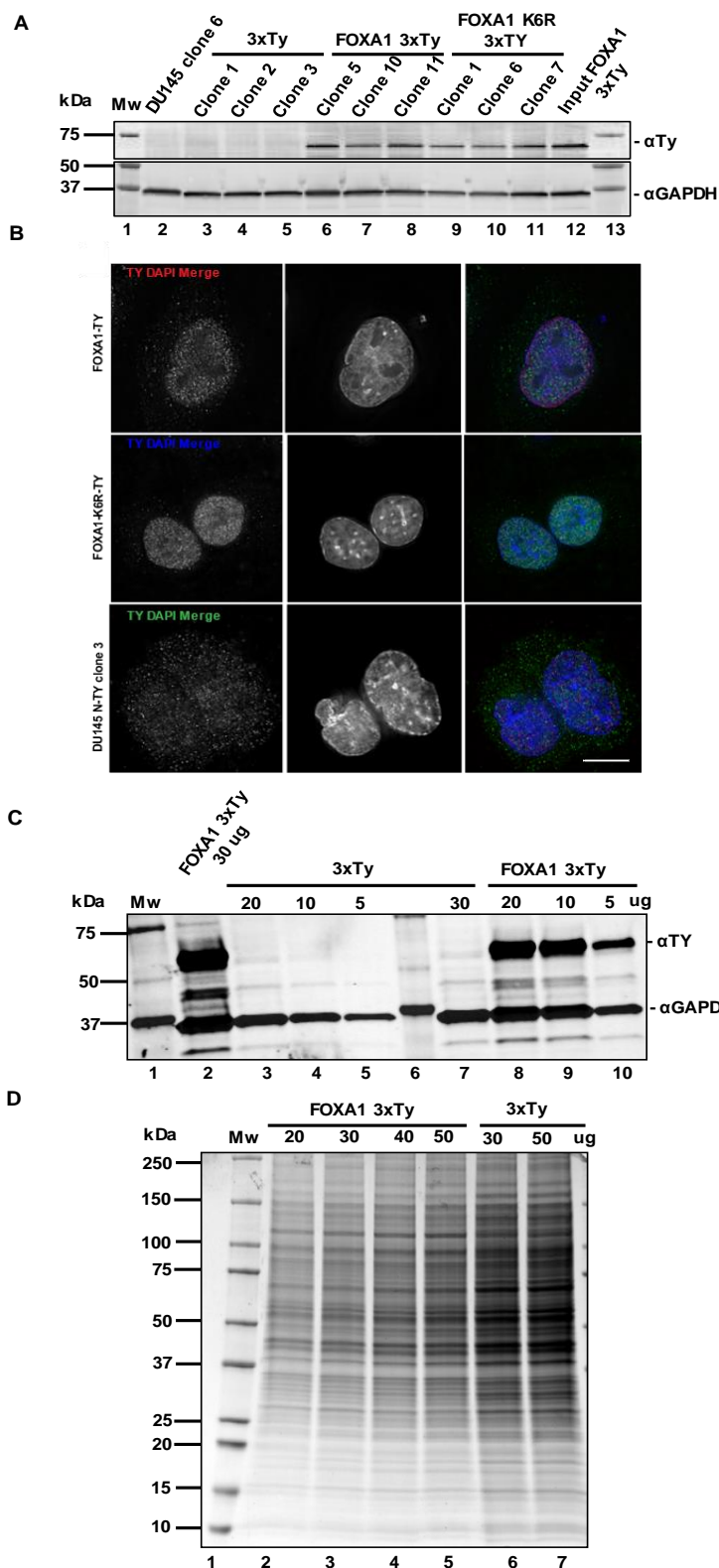
Figure 3.6. FOXA1 expression and localization.

(A) Western blot of 10 µg of DU145 cell extract for three clones for each construct with anti-Ty and GAPDH (Clone 1, 2 and 3 for 3xTy (lane 3 to 5), clone 5, 10 and 11 for FOXA1 3xTy (lane 6 to 8) and clone 1, 6 and 7 for FOXA1 K6R 3xTy (lane 9 to 11)). DU145 cell line used for initial transfection was used as negative control (lane 2) and DU145 overexpressed FOXA1 3xTy as positive (lane 12). (B) Immunofluorescence staining of DU145 cells with FOXA1 3xTy, FOXA1 K6R 3xTy or 3xTy overexpressing cell lines were fixed with 4 % paraformaldehyde, stained against αTy (1:100) and DAPI before subjected to LSM880 airyscan confocal imaging. Scale bar 10 uM. Pictures were taken with identical setting and display. Immunostain was performed by M. Rogne. (C) Western blot of purified nuclear extract of DU145 cells stably overexpressing FOXA1 3xTy or 3xTy, blotted against Ty and GAPDH. 5, 10, 20 and 30 µg of nuclear extract were loaded for FOXA1 3xTy (lane 2, 8, 9 and 10) and 3xTy (lane 3, 4 and 5). (D) SDS PAGE 4-20 % of nuclear extract stained with Colloidal Blue. 20, 30, 40 and 50 µg of DU145 nuclear extract of cells overexpress FOXA1 3xTy were loaded (lane 2 to 5) . 30 and 50 ug of 3xTy nuclear extract were loaded (lane 6 and 7).

3.3 Optimizations and validation of the catTFRE system by FOXA1 3xTy pulldown

To optimize the catTFRE pulldown system, nuclear extracts were incubated with biotinylated TFRE clones (clones 1 to 3) and the pGL4.26PvuII- control construct prior to washes with increasing NaCl salt concentrations (100, 300 and 500 mM). According to the design of the TFRE clones, the FOXA1 binding sites is present in construct 1 (a list with the TF binding sites included in the designed TFRE constructs is found in Appendix table 6.2.1). A Schematic illustration of the catTFRE experimental setup is showed in Figure 3.7.
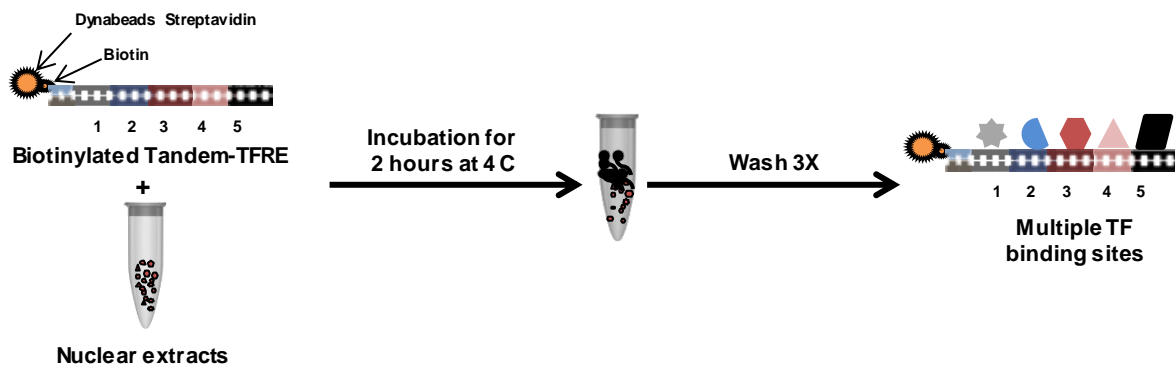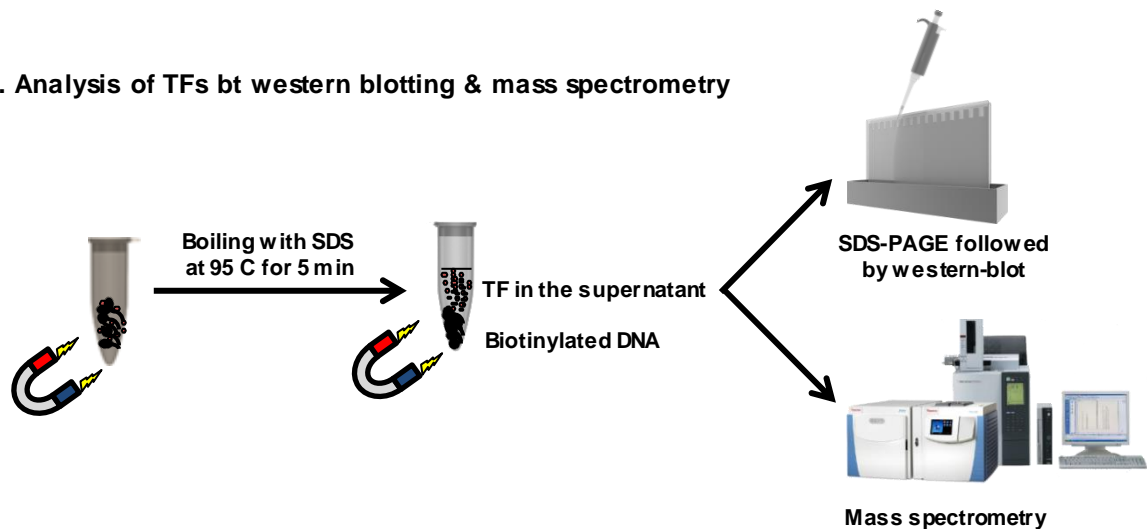
Figure 3.7. Illustration of the catTFRE experimental setup.

Immobilized biotinylated Tandem-TFRE (1 µg DNA) was incubated with 20 ug of nuclear extracts for 2 hours at 4 C. After incubation biotinylated DNA was washed three times with buffer C. Transcription factors bound to biotinylated DNA was pulled down using magnetic beads and boiled in SDS loading dye before separation on a 4-20 % SDS PAGE gel. TFs were visualized by western blot or will be analyzed by mass spectrometry.

Bound TFs were immunoblotted for the presence of FOXA1-3xTy (Figure 3.8A and B). At low salt concentrations unspecific FOXA1 3xTy bound to all TFRE clones and the pGL4.26PvuII- control construct. At higher salt concentrations only residual FOXA1 3xTy were present in the control construct. Surprisingly, FOXA1 bound to all three TFRE construct under all salt concentrations. This prompted us to examine if the designed catTFRE clone sequenced could contain additional unintended TF binding sites.
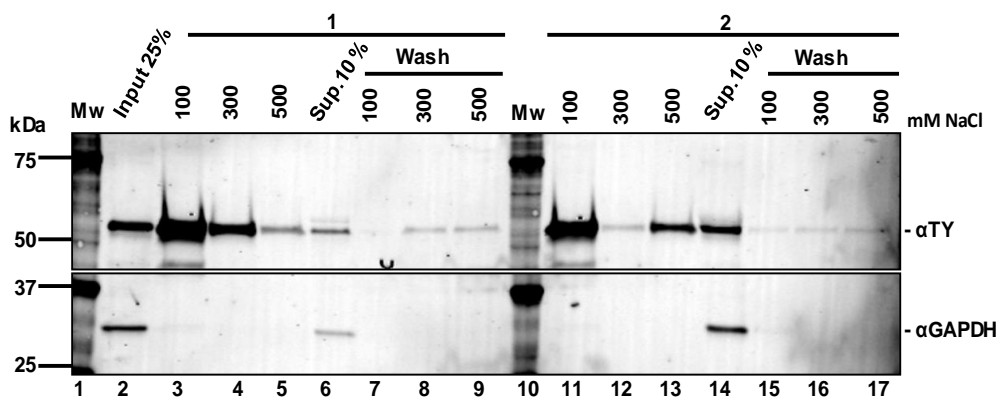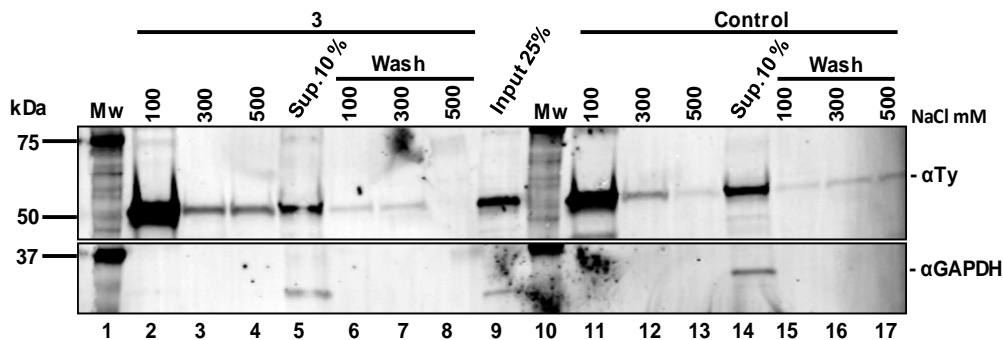
Figure 3.8. Optimization of FOXA1 3xTy pulldown by catTFRE system.

(A) Upper blot shows TFRE construct 1 and 2 with increasing salt concentration and 25 % of the input extract (5 ug) was loaded in lane 2. 10 % of the supernatant after incubation was loaded in lane 6 and 14 for construct 1 and 2 respectively (marked as Sup. 10 %). 6% of the wash buffer was loaded (marked as wash). (B) Lower blot shows TFRE construct 3 and control with increasing salt concentration and 25 % of the input extract (5 ug) was loaded in lane 9. 10 % of the supernatant after incubation was loaded in lane 5 and 14 for construct 3 and control respectively (marked as Sup. 10 %). 6 % of the wash buffer was loaded (marked as wash). Biotinylated Tandem-TFRE (clones 1 to 3) and the pGL4.26PvuII-control construct (1 µg DNA) were incubated with 100 ug of nuclear extracts of FOXA1 3xTy overexpressing DU145 cells (clone 10), washed with increasing salt concentration: 100, 300 and 500 mM NaCl before blotting against α-Ty and α-GAPDH (extract loading control)

An unbiased scan for transcription factor binding sites using position frequency matrices from the JASPAR database (8th release -2020) was performed on all TRFE construct using CiiiDER software (http://ciiider.com/). In total 730 TFs binding sites were identified combined for the TFRE constructs (Supplementary table 2). Several binding sites were identified for FOXA1 in all TFRE construct (Figure 3.9C). Moreover, most of the identified TF binding sites were common across all three TFRE constructs and with only a few unique TFs binding sites to each construct (List of unique TFs binding sites found in each TFRE are found in Table 3.1

Table 3.1. The number of total and unique TFs detected in each catTFRE clone

	Construct 1	Construct 2	Construct 3	pGL4.26PvuII-control
No. of TFs	632	652	642	396
No. of unique TFs	6	23	24	4
List of unique TFs	E2F3 EWSR1-FLI1 BCL6B CTCF NR2F6 (var.3) ZKSCAN5	NR1H2::RXRA NR2C2 NR2F1 (var.2) ZNF652 HLF HSF4 NFIB NFIC (var.2) NFIC::TLX1 NFIL3 NRF1 PAX9 HSF2 HSF1 MTF1 Rarg (var.2) Prdm15 RXRG (var.2) SIX1 ZBTB32 ZBTB6 ZNF135 ZNF682	PPARG NR1I2 ZBED1 EBF1 RFX2 RFX3 RFX4 RFX5 SOX21 SRF NR1D1 RXRA::VDR Nr2f6 (var.2) PLAG1 PROX1 RARA Rarb Rarg REST Sox11 Stat6 Wt1 ZNF410 NR1H4::RXRA	ZNF449 Nr2e3 ZBTB14 ZSCAN4

To examine if increased specificity could be achieved using various detergents and salt concentrations in the washes, nuclear extracts were incubated with biotinylated TFRE clones (clones 1 to 3) and the pGL4.26PvuII- control construct followed by indicated washes (Figure 3.9). We observed FOXA1 binding to all three TFRE construct under all salt and detergent concentrations. Although more stringent salt and detergent conditions reduced the unspecific binding of FOXA1 to the pGL4.26PvuII- control construct, the various conditions did not increase the specificity between the catTFRE clones (Figure 3.9A and B). A negative control with beads only did not show any FOXA1 3xTy binding.
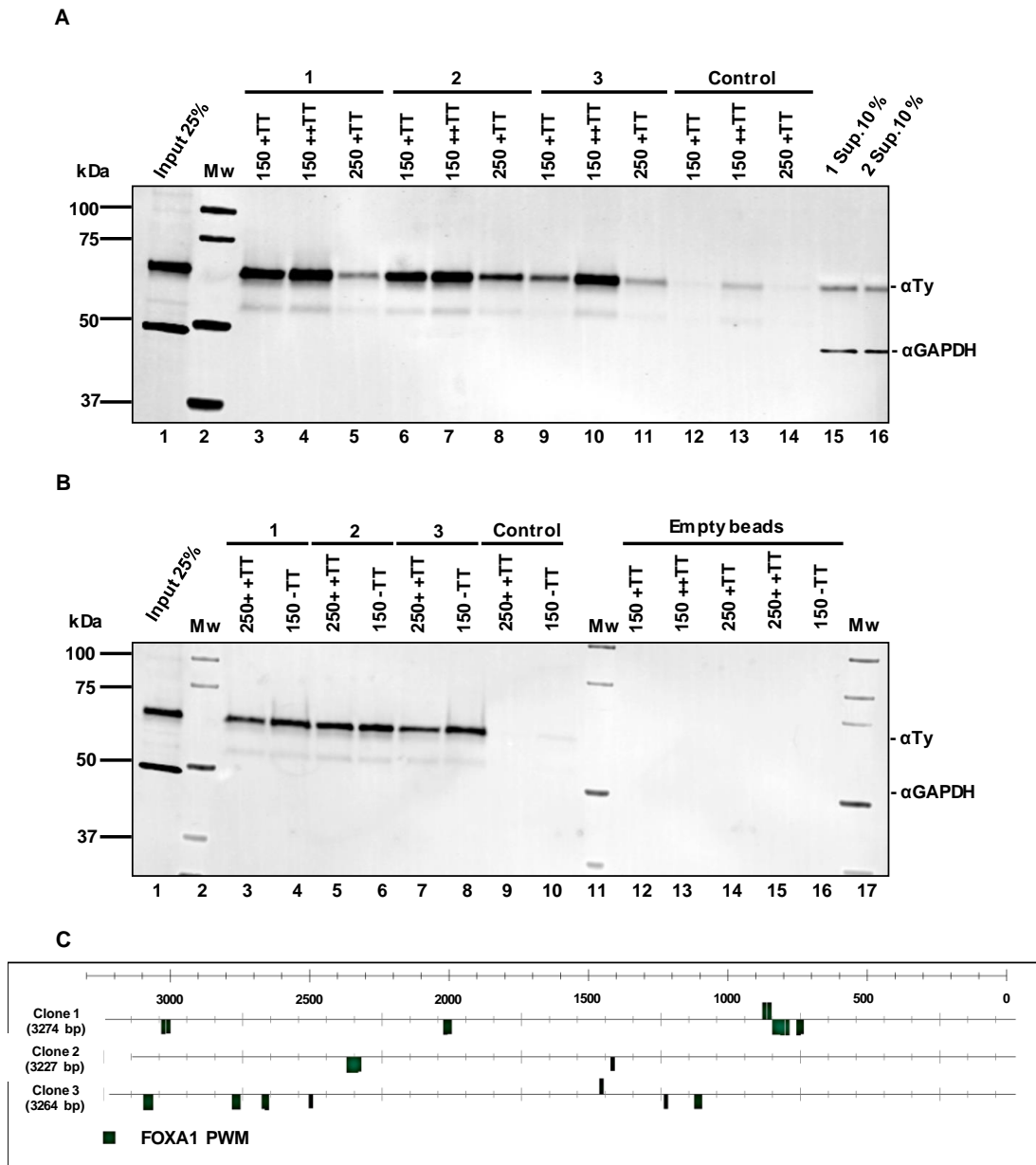
Figure 3.9. Binding specificity of FOXA1 3xTy to the different catTFRE constructs.

(A) Levels of FOXA1 3xTy pulldown by construct 1, 2, 3 and control under the indictable conditions. 25 % of the input extract was loaded in lane 1. 10 % of the supernatant after pulldown was loaded in lane 15 and 16. (B) Levels of FOXA1 3xTy pulldown by construct 1, 2, 3 and control under the indictable conditions. 25 % of the input extract was loaded in lane 1. Negative control with beads only incubated with the extract was loaded in lane 12 to 16. (C) Identified FOXA1 binding sites in the TFRE clone 1, 2 and 3. Figure was made using the CiiiDER software with default parameters. . The biotinylated Tandem-TFRE (marked as 1 to 3) and pGL4.26PvuII- control construct (1 µg DNA) were incubated with 20 µg of nuclear extracts. FOXA1 3xTy overexpressing DU145 cells (clone 10) nuclear extract was washed with 150 or 250 mM NaCl (Appendix 6.4). Various detergents and concentrations were included in washing including 0 .5 % Tween 20 and 0.1 % Triton x100 (marked as +TT) or 1 % Tween 20 and 0.5 % Triton x100 (marked as ++TT) or without extra detergents (marked as -TT) as described in Appendix 6.4.

3.4 Assessing the DNA binding of FOXA1 3xTy and FOXA1-K6R 3xTy by using the catTFRE system

It has previously been shown that SUMOylation deficient mutants of FOXA1 dysregulates its retention time on chromatin (Sutinen et al., 2014). To examine if the DNA binding features of the FOXA1-K6R mutation compared to wild-type FOXA1 3xTy using the catTFRE system we incubated nuclear extracts with biotinylated TFRE clones (clones 1 to 3) and the pGL4.26PvuII- control construct prior to washes in 150 and 250 mM NaCl. Immunoblots for the presence of FOXA1 3xTy and FOXA1-K6R 3xTy indicated a stronger binding tendency for the K6R mutants compared to wild-type FOXA1 3xTy (Figure 3.10A). Higher salt concentration reduced the binding of FOXA1 3xTy but not the SUMOylation deficient FOXA1-K6R 3xTy to the catTFRE clones (Figure 3.10B).

It is well known that FOXA1is a pioneer TF with the ability to rewire cells fate (Sérandour et al. 2011) and FOXA1 have several thousand reported binding sites in the human genome (Zhang et al. 2016). In this experiment we try to understand the role of SUMOylation on the DNA binding functions of FOXA1. In order to examine the total number of the proteins pulled down of FOXA1 3xTy and FOXA1-K3R 3xTy using the catTFRE system we incubated nuclear extracts with biotinylated TFRE clones (clones 1 to 3) and the pGL4.26PvuII- control construct. The precipitated complexes were run on an SDS PAGE and proteins visualized by Colloidal Blue stain. A large number of proteins were found in the pulldowns from both wild-type FOXA1 3xTy and FOXA1 K6R 3xTy nuclear extracts with a couple of unique bands found to be specific for each of the constructs (Figure 3.10C).

Figure 3.10. Binding specificity of FOXA1 3xTy and FOXA1 K6R 3xTy with catTFRE construct 1, 2 and 3.

Biotinylated TFRE (marked as Cons.1 to 3) and pGL4.26PvuII-control construct (1 μg DNA) pulldown from 20 μg of FOXA1 3xTy nuclear extracts (clone 10) or FOXA1 K6R 3xTy (clone 7), using 150 and 250 mM NaCl (table xxx) washing buffer.
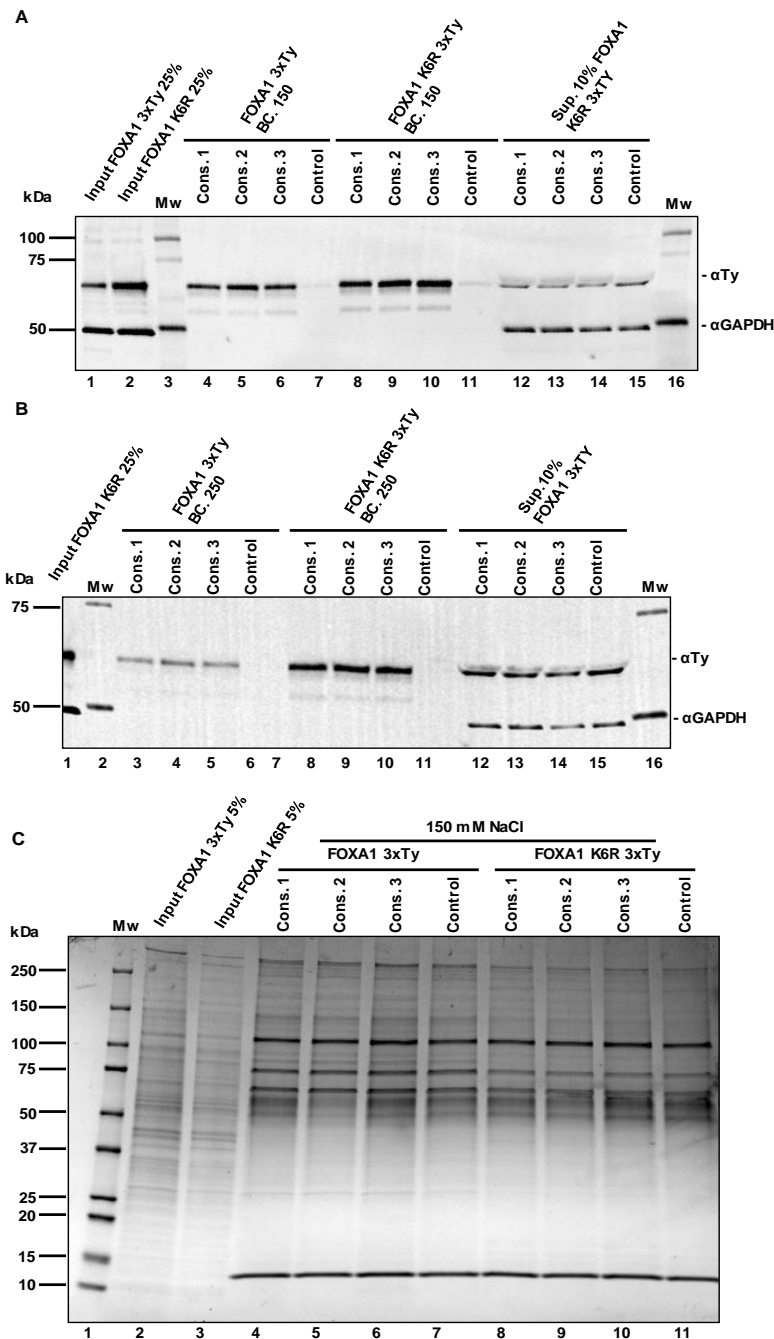
(A) Levels of FOXA1 3xTy and FOXA1 K6R 3xTy after catTFRE pulldown with indicated constructs and 150 mM NaCl buffer. 25 % from the input nuclear extract of FOXA1 3xTy and FOXA1 K6R 3xTy was loaded in lane 1 and 2 respectively. 10 % of the supernatant was loaded in lane 12 to 15. (B) Levels of FOXA1 3xTy and FOXA1 K6R 3xTy after catTFRE pull-down with indicated constructs and 250 mM NaCl buffer. 25 % from the input nuclear extract of FOXA1 K6R 3xTy was loaded in lane 1. 10 % of the supernatant remaining after pull-down was loaded in lane 12 to 15. (C) Two pulldowns of biotinylated Tandem-TFRE (marked as Cons.1 to 3) and pGL4.26PvuII- control construct (lane 7 to 11). DNA (1 μg) was incubated with 100 μg of nuclear extracts from DU145 cells overexpressing FOXA1 3xTy (clone 10) loaded in lane 4 to 7 and FOXA1 K6R 3xTy (clone 7) loaded in lane 8 to 11. 5 % from the input nuclear extract of FOXA1 and FOXA1 K6R was loaded in lane 2 and 3 respectively. 4-20 % SDS PAGE stained with Colloidal Blue stain to validate the efficiency of the catTFRE system. Black arrows indicates unique band specific to each cell line.

3.5 CTCF expression and pull down

One of the few TFs binding sites that were uniquely present in TFRE construct 1 was CTCF (Table 3.1, Figure 3.11A) and as therefore CTCF was an ideal protein to assess the specificity of the catTFRE clones. To examine CTCF binding to the TFRE clones we incubated nuclear extracts with biotinylated TFRE clones (clones 1 to 3) and the pGL4.26PvuII- control construct, washed with increasing salt concentrations and immunoblotted for the presence of CTCF (Figure 3.11B). We observed unspecific binding of CTCF to TFRE clones including the control at low salt concentration (100 mM NaCl). However, specific CTCF binding to construct 1 were achieved for the 300 mM NaCl washing condition. Washes with high salt concentration (500 mM NaCl) abolished CTCF binding to all the TFRE constructs.

It have previously been demonstrated a mutually exclusive binding of FOXA1 and CTCF to some genomic regions (Fei et al. 2019). In order to compare the *in vitro* CTCF binding to the catTFRE clones in nuclear extracts from DU145 cells stably overexpressing FOXA1 3xTy or FOXA1 K6R 3xTy, catTFRE pulldown were performed as described above. Specific CTCF binding to catTFRE clone 1 was achieved in the 250 mM NaCl washing condition (Figure 3.11.B, lower panel). Moreover, we observed increased CTCF binding in SUMOylation mutant deficient FOXA1-K6R 3xTy over expressing cells compare to wild-type FOXA1 3xTy (Figure 3.11.C).
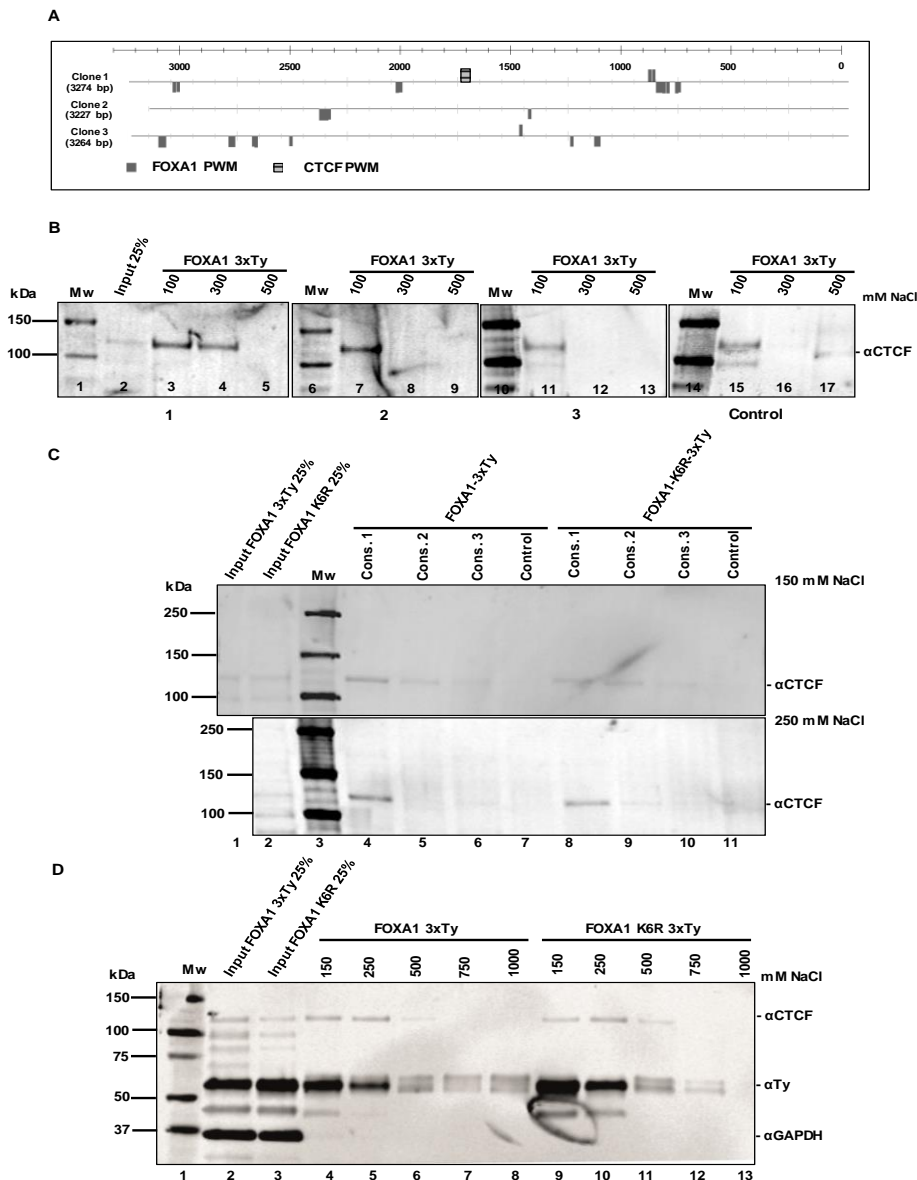
Figure 3.11. Binding specificity of CTCF with catTFRE construct 1, 2 and 3.

(A) Identified FOXA1 PWM (dark grey) and CTCF PWM (light grey) binding sites in TFRE clone 1, 2 and 3. Figure was made using the CiiiDER software with default parameters. (B) Levels of CTCF after catTFRE pulldown with indicated constructs incubated with 20 ug of nuclear extracts from FOXA1 3xTy overexpressing DU145 cells (clone 10), washed with increasing salt concentrations (100, 300, 500 mM NaCl) (Appendix 6.4). 25 % from the input nuclear extract of FOXA1 was loaded in lane 2. (C) Levels of CTCF after catTFRE pulldown with indicated constructs incubated with 20 ug of nuclear extracts from FOXA1 3xTy (clone 10) or FOXA1 K6R 3xTy (clone 7) overexpressing DU145 cells. Washed with buffer contain 150 and 250 mM NaCl (Appendix 6.4). (D) Levels of CTCF, FOXA1 3xTy and FOXA1 K6R 3xTy after catTFRE pulldown with indicated constructs. 20 ug of FOXA1 3xTy (clone 10) or FOXA1 K6R 3xTy (clone 7) nuclear extracts from overexpressing DU145 cells. Washed with increasing salt concentrations (150, 250, 500, 750 or 1000 mM NaCl) (Appendix 6.4). 25 % from the input nuclear extract of FOXA1 and FOXA1 K6R was loaded in lane 2 and 3 respectively.

52

3.5 RNA-seq analysis

RNA sequencing and the data analysis was performed as described in section 2.5. To assess the quality of the sequence before and after mapping reads to the references genome several parameters were taken into consideration. A summary of the main data quality parameters also the total number of TFs in the catTFRE system that represented more than one Reads per Kilobase of transcript, per Million mapped reads (RPKM) were shown in table 3.2. Based on the data represented in table 3.2, more than 95 % of the reads were mapped to the reference human genome and the average read depth per gene was between 326.7 and 994.7. All replicates included in this experiment showed more than 8 as the average RPKM per gene. In addition, several TFs in the catTFRE system (more than 135 TFs) were represented with more than one RPKM. However, few TFs in the catTFRE system were differentially expressed (DE) between different cell lines. Most of the expressed TFs in the catTFRE system had average RPKM around 2, which is lower than the average of the total number of the expressed genes

Table 3.2. Quantification of the read quality filtering, mapping, protein coding annotation statistics of paired-end mRNA-seq libraries.

Sample Name	Rep. No.	Total raw reads (millions)	Uniquely aligned reads (millions)	Average read depth / gene	Average RPKM [1]/ gene	TFs in catTFRE [2]	Average RPKM / TFs in catTFRE
Control	1	31.8	31.3	662.3	9.0	147	2.2
Control	2	16.2	15.9	326.7	9.4	137	1.9
Control	3	25.3	24.7	512.0	10.4	135	1.8
Control	4	47.8	47.0	947.1	9.3	158	2.2
FOXA1	1	48.0	47.2	986.7	8.9	152	2.5
FOXA1	2	31.0	30.6	616.3	9.6	150	2.2
FOXA1	3	31.9	31.3	669.0	9.1	139	2.0
FOXA1	4	35.0	34.5	707.4	9.7	140	1.8
FOXA1 K6R	1	35.3	34.8	709.7	8.9	152	2.4
FOXA1 K6R	2	43.4	42.8	863.0	9.3	154	2.1
FOXA1 K6R	3	35.6	35.0	716.4	9.5	168	2.8
FOXA1 K6R	4	49.0	48.3	994.7	8.4	153	2.3

[1] The average Reads Per Kilobase of transcript, per Million mapped reads (RPKM) for all detected genes

[2] The number of TFs in the catTFRE system that represented by more than one RPKM value

Differential Expression (DE) analysis showed large variation between the cell lines in the number of significantly differentially expressed genes (p-value < 0.1) as represented in figure 3.12. Only 41 genes were DE between non transfected DU145 cells (control) and the ectopically expressing FOXA1 3xTy cells, in which 31 genes were up regulated and only 10 genes were down regulated. The largest number of DE was found between control and FOXA1 K6R 3xTy cells, where 1353 genes were DE and slightly higher number of up regulated genes (686) compares to down regulated genes (667). In addition, 870 DE genes were found in the ectopically expressing FOXA1 3xTy relative to FOXA1 K6R 3xTy DU145 cells, with much higher number of up regulated genes (500) compares to down regulated (370) genes.
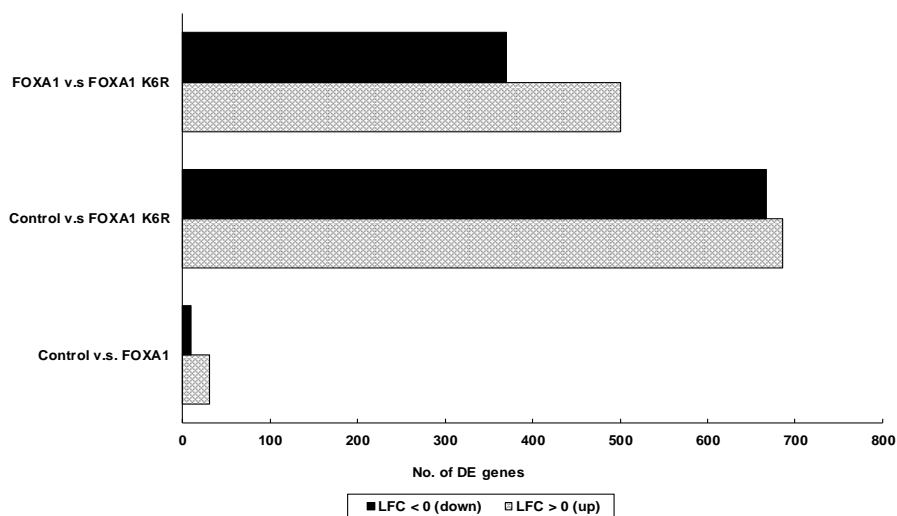
Figure 3.12. The number of significantly differentially expressed genes (p-value < 0.1) from RNA-seq analysis. Significantly DE genes in the ectopically expressing FOXA1 3xTy relative to FOXA1 K6R 3xTy DU145 cell lines and between each condition non transfected DU145 cells. Up regulated DE genes (Log2 Fold Change LFC > 0) are shown in grey and down regulates DE genes (Log2 Fold Change LFC < 0) are shown in black.

The number of DE genes relative to their log2 fold change and false discovery rate FDR for each cell line was visualized by volcano plot as represented in figure 3.13. The number of DE genes with FDR lower than 0.05 and fold change higher than 1.5 as represented in red points in the volcano plots varied between the cell lines compared with most differential genes observed for FOXA1 K6R 3XTy relative to control. The log2 fold change of top 30 significant differentially expressed genes (p-value < 0.1) was visualized using heatmap and hierarchical clustering. A clear grouping of replicates representing each cell line was found between all cell lines and the highly significant genes showed a consistent pattern over all replicates (figure 3.13)
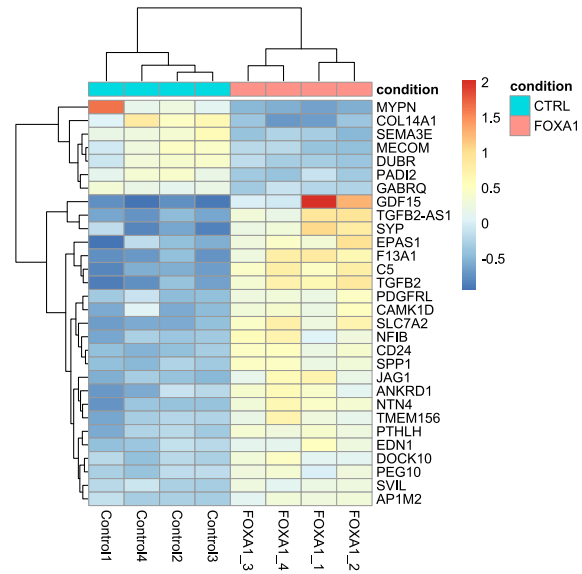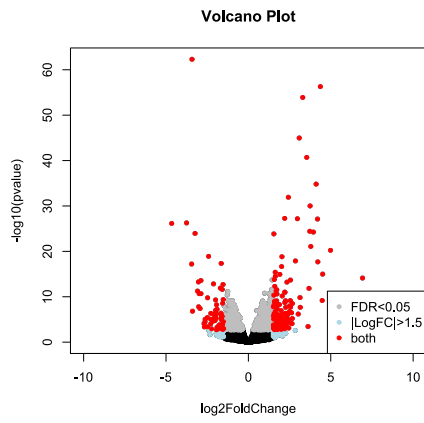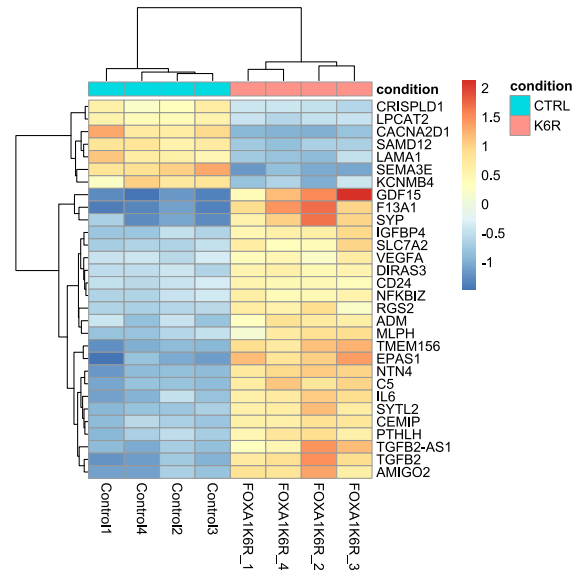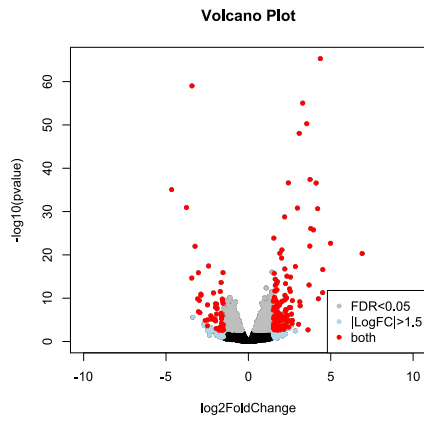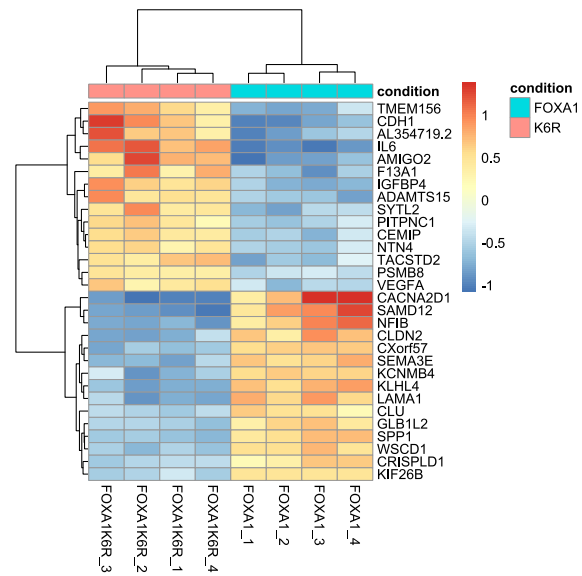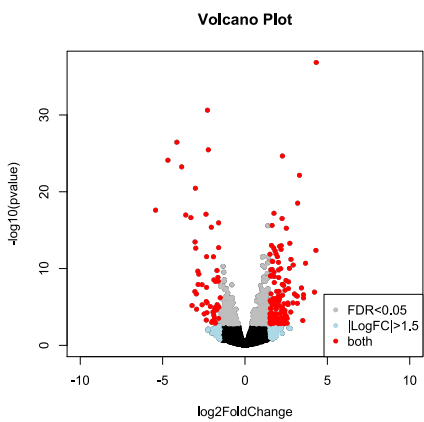
Figure 3.13. Top significant differential expressed genes in DU145 cell lines

(A and B) Differentially Expressed (DE) genes in DU145 cell line overexpressing FOXA1 3xTy or FOXA1 K6R 3xTy relative to the control DU145 cell line respectively. (C) DE genes for DU145 cell line overexpressing FOXA1 relative to FOXA1 mutant K6R. To the right is a heatmap showing hierarchical clustering of top 30 significant differentially expressed genes (p-value < 0.1), with each row representing a gene colored according to its fold change. Volcano plot shows each gene is represented by one point. Fold change is indicated on the X-axis and the statistical significance is shown on the Y-axis. Genes with false discovery rate FDR lower than 0.05 are represented by grey points, while genes with a fold change higher than 1.5 are represented by sky-blue points and genes that fit both thresholds are represented by red points.

Several important cancer related genes were upregulated in FOXA1 and FOXA1 mutant K6R compare to control such as *GDF15, TGFB, NFIB, NFKB, CD24* and *EPAS1*. Among the DE genes upregulated in FOXA1 mutant K6R cells compared to FOXA1 cells, *CDH1* and *IL6* have been shown to contribute to prostate cancer metastatic ability and radiation resistance (source: http://www.cancer-genetics.org/). To visualize the common DE genes between the cell lines, Venn diagrams was used as represented in Figure 3.14A. Only significant differentially expressed genes (p-value < 0.1) were used to illustrate this intersect. Few genes were common between cell lines, FOXA1 3xTy relative to FOXA1 K6R 3xTy and between each condition relative to the control DU145 cell line, which was 9, 7 and 12 common DE gene respectively. The total of 450 DE genes was found common between control vs. FOXA1 K6R and FOXA1 vs. FOXA1 K6R condition. A total of 39 significant DE TFs included in the catTFRE system were found common between all cell lines as represented with Venn diagram in Figure 3.14B. Of these, 13 TFs genes were common between control vs. FOXA1 K6R and FOXA1 vs. FOXA1 K6R and only *NFIB* was common DE TF genes between control vs. FOXA1 and FOXA1 vs. FOXA1 K6R cell lines. The fold change of the significant DE TFs included in the catTFRE system was represented in Figure 3.14C. All DE TFs genes show the same pattern across cell lines except *NFIB* which was up regulated in control vs. FOXA1 condition and downregulated in FOXA1 vs. FOXA1 K6R condition.

A

Cont. v.s FOXA1 (41)

Cont. v.s K6R (1121)

FOXA1 v.s K6R (870)

B

Cont. v.s FOXA1 (1)

Cont. v.s K6R (27)

FOXA1 v.s K6R (25)
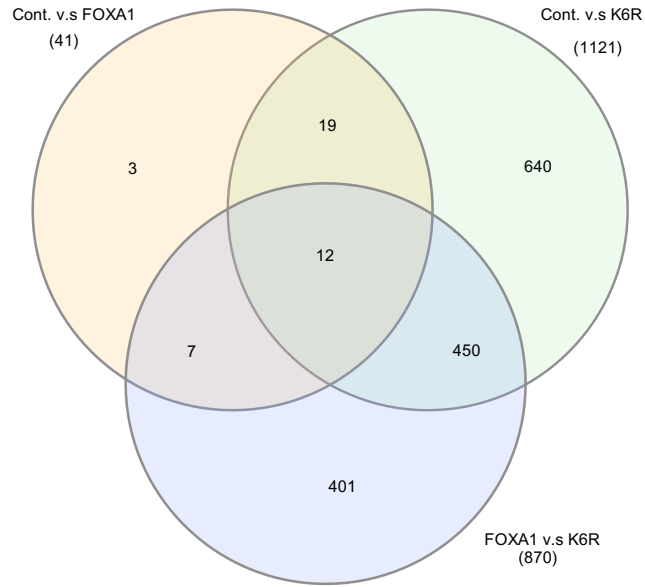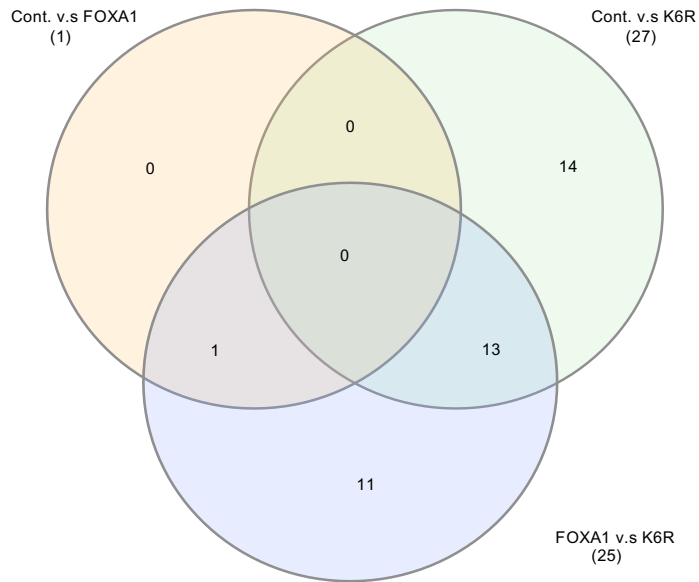
C
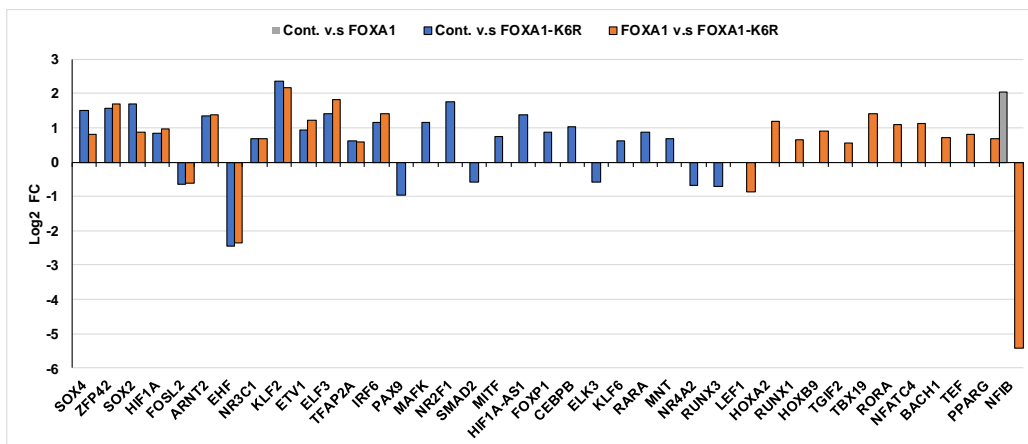
Figure 3.14. Differentially expressed transcription factors included in the catTFRE system by RNA-seq

(A) The intersection between the significant differentially expressed genes (p-value < 0.1) upon DU145 cell line overexpressing FOXA1 3xTy relative to FOXA1 K6R 3xTy(blue) and between each condition relative to the control DU145 cell line (yellow and green respectively). (B) The intersection between the significant differentially expressed (p-value < 0.1) transcription factors included in the catTFRE system upon DU145 cell line overexpressing FOXA1 3xTy relative to FOXA1 K6R 3xTy(blue) and between each condition relative to the control DU145 cell line (yellow and green respectively). (C) The fold change of the significant differentially expressed transcription factors (p-value < 0.1) included in the catTFRE system upon DU145 cell line overexpressing FOXA1 3xTy relative to FOXA1 K6R 3xTy (orange) and between each condition relative to the control DU145 cell line (grey and blue respectively).

To understand the role of the significant DE TFs included in the catTFRE system in different biological functions, a gene network analysis was conducted. The gene network of the significant DE TFs in the catTFRE system upon DU145 cell line overexpressing FOXA1 K6R relative to control was represented in Figure 3.15A with 46 node and 308 edges, while for cell line overexpress FOXA1 relative to FOXA1 K6R 3xTy was represented in Figure 3.15B with 45 node and 259 edges. Several of these TFs play an important role in cancer development and especially in prostate cancer such as *NFIB, KLF6* and *PPARG*. The network analysis showed the co-expression of these DE TFs and several prostate cancer related genes such as *FOXA2, MAF* and *FOXP4* even though these genes were not DE in any condition of inhouse RNA-seq experiment.

Gene expression

Log2 FC

− 5.34   − 0.04   5.34
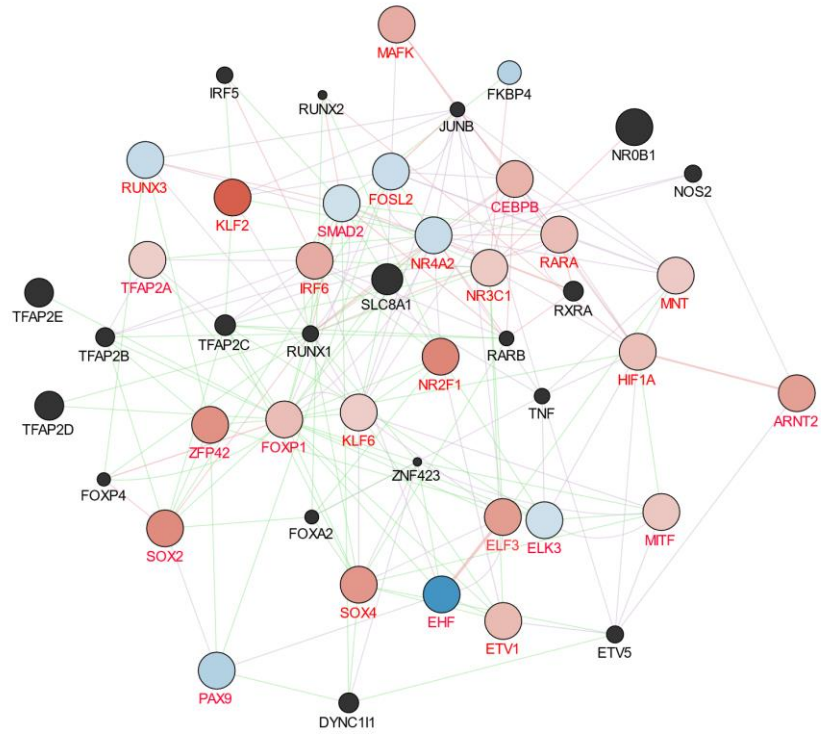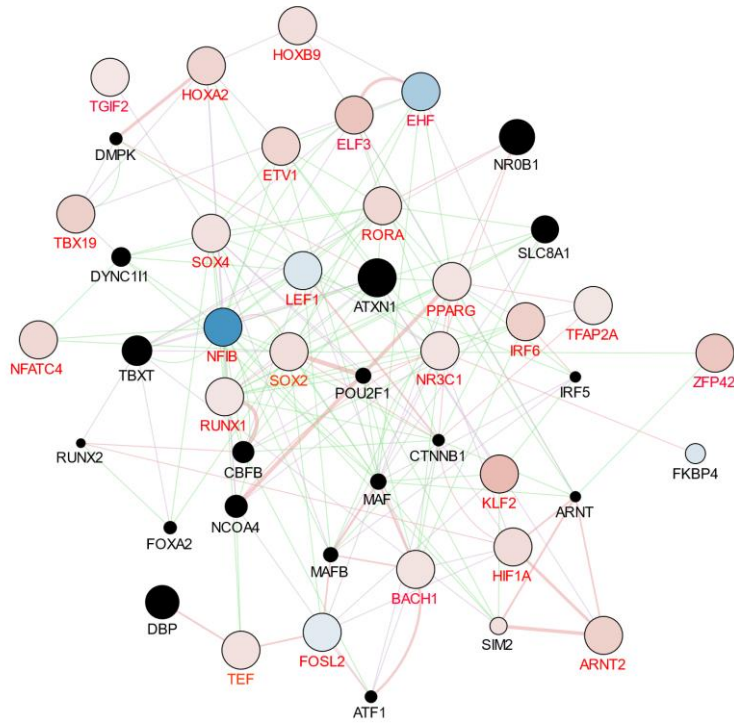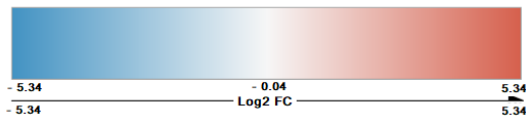
Edge Width Mapping

Normalized max weight
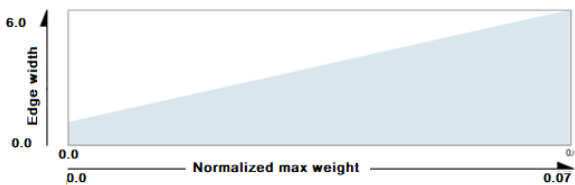
Edge Stroke Color

Edge Stroke Color	data type
	Co-expression
	Genetic Interaction
	Physical Interaction

Node Size Mapping

score
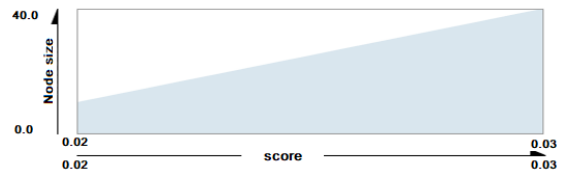
Figure 3.15. Gene network analysis of the significant differentially expressed transcription factors included in the catTFRE system

(A) Gene network analysis of the significant differentially expressed transcription factors included in the catTFRE system (red labelled) in DU145 cell line overexpressing FOXA1 mutant K6R 3xTy relative to control DU145 cell line. (B) Gene network analysis of the significant differentially expressed transcription factors included in the catTFRE system in (red labelled) DU145 cell line overexpressing FOXA1 3xTy relative to FOXA1 mutant K6R 3xTy. The network analysis shows the functional features: co-expression, genetic interactions and physical interactions. Genes are generally represented as circles colored according to their fold change up-regulated (red), down-regulated (green) and not significantly differentially expressed (black). Interactions are represented as edges (edge thickness represents normalized link weights) and edge color indicates physical interaction (magenta), co-expression (purple) and genetic interactions (light green).

4. Discussion

The Regulation of the transcription process is crucial to maintain cell biological functions and identity. Transcription factors can control the cell-specific transcription process by binding to regulatory regions spread throughout the genome (Fournier et al. 2016). Screening TFs activities at proteome scale is considered a challenge to study due to low protein expression levels and complicated gene regulation networks. In the current study, a synthetic DNA containing a TF response elements was used to evaluation of the DNA-binding activity of the TF FOXA1 at the global scale in prostate cancer cell line DU145 overexpressing FOXA1. In addition, the effect of the SUMOylation deficient FOXA1 K6R mutant on the TFs sub-transcriptome level was investigated using RNA-seq.

4.1   Establishing stable DU145 cell lines

To study FOXA1 activity and the role of the FOXA1 K6R mutant, ectopic expression in DU145 cells with the wild type or the mutated sequence were fused to elongation factor 1 alpha (EF1α) promoters were cloned. The reason we chose this promoter is that the EF-1α promoter is more stable in long-term culture, even in the absence of selection pressure (Wang et al. 2017). The most important characteristic of stable cell lines is that they are genetically similar in culture and the same genetic characteristics are presented in all cells after new generations have been produced. Cancer cell lines such as DU145 should be stable for long term studies also for investigating the role of FOXA1 on the cellular behavior.

To generate stable cell lines several factors must be taken into consideration such as the DNA vector used for transfection, which cell line to use, reagents to add and the choice of transfection method. In the current study, specific transfection method, culture media and reagents were used to ensure better response for transfection. The development of stable cell line is a laborious procedure where identification of positive clones and characterization of clonal cell lines takes many weeks. As part of this study, three single clones from each of the constructs with comparable protein expression levels were selected to ensure that we have biological replicates of generated cell lines. Similar level of expression were found between these selected clones after it was visualized by immunoblotting (Figure 3.6A). To assess the nucleoplasmic localization of the overexpressed proteins, immunolocalization studies was performed with

specific of 3xTy FOXA1 and 3xTy FOXA1 K6R (performed by M. Rogne). Both 3xTy FOXA and 3xTy FOXA1 K6R localized to the nucleus whereas 3xTy expressing cells showed a diffuse cytoplasmic localization (Figure 3.6B). The result confirms the ability of our overexpressed proteins to specifically localize to the nucleus.

In order to study the DNA binding ability of TFs nuclear protein extraction was performed. To maintain the activity of TFs all extraction steps were implemented on ice and with protease inhibitors. In all buffers used during the extraction we add Phenylmethylsulfonyl fluoride (PMSF) as serine protease inhibitor and Dithiothreitol (DTT) as reducing agent to prevent the oxidation damage. After the nuclear fraction was obtain the nuclear proteins was extracted by Bioruptor® Pico sonication system. This system contain sonication water bath with controlled temperature (4 °C) to generate indirect sonication wave in order to reduce the operating temperature and the damage to native TFs so we can have an increase of the extraction yield. A whole cell and nuclear protein extracts were used to assets the quality of proteins of interest using western blotting and SDS PAGE stained with Colloidal Blue. Both extracts (whole cell and nuclear) showed a similar level of protein expression between all selected clones. Besides, SDS PAGE showed a protein rich lysate with numerous proteins after being stained with Colloidal Blue. (Figure 3.6) All previous results made us confident enough to proceed with to test the catTFRE pulldown system

4.2   Design of the catTFRE system

In the present study, an in house design of a synthetic DNA containing a concatenated tandem array of consensus TF response elements (TFREs) covering most known TF families was used. The design of the consensus TFREs for different TF was separated to 3 constructs with size approximately 3200 bp each and all constructs were cloned into pUC57- vectors to facilitate the cloning and the amplification of the TFREs (performed by Trung Tran). However, the original design of these constructs did not take in to account the degree of similarity between TFs binding sites, so TFs with high degree of similarity in DNA binding sites should be distributed to different constructs. The catTFRE system is consider as *in vitro* assay that uses synthesis DNA template to measure the potential DNA-binding activities of TFs. Therefore, the DNA binding activity of a TF as determined by catTFRE may not consider its real activity in on the chromosome, where TF response elements may be blocked by nucleosome or by other

histone/DNA modifications. In the current TFREs design was have around 100 different TFs binding sites. This number can be reduced by splitting these long TFREs to shorter constructs to avoid unspecific binding. Also, in the current TFREs design, each DNA binding site was separated by a spacer of three nucleotides (ATC). The design of this spacer region should take in to account the degree of similarity in DNA binding sites to avoid unspecific binding. The pGL4.26PvuII- control plasmid was modified to display minimal TF binding capacity (designed by R. Eskeland). However, later experiments showed that low salt concentration during washing can results to unspecific binding of TFs to this construct.

Biotinylation of the synthesized TFRE-plasmids is an important step to facilitate TFs pulldown. Different compensation of restriction enzymes was tested to obtain specific biotinylation of TFRE construct and not the plasmid backbone. The digested TFRE construct were subjected to biotinylation using the Klenow (3′→ 5′ exo-) enzyme. In This enzyme is an N-terminal fragment of DNA polymerase I that maintains its polymerase activity, but has lost both 5′→ 3′ and 3′→ 5′ exonuclease activity. Ding et al. (2013) used PCR amplification with biotinylated primers for the biotinylation of TFRE construct. By using Klenow enzyme we avoid sequence mismatch by DNA-polymerase during the PCR, increase the amount of DNA for each biotinylation reaction and reduce the cost for the biotinylation process. Therefore, the method described in our study can be used to perform DNA biotinylation in efficient and high throughput manner.

Immobilization of the biotinylated clones onto Dynabeads™ was performed using kilobaseBINDER™ Kit to ensure specific and efferent binding of the biotinylated TFRE construct to the beads. This process allows us to separate biotinylated TFRE construct or/and it's binding TFs from the rest of the reaction mixture in rapid and affordable way. The use of kilobaseBINDER Kit is recommended for immobilizing long double stranded DNA with size >2 kb. According to the EMBL, Heidelberg group who developed this kit (Source: https://www.thermofisher.com/order/catalog/product/60101) it is possible to immobilize 70 pmoles of a 4 kb biotinylated DNA fragment and 80 pmoles of a 10 kb fragment per mg beads. To evaluation the best binding ratio between biotinylated DNA and the Dynabeads, 1μg of the DNA from each of the catTFRE constructs and a pGL4.26PvuII- control plasmid was immobilized onto increasing amounts of Dynabeads. The result of this experiment was important for all downstream TFs pulldown. 1μg of biotinylated DNA (0.46 pmols of DNA) with 4 μl of Dynabeads was the suitable ratio to reach the 100 % binding capacity of the

Dynabeads. Former findings were enough to continue with the testing for TFs pulldown using the catTFRE system. This ratio was more than the recommended ratio by the manufacture.

4.3 Validation of the catTFRE system

To optimize the TFs pulldown, biotinylated TFRE DNA were incubated with nuclear extracts as described by Ding et al. (2013). The washing buffer was adjusted as part of these experiments to achieve both specificity and sensitivity of the catTFRE system. TFs-DNA binding is highly dependent on salt concentration (Saecker et al. 2002). DNA binding ability of TFs is decreases with increasing salt concentrations. Salt concentration can affect protein conformation, subunit polymerization and the balance between the association and the dissociation of specific and non-specific protein-DNA complexes (Butcher et al. 1994). Based on our results, increasing salt concentration can increase the catTFRE specificity but reduce sensitivity. Also, detergent concentration is an important factor, with increasing detergent concentration the sensitivity of the catTFRE was increased but the specificity was reduced. Wash buffer with low salt concentrations showed unspecific binding of 3xTy FOXA1 to all TFRE constructs and to the pGL4.26PvuII- control construct. However, higher salt concentration buffer reduces the unspecific binding of FOXA1 to the control construct. Based on these experiments we found that salt concentration 150 ~ 250 mM NaCl with 0.5% 4-Nonylphenyl-polyethylene glycol (NP40) is the most suitable buffer to maintain the sensitivity and specificity of catTFRE system. According to Zhou et al. (2017) salt concentration of 200-250 mM NaCl was used to pull down TFs using catTFRE system to characterize TFs in the major organs of the mouse, suggesting we have similar conditions.

Moreover, 3xTy FOXA1 bound to all three TFRE construct under all salt concentrations. This finding directed us to inspect the designed TFRE constructs if they contain additional unintended TF binding sites. A scan for transcription factor binding sites using position frequency matrices from JASPAR database (8th release -2020) was performed on all TRFE construct using CiiiDER software (http://ciiider.com/). A total of 730 TFs binding sites were found in all TFRE constructs with more than 85% sequence similarity (Supplementary table 2). According to Ding et al. (2013) 400 and 878 TFs were identified from a single mammalian cell line and from 11 mammalian cell types respectively using a single TFRE containing only 100 TF binding sites. However, the number of TFs found in this study significantly exceeded the original design in the concatenated tandem array. Three main reasons can be the cause for

that, first is the linkers of 3-bp between TFs binding sites may create additional unspecific binding. Second is the tandem TFRE may also create additional binding sites and third is the flexibility of TFs recognition of the binding sites (Ding et al. 2013). According to Ding et al. (2013) the array designed during this study will not be suitable to find the differences in DNA binding among TFs of the same family. After the validation of the sensitivity and the specificity of the catTFRE system, enough information was available to test this system to answer a biological question about the effect of SUMOylation deficient FOXA1 K6R mutant on FOXA1 DNA binding activity.

4.4   DNA binding of FOXA1 and FOXA1 K6R

The reason for choosing FOXA1 for this study was the following: FOXA1 has been described as a pioneer transcription factor because its ability to unpack compact chromatin structures and facilitates the binding of other transcription factors such as nuclear receptors (Zhang et al. 2016). An important member of the nuclear receptors is the AR its main function is to mediate the effects of androgens through the regulation of gene transcription (Sérandour et al. 2011). The dysfunction of AR can lead to androgen insensitivity syndrome, which can contribute to prostate cancer progression and the upregulation of AR activity (Yegnasubramanian et al. 2019). FOXA1 facilitate the binding of AR to open chromatin (Parolia et al. 2019). FOXA1 alteration can contribute to the transcriptional profile of prostate cancer, whereas the high levels of FOXA1 transcription is linked to poor prognosis (Parolia et al. 2019). The mechanisms controlling the nuclear activity of FOXA1 and the role post-translational modifications such as SUMOylation on this process are still under investigation (Sutinen et al. 2014). Based on the information that SUMOylation regulates essential TFs in prostate cancer (Bettermann et al. 2012), it was relevant in the course of this study to further investigate the role of FOXA1 SUMOylation and its regulatory effect. Moreover, DU145 cell line was chosen because of its express very low levels of endogenous FOXA1 and do not express AR (Jin et al. 2013), so we can investigate the role of ectopic FOXA1 or FOXA1 K6R overexpression thought independent of the AR.

SUMOylation can modify the nuclear mobility of transcription factors and regulate its target genes. However, SUMOylation can be reversed by SUMO-specific proteases SENPs (Rosonina et al. 2017). According to Sutinen et al. (2014) three potential SUMOylation sites

at lysines K6, K267, and K389 were conserved between human, mouse, and zebrafish. To study the potential of this site we introduce a mutation at lysine K6 to mutate into arginine K6R and mutant gene was overexpressed with 3xTy tag in prostate cancer cell line DU145 (this work was part of an ongoing project in R. Eskeland group).

Based on our results we found a stronger binding tendency for the K6R mutants compared to wild-type FOXA1 3xTy on the catTFRE system (Figure 3.10A). The DNA binding of FOXA1 was reduced by using higher salt concentration, while the DNA binding of SUMOylation deficient FOXA1-K6R was more stable (Figure 3.10). The FOXA1 K6 SUMOylation site is located in a highly disorder region near FOXA1 N-terminal transactivation domain (Figure 3.3A). Also based on protein intrinsic disorder prediction model that enables us to detect the protein-binding sites of FOXA1, K6 was the only SUMOylation site with high confidence score for protein binding (Figure 3.4B).Based on these finding we suggest that K6 SUMOylation site might affect the DNA binding of FOXA1 through the interaction with its DNA binding domain. Similar result was found by Sutinen et al. (2014) that found that SUMOylation-deficient FOXA1 was more active than the wild-type FOXA1 in binding to regulatory regions of AR target genes in prostate cancer cells. However, our study only uses naked DNA to assess the DNA-binding of FOXA1 and the DU145 cell line is consider AR negative.

4.5   DNA binding of CTCF

To assess the specificity of catTFRE system, we chose to test the DNA binding activity of CTCF under different pulldown conditions, which was uniquely present in TFRE construct 1. Our results showed that washing buffer with low salt concentration reduce the specificity of all TFRE constructs towards CTCF. Besides, using washing buffer with too high salt concentration can remove any CTCF bound to the TFRE construct. Moreover, using 250 mM NaCl specific CTCF binding was achieved for construct 1. We also compare between CTCF binding activity in DU145 cells overexpressing FOXA1 and FOXA1 K6R Increasing CTCF binding was found in FOXA1 K6R compare to FOXA1 especially with higher salt concentration (Figure 3.10). Up regulation of CTCF gene expression has been associated with poor prognosis in prostate cancer, which suggests that CTCF expression can be used as a biomarker for prognosis assessment in prostate cancer (Höflmayer et al. 2020). Recently

10,000 FOXA1 and CTCF binding sites in breast and prostate cancer cells was systematically investigated using CRISPR screens by Fei et al. (2019). This study found that the essential binding of sites FOXA1 and CTCF tend to be close to each other. Also, FOXA1 and CTCF binding sites can share similar hallmarks of canonical enhancers such as strong DNase I and H3K27ac signals (Fei et al. 2019). Interestingly, the study also found that CTCF binding strength can be correlated with the essentiality of a binding site, which could be related to our finding that CTCF binding was higher in wtFOXA1. Therefore, our study suggests that the stronger DNA binding of SUMOylation-deficient FOXA1 might interfere with CTCF binding.

4.6   RNA-seq and Differential Expression analysis

Analysis based on RNA-seq data showed a large variation between the cell lines in the number of significantly differentially expressed (DE) genes. Only 41 genes were DE between non transfected DU145 cells (control) and the ectopically expressing FOXA1 3xTy cells, in which 31 were up regulated and only 10 genes were down regulated. While 1353 genes were DE between control and FOXA1 K6R 3xTy cells. This could be related to the level of the endogenous FOXA1 expression in the ectopically expressing FOXA1 3xTy cells. However, DU145 cell line show low expression level of endogenous FOXA1 (Jin et al. 2013), the ectopically expression of FOXA1 was at a proximity the same level as endogenous FOXA1. In total 870 DE genes were found in the ectopically expressing FOXA1 3xTy relative to FOXA1 K6R 3xTy DU145 cells, with much higher number of up regulated genes (500) compares to down regulated (370) genes. Several important cancer related genes were up regulated in FOXA1 mutant K6R compared to control such as growth differentiation factor 15 (*GDF15*) which associated with inhibition of cell proliferation and induction of apoptosis in prostate cancer (Zhang et al. 2019b) and *CD24* which associated with tumorigenesis and bone metastasis in prostate cancer (Weng et al. 2019). Also several of the up regulated genes were assisted with prestart cancer development such as secreted phosphoprotein 1 (*SPP1*) which is highly expressed in prostate cancer metastasis and become a novel biomarker for the diagnosis of prostate cancer metastasis and a new target for its treatment (Li et al. 2014) and interleukin 6 (*IL6*) which associated with radioresistance of prostate cancer (Chen et al. 2019). In general, FOXA1 expression is up regulated in localized prostate cancer, but in the later stage in cancer development (metastatic stage) it's significantly down regulated (Jin et al. 2013). Overexpression of FOXA1 showed no effects on DU145 cell growth; however, it reduced cell

motility (Jin et al. 2013). Moreover, the results of our experiments suggested that SUMOylation of FOXA1 can play a role in its function as pioneer TF in prostate cancer development during metastasis

Few DE genes were common between cell lines, FOXA1 3xTy relative to FOXA1 K6R 3xTy and between each condition relative to the control DU145 cell line, which was 9, 7 and 12 common DE genes, respectively. This could be because the low number of DE genes found in the ectopically expressing FOXA1 3xTy cells. However, more common DE genes (450 genes) were found between control vs. FOXA1 K6R and FOXA1 vs. FOXA1 K6R. These results showed that the effect of SUMOylation deficient FOXA1-K6R mutant on the global gene expression was more than the effect of endogenous and ectopically expressed FOXA1.

4.7   TFs Gene network analysis

In this analysis we focus on the significant DE TFs included in the catTFRE system. Only 39 significant DE TFs were found in all cell lines. catTFRE system allows for the enrichment and identification of many TFs. However, this method cannot differentiate between the single TF and TF complex DNA binding activities. In order to better understand the relation between Transcriptional co-regulators (TC) and TF complex, we may need to develop more precise methods. In this method we can use a single TFRE to identify particular TF–TC complexes. Our RNA-seq results showed that 13 DE TFs genes were common between control vs. FOXA1 K6R and FOXA1 vs. FOXA1 K6R. Only nuclear factor I B (*NFIB*) was common DE TF genes between control vs. FOXA1 and FOXA1 vs. FOXA1 K6R. *NFIB* regulates androgen receptor (*AR*) activity and *NFIB* over-expression in castration-resistant prostate cancer is correlated with the increase of *AR* activity (Nanda et al. 2019). In our study all DE TFs genes show the same pattern across cell lines. This could be due to the role of SUMOylation in protein-protein interaction between TFs and other co-regulators rather than the direct TFs effect on gene expression. Only TF NFIB was up regulated in control vs. FOXA1 condition and down regulated in FOXA1 vs. FOXA1 K6R condition. This interesting result can be due to the role of NFIB in prostate cancer castration-resistant , which consider to be regulated by FOXA1 (Parolia et al. 2019) and the role of SUMOylation in FOXA1 interaction with other TFs.

The gene network analysis was conducted to understand the role of these TFs in different biological function in repose to each condition. The first network was for significant DE TFs in the catTFRE system upon DU145 cell line overexpresses FOXA1 K6R relative to control and the network was represented with 46 node and 308 edges. The second network was for cell line overexpress FOXA1 relative to FOXA1 K6R 3xTy and the network was represented with 45 node and 259 edges. Several of these DE TFs play an important role in cancer development and especially in prostate cancer such as Kruppel like factor 6 (*KLF6*) which consider as tumor suppressor gene that is often mutated in prostate cancer (Narla et al. 2001) and the overexpression of a splice variant known as *KLF6-SV1* is associated with prostate cancer progression and metastasis in human and mouse (Narla et al. 2008), also peroxisome proliferator activated receptor gamma (*PPARG*), which acts as an oncogene its elevated levels strongly correlate with prostate cancer tumorigenesis and poor prognosis (Ahmad et al. 2016; Salgia et al. 2019). The network analysis showed the co-expression of these DE TFs and several prostate cancer related genes even though these genes were not DE in the current RNA-seq experiment. Few of these genes belong to the FOX TFs family such as *FOXA2* and *FOXP4*, which activates AR signaling in castration resistant prostate cancer and facilitates prostate cancer progression (Connelly et al. 2017; Huang et al. 2019). According to Sutinen et al. (2014) FOXA1 SUMOylation negatively regulates FOXA1 association with AR. However in our study we used AR negative cell line, Jin et al. (2013) showed that FOXA1control the expression of genes involve in cell motility and epithelial to mesenchymal cell transition through AR-independent process. Our findings also support the role of FOXA1 SUMOylation in cancer development in the absence of AR

4.8   Conclusions

The main aim of this thesis was to investigate the role of FOXA1 SUMOylation in prostate cancer by the enrichment of TFs using catTFRE system. For this purpose, we have established stable DU145 cell lines overexpressing 3xTy tag, 3xTy FOXA1 or SUMOylation deficient FOXA1 K6R mutant; also we have optimized the DNA binding conditions for the catTFRE system. Following this, I characterized the effect of FOXA1 mutant and the most important conclusion from this experiment was that FOXA1 K6R DNA binding was more stable under higher salt concentrations, which might increase FOXA1 residence time on the chromatin as also presented by other studies. Based on our results we demonstrate that catTFRE is a very useful method to study TF DNA binding activity. The main advantage of this method is the ability to screen the DNA binding activity of a large number of TFs in a robust and high throughput way. However, the main limitation of this method is that TFs from the same family or share similar DNA binding site will be difficult to distinguish their DNA binding activity. The specificity of the catTFRE system was demonstrated by analyzing the DNA binding activity of CTCF under different pulldown conditions. In addition to that, our global RNA-seq analysis revealed that several important cancer related genes were up regulated and down regulated in DU145 cell lines overexpressing FOXA1 and FOXA1 mutant K6R compare to control.

4.9   Future perspective

In the future we would like to investigate our findings using catTFRE followed with MS identification of TFs. However, the results represented in this thesis were the base to start another study to characterize TFs network in osteosarcoma using MS method. Also studying the effect of other two FOXA1 SUMOylation sites will help us to have better picture on the role of SUMOylation. Moreover, we would like to continue the investigation of the role of SUMOylation deficient FOXA1 K6R mutant by other methods such ChIP-seq or ATAC-seq to assess genome-wide chromatin accessibility as effect of this FOXA1 mutation. The integration of all these datasets will also be our focus. Therefore, we can understand the biological role of SUMOylation in functional interaction between the FOXA1 and the AR in the regulation of AR target genes in prostate cancer cell. Some of these suggestions are already performed as part of the research interest in FOXA1 SUMOylation at Eskeland group.

5 References

Ahmad, Imran, et al. (2016), 'Sleeping Beauty screen reveals Pparg activation in metastatic prostate cancer', *Proceedings of the National Academy of Sciences,* 113 (29), 8290-95.

Akhtar, Junaid, et al. (2019), 'TAF-ChIP: an ultra-low input approach for genome-wide chromatin immunoprecipitation assay', *Life science alliance,* 2 (4).

Aksnessæther, Bjørg Y, et al. (2019), 'Second cancers in radically treated Norwegian prostate cancer patients', *Acta Oncologica,* 58 (6), 838-44.

Anders, Simon, Pyl, Paul Theodor, and Huber, Wolfgang (2015), 'HTSeq—a Python framework to work with high-throughput sequencing data', *Bioinformatics,* 31 (2), 166-69.

Andersson, Robin and Sandelin, Albin (2019), 'Determinants of enhancer and promoter activities of regulatory elements', *Nature Reviews Genetics*, 1-17.

Andrews, Simon (2010), 'FastQC: a quality control tool for high throughput sequence data', (Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom).

Arnold, Cosmas D, et al. (2018), 'A high-throughput method to identify trans-activation domains within transcription factor sequences', *The EMBO journal,* 37 (16).

Benayoun, Bérénice A and Veitia, Reiner A (2009), 'A post-translational modification code for transcription factors: sorting through a sea of signals', *Trends in cell biology,* 19 (5), 189-97.

Benayoun, Bérénice A, Caburet, Sandrine, and Veitia, Reiner A (2011), 'Forkhead transcription factors: key players in health and disease', *Trends in Genetics,* 27 (6), 224-32.

Berger, Adeline and Rickman, David S (2018), 'The Role of Androgen Receptor in Prostate Cancer', *Precision Molecular Pathology of Prostate Cancer* (Springer), 345-65.

Bettermann, Kira, et al. (2012), 'SUMOylation in carcinogenesis', *Cancer letters,* 316 (2), 113-25.

Bhagwat, Anand S and Vakoc, Christopher R (2015), 'Targeting transcription factors in cancer', *Trends in cancer,* 1 (1), 53-65.

Butcher, Stephen, Hainaut, Pierre, and Milner, Jo (1994), 'Increased salt concentration reversibly destabilizes p53 quaternary structure and sequence-specific DNA binding', *Biochemical Journal,* 298 (Pt 3), 513.

Calnan, DR and Brunet, A (2008), 'The foxo code', *Oncogene,* 27 (16), 2276.

Carlsson, Peter and Mahlapuu, Margit (2002), 'Forkhead transcription factors: key players in development and metabolism', *Developmental biology,* 250 (1), 1-23.

Chen, Haitao, et al. (2018a), 'Genetic factors influencing prostate cancer risk in Norwegian men', *The Prostate,* 78 (3), 186-92.

Chen, Han, et al. (2018b), 'A pan-cancer analysis of enhancer expression in nearly 9000 patient samples', *Cell,* 173 (2), 386-99. e12.

Chen, Xiaodong, et al. (2019), 'IL-6 signaling contributes to radioresistance of prostate cancer through key DNA repair-associated molecules ATM, ATR, and BRCA 1/2', *Journal of cancer research and clinical oncology,* 145 (6), 1471-84.

Choi, Woosuk, Choe, Shawn, and Lau, Gee W (2020), 'Inactivation of FOXA2 by respiratory bacterial pathogens and dysregulation of pulmonary mucus homeostasis', *Frontiers in Immunology,* 11.

Chronis, Constantinos, et al. (2017), 'Cooperative binding of transcription factors orchestrates reprogramming', *Cell,* 168 (3), 442-59. e20.

Chupreta, Sergey, et al. (2007), 'Sumoylation-dependent Control of Homotypic and Heterotypic Synergy by the Krüppel-type Zinc Finger Protein ZBP-89', *Journal of Biological Chemistry,* 282 (50), 36155-66.

Cirillo, Lisa A and Zaret, Kenneth S (2007), 'Specific interactions of the wing domains of FOXA1 transcription factor with DNA', *Journal of molecular biology,* 366 (3), 720-24.

Cirillo, Lisa A, et al. (1998), 'Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome', *The EMBO journal,* 17 (1), 244-54.

Cirillo, Lisa Ann, et al. (2002), 'Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4', *Molecular cell,* 9 (2), 279-89.

Collas, Philippe (2010), 'The current state of chromatin immunoprecipitation', *Molecular biotechnology,* 45 (1), 87-100.

Conesa, Ana, et al. (2016), 'A survey of best practices for RNA-seq data analysis', *Genome biology,* 17 (1), 13.

Connelly, Zachary M, et al. (2017), 'Wnt/beta-catenin and Foxa2 axis activates AR signaling in castration resistant prostate cancer', (AACR).

Cooper, Geoffrey M, Hausman, Robert E, and Hausman, Robert E (2000), *The cell: a molecular approach* (10: ASM press Washington, DC).

Creyghton, Menno P, et al. (2010), 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proceedings of the National Academy of Sciences,* 107 (50), 21931-36.

Criscione, Steven W, Teo, Yee Voan, and Neretti, Nicola (2016), 'The chromatin landscape of cellular senescence', *Trends in Genetics,* 32 (11), 751-61.

Csizmok, Veronika and Forman-Kay, Julie D (2018), 'Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications', *Current opinion in structural biology,* 48, 58-67.

Darmostuk, Mariia, et al. (2015), 'Current approaches in SELEX: An update to aptamer selection technology', *Biotechnology advances,* 33 (6), 1141-61.

Darnell, James E (2002), 'Transcription factors as targets for cancer therapy', *Nature Reviews Cancer,* 2 (10), 740-49.

Davey, Rachel A and Grossmann, Mathis (2016), 'Androgen receptor structure, function and biology: from bench to bedside', *The Clinical Biochemist Reviews,* 37 (1), 3.

Ding, Chen, et al. (2013), 'Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements', *Proceedings of the National Academy of Sciences,* 110 (17), 6771-76.

Dobersch, Stephanie, Rubio, Karla, and Barreto, Guillermo (2019), 'Pioneer factors and architectural proteins mediating embryonic expression signatures in cancer', *Trends in molecular medicine*.

Dobin, Alexander, et al. (2013), 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics,* 29 (1), 15-21.

Everett, Logan, Vo, Antony, and Hannenhalli, Sridhar (2009), 'PTM-Switchboard—a database of posttranslational modifications of transcription factors, the mediating enzymes and target genes', *Nucleic acids research,* 37 (suppl_1), D66-D71.

Fei, Teng, et al. (2019), 'Deciphering essential cistromes using genome-wide CRISPR screens', *Proceedings of the National Academy of Sciences,* 116 (50), 25186-95.

Filtz, Theresa M, Vogel, Walter K, and Leid, Mark (2014), 'Regulation of transcription factor activity by interconnected post-translational modifications', *Trends in pharmacological sciences,* 35 (2), 76-85.

Fournier, Michèle, et al. (2016), 'FOXA and master transcription factors recruit Mediator and Cohesin to the core transcriptional regulatory circuitry of cancer cells', *Scientific reports,* 6, 34962.

Franz, Max, et al. (2018), 'GeneMANIA update 2018', *Nucleic acids research,* 46 (W1), W60-W64.

Fujita, Kazutoshi and Nonomura, Norio (2019), 'Role of androgen receptor in prostate cancer: a review', *The world journal of men's health,* 37 (3), 288-95.

Gao, Weiwu, et al. (2020), 'Genome-wide profiling of nucleosome position and chromatin accessibility in single cells using scMNase-seq', *Nature Protocols,* 15 (1), 68-85.

Geiss-Friedlander, Ruth and Melchior, Frauke (2007), 'Concepts in sumoylation: a decade on', *Nature reviews Molecular cell biology,* 8 (12), 947.

Gill, Grace (2005), 'Something about SUMO inhibits transcription', *Current opinion in genetics & development,* 15 (5), 536-41.

Golson, Maria L and Kaestner, Klaus H (2016), 'Fox transcription factors: from development to disease', *Development,* 143 (24), 4558-70.

Gutiérrez, Gabriel, et al. (2017), 'Subtracting the sequence bias from partially digested MNase-seq data reveals a general contribution of TFIIS to nucleosome positioning', *Epigenetics & chromatin,* 10 (1), 58.

Han, Dong, et al. (2019), 'Lysine methylation of transcription factors in cancer', *Cell death & disease,* 10 (4), 1-11.

Hanahan, Douglas and Weinberg, Robert A (2011), 'Hallmarks of cancer: the next generation', *cell,* 144 (5), 646-74.

Harbison, Christopher T, et al. (2004), 'Transcriptional regulatory code of a eukaryotic genome', *Nature,* 431 (7004), 99-104.

Harteis, Sabrina and Schneider, Sabine (2014), 'Making the bend: DNA tertiary structure and protein-DNA interactions', *International journal of molecular sciences,* 15 (7), 12335-63.

Hassanpour, SH, et al. (2018), 'A Mini Review of Cancer Treatment and Epigenetic', *J Cancer Sci Ther,* 10, 005-07.

Ho, Shiuh-Rong, et al. (2010), 'O-GlcNAcylation enhances FOXO4 transcriptional regulation in response to stress', *FEBS letters,* 584 (1), 49-54.

Höflmayer, Doris, et al. (2020), 'Expression of CCCTC-binding factor (CTCF) is linked to poor prognosis in prostate cancer', *Molecular oncology,* 14 (1), 129.

Hong, Yiling, et al. (2001), 'Regulation of heat shock transcription factor 1 by stress-induced SUMO-1 modification', *Journal of Biological Chemistry,* 276 (43), 40263-67.

Huang, Changkun, et al. (2019), 'Circular RNA circABCC4 as the ceRNA of miR-1182 facilitates prostate cancer progression by promoting FOXP4 expression', *Journal of cellular and molecular medicine,* 23 (9), 6112-19.

Hutter, Carolyn and Zenklusen, Jean Claude (2018), 'The cancer genome atlas: creating lasting value beyond its data', *Cell,* 173 (2), 283-85.

Iwafuchi-Doi, Makiko and Zaret, Kenneth S (2014), 'Pioneer transcription factors in cell reprogramming', *Genes & development,* 28 (24), 2679-92.

Iwafuchi-Doi, Makiko, et al. (2016), 'The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation', *Molecular cell,* 62 (1), 79-91.

Jaenisch, Rudolf and Bird, Adrian (2003), 'Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals', *Nature genetics,* 33 (3), 245-54.

Jenuwein, Thomas and Allis, C David (2001), 'Translating the histone code', *Science,* 293 (5532), 1074-80.

Jiang, Cizhong and Pugh, B Franklin (2009), 'Nucleosome positioning and gene regulation: advances through genomics', *Nature Reviews Genetics,* 10 (3), 161.

Jin, Hong-Jian, et al. (2013), 'Androgen receptor-independent function of FoxA1 in prostate cancer metastasis', *Cancer research,* 73 (12), 3725-36.

Johnson, Erica S (2004), 'Protein modification by SUMO', *Annual review of biochemistry,* 73 (1), 355-82.

Jones, David T and Cozzetto, Domenico (2015), 'DISOPRED3: precise disordered region predictions with annotated protein-binding activity', *Bioinformatics,* 31 (6), 857-63.

Jones, Peter A, Issa, Jean-Pierre J, and Baylin, Stephen (2016), 'Targeting the cancer epigenome for therapy', *Nature Reviews Genetics,* 17 (10), 630.

Jung, Yoon Hee, et al. (2019), 'Maintenance of CTCF-and transcription factor-mediated interactions from the gametes to the early mouse embryo', *Molecular cell,* 75 (1), 154-71. e5.

Kaestner, Klaus H, Knöchel, Walter, and Martínez, Daniel E (2000), 'Unified nomenclature for the winged helix/forkhead transcription factors', *Genes & development,* 14 (2), 142-46.

Khang, Tsung Fei and Lau, Ching Yee (2015), 'Getting the most out of RNA-seq data analysis', *PeerJ,* 3, e1360.

Klemm, Sandy L, Shipony, Zohar, and Greenleaf, William J (2019), 'Chromatin accessibility and the regulatory epigenome', *Nature Reviews Genetics,* 20 (4), 207-20.

Krueger, Felix (2015), 'Trim galore', *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files,* 516, 517.

Laissue, Paul (2019), 'The forkhead-box family of transcription factors: key molecular players in colorectal cancer pathogenesis', *Molecular cancer,* 18 (1), 5.

Lambert, Samuel A, et al. (2018), 'The human transcription factors', *Cell,* 172 (4), 650-65.

Le Lay, John and Kaestner, Klaus H (2010), 'The Fox genes in the liver: from organogenesis to functional integration', *Physiological reviews,* 90 (1), 1-22.

Lee, Jason S, Choi, Hee June, and Baek, Sung Hee (2017), 'Sumoylation and Its contribution to cancer', *SUMO Regulation of Cellular Processes* (Springer), 283-98.

Li, Bin, et al. (2007), 'FOXP3 interactions with histone acetyltransferase and class II histone deacetylases are required for repression', *Proceedings of the National Academy of Sciences,* 104 (11), 4571-76.

Li, Jun, et al. (2017), 'Structure of the forkhead domain of FOXA2 bound to a complete DNA consensus site', *Biochemistry,* 56 (29), 3745-53.

Li, Sen, et al. (2019a), 'A time-resolved proteomic analysis of transcription factors regulating adipogenesis of human adipose derived stem cells', *Biochemical and biophysical research communications,* 511 (4), 855-61.

Li, Shen, et al. (2019b), 'Cytokine-induced post-translational modifications of FOXA1 affect enhancer selection and estrogen signaling in breast cancer cells', (AACR).

Li, TQ, et al. (2014), 'The highly expressed secreted phosphoprotein 1 gene in prostate cancer metastasis: a microarray-based bioinformatic analysis', *Zhonghua nan ke xue= National journal of andrology,* 20 (11), 984-90.

Lion, Mattia, Tolstorukov, Michael Y, and Oettinger, Marjorie A (2019), 'Low-Input MNase Accessibility of Chromatin (Low-Input MACC)', *Current Protocols in Molecular Biology,* 127 (1), e91.

Liu, Yubo, et al. (2019), 'Resistance to bortezomib in breast cancer cells that downregulate Bim through FOXA1 O-GlcNAcylation', *Journal of cellular physiology*.

Love, Michael I, Huber, Wolfgang, and Anders, Simon (2014), 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology,* 15 (12), 550.

Magnani, Luca, Eeckhoute, Jérôme, and Lupien, Mathieu (2011), 'Pioneer factors: directing transcriptional regulators within the chromatin environment', *Trends in Genetics,* 27 (11), 465-74.

Mann, Matthias (2006), 'Functional and quantitative proteomics using SILAC', *Nature reviews Molecular cell biology,* 7 (12), 952-58.

Mansoori, Behzad, et al. (2017), 'The different mechanisms of cancer drug resistance: a brief review', *Advanced pharmaceutical bulletin,* 7 (3), 339.

Mariño-Ramírez, Leonardo, et al. (2005), 'Histone structure and nucleosome stability', *Expert review of proteomics,* 2 (5), 719-29.

Mayran, Alexandre and Drouin, Jacques (2018), 'Pioneer transcription factors shape the epigenetic landscape', *Journal of Biological Chemistry,* 293 (36), 13795-804.

Mieczkowski, Jakub, et al. (2016), 'MNase titration reveals differences between nucleosome occupancy and chromatin accessibility', *Nature communications,* 7, 11485.

Mirecka, Alicja, Morawiec, Zbigniew, and Wozniak, Katarzyna (2016), 'Genetic Polymorphism of SUMO-Specific Cysteine Proteases− SENP1 and SENP2 in Breast Cancer', *Pathology & Oncology Research,* 22 (4), 817-23.

Mo, Yin-Yuan, et al. (2005), 'A role for Ubc9 in tumorigenesis', *Oncogene,* 24 (16), 2677.

Molværsmyr, Ann-Kristin, et al. (2010), 'A SUMO-regulated activation function controls synergy of c-Myb through a repressor–activator switch leading to differential p300 recruitment', *Nucleic acids research,* 38 (15), 4970-84.

Müller, Manuel M (2017), 'Post-translational modifications of protein backbones: unique functions, mechanisms, and challenges', *Biochemistry,* 57 (2), 177-85.

Muratani, Masafumi and Tansey, William P (2003), 'How the ubiquitin–proteasome system controls transcription', *Nature reviews Molecular cell biology,* 4 (3), 192.

Nanda, Jagpreet S, et al. (2019), 'Nuclear factor I/B increases in prostate cancer to support androgen receptor activation', *bioRxiv*, 684472.

Narla, Goutham, et al. (2001), 'KLF6, a candidate tumor suppressor gene mutated in prostate cancer', *Science,* 294 (5551), 2563-66.

Narla, Goutham, et al. (2008), 'KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis', *The Journal of clinical investigation,* 118 (8), 2711-21.

Nicolas, Damien, Phillips, Nick E, and Naef, Felix (2017), 'What shapes eukaryotic transcriptional bursting?', *Molecular BioSystems,* 13 (7), 1280-90.

Obsil, T and Obsilova, V (2008), 'Structure/function relationships underlying regulation of FOXO transcription factors', *Oncogene,* 27 (16), 2263.

Ohuchi, Shoji (2012), 'Cell-SELEX technology', *BioResearch open access,* 1 (6), 265-72.

Ong, Shao-En (2012), 'The expanding field of SILAC', *Analytical and bioanalytical chemistry,* 404 (4), 967-76.

Pan, Yongping, et al. (2010), 'Mechanisms of transcription factor selectivity', *Trends in Genetics,* 26 (2), 75-83.

Papaccio, Federica, et al. (2017), 'Concise review: cancer cells, cancer stem cells, and mesenchymal stem cells: influence in cancer development', *Stem cells translational medicine,* 6 (12), 2115-25.

Parolia, Abhijit, et al. (2019), 'Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer', *Nature,* 571 (7765), 413-18.

Pichler, Andrea, et al. (2017), 'SUMO conjugation–a mechanistic view', *Biomolecular concepts,* 8 (1), 13-36.

Rabellino, Andrea and Khanna, Kum Kum (2020), 'The implication of the SUMOylation pathway in breast cancer pathogenesis and treatment', *Critical Reviews in Biochemistry and Molecular Biology*, 1-17.

Rebbeck, Timothy R (2017), 'Prostate cancer genetics: variation by race, ethnicity, and geography', *Seminars in radiation oncology* (27: Elsevier), 3-10.

Rosonina, Emanuel, et al. (2017), 'Regulation of transcription factors by sumoylation', *Transcription,* 8 (4), 220-31.

Rowley, M Jordan, et al. (2017), 'Evolutionarily conserved principles predict 3D chromatin organization', *Molecular cell,* 67 (5), 837-52. e7.

Saecker, Ruth M and Record Jr, M Thomas (2002), 'Protein surface salt bridges and paths for DNA wrapping', *Current opinion in structural biology,* 12 (3), 311-19.

Salgia, Meghan M, et al. (2019), 'Different roles of peroxisome proliferator-activated receptor gamma isoforms in prostate cancer', *American journal of clinical and experimental urology,* 7 (3), 98.

Sarkar, Sibaji, et al. (2013), 'Cancer development, progression, and therapy: an epigenetic overview', *International journal of molecular sciences,* 14 (10), 21087-113.

Seeler, Jacob-Sebastian and Dejean, Anne (2017), 'SUMO and the robustness of cancer', *Nature Reviews Cancer,* 17 (3), 184.

Sekiya, Takashi, et al. (2009), 'Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA', *Genes & development,* 23 (7), 804-09.

Sérandour, Aurélien A, et al. (2011), 'Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers', *Genome research,* 21 (4), 555-65.

Shen, Michael M and Abate-Shen, Cory (2010), 'Molecular genetics of prostate cancer: new prospects for old challenges', *Genes & development,* 24 (18), 1967-2000.

Shen, Ning, et al. (2018), 'Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding', *Cell systems,* 6 (4), 470-83. e8.

Simicevic, Jovan and Deplancke, Bart (2017), 'Transcription factor proteomics—Tools, applications, and challenges', *Proteomics,* 17 (3-4), 1600317.

Slattery, Matthew, et al. (2014), 'Absence of a simple code: how transcription factors read the genome', *Trends in biochemical sciences,* 39 (9), 381-99.

Snider, Justin, et al. (2019), 'Pulse SILAC Approaches to the Measurement of Cellular Dynamics', *Advancements of Mass Spectrometry in Biomedical Research* (Springer), 575-83.

Sumanasuriya, Semini and De Bono, Johann (2018), 'Treatment of advanced prostate cancer—A review of current therapies and future promise', *Cold Spring Harbor perspectives in medicine,* 8 (6), a030635.

Sutinen, Päivi, et al. (2014), 'Nuclear mobility and activity of FOXA1 with androgen receptor are regulated by SUMOylation', *Molecular endocrinology,* 28 (10), 1719-28.

Swinstead, Erin E, et al. (2016), 'Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors', *Bioessays,* 38 (11), 1150-57.

Thomas, Mary C and Chiang, Cheng-Ming (2006), 'The general transcription machinery and general cofactors', *Critical reviews in biochemistry and molecular biology,* 41 (3), 105-78.

Tomasi, Maria Lauda and Ramani, Komal (2018), 'SUMOylation and phosphorylation cross-talk in hepatocellular carcinoma', *Translational gastroenterology and hepatology,* 3.

Torres, Sofía, et al. (2018), 'Proteomic characterization of transcription and splicing factors associated with a metastatic phenotype in colorectal cancer', *Journal of proteome research,* 17 (1), 252-64.

Tuerk, Craig and Gold, Larry (1990), 'Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase', *science,* 249 (4968), 505-10.

van der Horst, Armando, et al. (2006), 'FOXO4 transcriptional activity is regulated by monoubiquitination and USP7/HAUSP', *Nature cell biology,* 8 (10), 1064.

Vaquerizas, Juan M, et al. (2009), 'A census of human transcription factors: function, expression and evolution', *Nature Reviews Genetics,* 10 (4), 252.

Venter, J Craig, et al. (2001), 'The sequence of the human genome', *science,* 291 (5507), 1304-51.

Venters, Bryan J and Pugh, B Franklin (2009), 'How eukaryotic genes are transcribed', *Critical reviews in biochemistry and molecular biology,* 44 (2-3), 117-41.

Viola, Ivana L and Gonzalez, Daniel H (2016), 'Methods to Study Transcription Factor Structure and Function', *Plant Transcription Factors* (Elsevier), 13-33.

Vogelstein, Bert and Kinzler, Kenneth W (2004), 'Cancer genes and the pathways they control', *Nature medicine,* 10 (8), 789-99.

Völker-Albert, Moritz Carl, et al. (2018), 'Detection of Histone Modification Dynamics during the Cell Cycle by MS-Based Proteomics', *Histone Variants* (Springer), 61-74.

Wade, Joseph T (2015), 'Mapping transcription regulatory networks with ChIP-seq and RNA-seq', *Prokaryotic systems biology* (Springer), 119-34.

Wagner, Meike, et al. (2016), 'Chromatin immunoprecipitation assay to identify genomic binding sites of regulatory factors', *Estrogen Receptors* (Springer), 53-65.

Walsh, Christopher T, Garneau-Tsodikova, Sylvie, and Gatto Jr, Gregory J (2005), 'Protein posttranslational modifications: the chemistry of proteome diversifications', *Angewandte Chemie International Edition,* 44 (45), 7342-72.

Wang, Guocan, et al. (2018a), 'Genetics and biology of prostate cancer', *Genes & development,* 32 (17-18), 1105-40.

Wang, Mangyuan, et al. (2019), 'Proteomic profiling of key transcription factors in the process of neonatal mouse cardiac regeneration capacity loss', *Cell biology international,* 43 (12), 1435-42.

Wang, Xiaoyin, et al. (2017), 'The EF-1α promoter maintains high-level transgene expression from episomal vectors in transfected CHO-K1 cells', *Journal of cellular and molecular medicine,* 21 (11), 3044-54.

Wang, Yunzhi, et al. (2018b), 'A proteomics landscape of circadian clock in mouse liver', *Nature communications,* 9 (1), 1-16.

Warde-Farley, David, et al. (2010), 'The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function', *Nucleic acids research,* 38 (suppl_2), W214-W20.

Wei, Bei, et al. (2017), 'Strong binding activity of few transcription factors is a major determinant of open chromatin', *bioRxiv*, 204743.

Wei, Junying, et al. (2019), 'Proteomic Investigations of Transcription Factors Critical in Geniposide-Mediated Suppression of Alcoholic Steatosis and in Overdose-Induced Hepatotoxicity on Liver in Rats', *Journal of proteome research,* 18 (11), 3821-30.

Weigel, Detlef, et al. (1989), 'The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the Drosophila embryo', *Cell,* 57 (4), 645-58.

Weinberg, Robert A (2013), *The Biology of Cancer: Second International Student Edition* (WW Norton & Company).

Weng, Ching-Chieh, et al. (2019), 'Mutant Kras-induced upregulation of CD24 enhances prostate cancer stemness and bone metastasis', *Oncogene,* 38 (12), 2005-19.

Wingender, Edgar, et al. (2015), 'TFClass: a classification of human transcription factors and their rodent orthologs', *Nucleic acids research,* 43 (D1), D97-D102.

Wotton, Karl R and Shimeld, Sebastian M (2006), 'Comparative genomics of vertebrate Fox cluster loci', *BMC genomics,* 7 (1), 271.

Yamagata, Kazuyuki, et al. (2008), 'Arginine methylation of FOXO transcription factors inhibits their phosphorylation by Akt', *Molecular cell,* 32 (2), 221-31.

Yamaguchi, Noritaka, et al. (2017), 'Tyrosine phosphorylation of the pioneer transcription factor FoxA1 promotes activation of estrogen signaling', *Journal of cellular biochemistry,* 118 (6), 1453-61.

Yamamoto, Hideki, et al. (2003), 'Sumoylation is involved in β-catenin-dependent activation of Tcf-4', *The EMBO journal,* 22 (9), 2047-59.

Yang, Yanfang, et al. (2017), 'Protein SUMOylation modification and its associations with disease', *Open biology,* 7 (10), 170167.

Yegnasubramanian, Srinivasan, De Marzo, Angelo M, and Nelson, William G (2019), 'Prostate Cancer Epigenetics: From Basic Mechanisms to Clinical Implications', *Cold Spring Harbor perspectives in medicine,* 9 (4), a030445.

Zaret, Kenneth S and Mango, Susan E (2016), 'Pioneer transcription factors, chromatin dynamics, and cell fate control', *Current opinion in genetics & development,* 37, 76-81.

Zhang, Gaihua, et al. (2016), 'FOXA1 defines cancer cell specificity', *Science advances,* 2 (3), e1501473.

Zhang, Jingjing, et al. (2017), 'An integrated approach to identify critical transcription factors in the protection against hydrogen peroxide-induced oxidative stress by Danhong injection', *Free Radical Biology and Medicine,* 112, 480-93.

Zhang, Lirong, et al. (2018a), 'Revealing transcription factor and histone modification co-localization and dynamics across cell lines by integrating ChIP-seq and RNA-seq data', *BMC genomics,* 19 (10), 914.

Zhang, Qingmiao, et al. (2018b), 'The reduction in FOXA 2 activity during lung development in fetuses from diabetic rat mothers is reversed by Akt inhibition', *FEBS open bio,* 8 (10), 1594-604.

Zhang, Siqi, et al. (2019a), 'Identification of the O-GalNAcylation site (s) on FOXA1 catalyzed by ppGalNAc-T2 enzyme in vitro', *Biochemical and biophysical research communications,* 514 (1), 157-65.

Zhang, Wenbo, et al. (2019b), 'Role of GDF15 in methylseleninic acid-mediated inhibition of cell proliferation and induction of apoptosis in prostate cancer cells', *PloS one,* 14 (9).

Zhao, Xiaolan (2018), 'SUMO-mediated regulation of nuclear functions and signaling processes', *Molecular cell,* 71 (3), 409-18.

Zhou, Quan, et al. (2017), 'A mouse tissue transcription factor atlas', *Nature communications,* 8 (1), 1-15.

6 Appendix

6.1 Appendix 1: Abbreviations

List of commonly used abbreviations

bp	Base pairs
AR	Androgen receptor
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
BSA	Bovine serum albumin
catTFRE	Concatenated tandem array of the consensus TFREs
ChIP	Chromatin immunoprecipitation
CTCF	CCCTC-binding factor
DAPI	4′,6-diamidino-2-phenylindole
DBD	DNA binding domain
DE	Differential expression
DMEM	Dulbecco's modified eagle medium
DMSO	Dimethyl sulfoxide
dNTP	Deoxynucleotide
dsDNA	Double stranded DNA
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol-bis(2-aminoethylether)-N,N,N′,N′-tetraacetic acid
EtBr	Ethidium bromide
FC	Fold change
FBS	Foetal bovine serum
FDR	False discovery rate
FOX	Forkhead box
gDNA	Genomic DNA
H3K9	Histone 3 lysine 9
kDa	Kilo Dalton
LB	Lysogeny broth
mRNA	Messenger RNA
MS	Mass spectrometry
NEAA	Non-essential amino acids

NP40	4-Nonylphenyl-polyethylene glycol
O/N	Overnight
OD	Optical density
ORF	Open reading frame
P/S	Penicillin/Streptomycin
PBS	Phosphate-Buffered Saline
PCR	Polymerase chain reaction
PTM	Post-translational modification
PVDF	Polyvinylidene fluoride
RCF	Relative centrifugal force
RE	Restriction enzyme
RPM	Revolutions per minute
RPKM	Reads Per Kilobase of transcript, per Million mapped reads
RT	Room temperature
SDS	Sodium dodecyl sulphate
SNP	Single nucleotide polymorphism
TAD	Trans-activation domain
TAE	Tris-acetate-EDTA
TBE	Tris-borate-EDTA
TBS-T	Tris-buffered saline- Tween
TE	Tris-EDTA
TFRE	Transcription factor response elements
TF	Transcription factor
TGX	Tris-glycine extended
UV	Ultraviolet
WT	Wild-type

6.2 Appendix 2: Supplementary tables

Supplementary Table 6.2.1: List of TFs included in each catTFRE clone

Clone	TF	Start	End	Seq.
1	ALX3	475	484	CTAATTAAA
1	IRF9	501	515	ACGAAACCGAAACT
1	ONECUT3	537	550	AAAAATCAATAAT
1	MAFK	571	585	TGAGTCAGCAATTT
1	HOXD12	607	617	GTCGTAAAAA
1	HOXC12	635	645	GTCGTAAAAA
1	JUND(Var.2)	663	677	AAGATGATGTCATC
1	LHX6	699	708	CTAATTAGC
1	MEF2C	725	739	TGCTAAAAATAGAA
1	MINT	761	770	CCACGTGCC
1	PDX1	787	794	TAATTAG
1	ARNT:HIF1A	809	816	GACGTGC
1	KLF16-SP3	831	841	CCACGCCCCC
1	OTX2	859	866	TAATCCT
1	NFKB2	881	893	GGGGATTCCCCT
1	MEIS3	913	920	TGACAGG
1	PAX6	935	948	TCACGCATGAGTT
1	LBX1	969	976	TAATTAG
1	KLF14	991	1004	GCCACGCCCCCTT
1	NR2F1	1025	1037	AAAGGTCAAGGG
1	JDP2	1057	1065	TGACTCAT
1	NFYB	1081	1095	AATGGACCAATCAG
1	ZIC1	1117	1130	ACCCCCCGCTGTG
1	TBX2	1151	1161	AAGGTGTGAA
1	POU2F1	1179	1190	ATATGCAAATT
1	MEOX1	1209	1218	CTAATTAAC
1	MYBL2	1235	1249	ACCGTTAAACGGTC
1	MZF1	1271	1276	GGGGA
1	NEUROG2	1289	1298	ACATATGTC
1	NFATC3	1315	1324	TTTTCCATT
1	NKX3-2	1341	1349	CCACTTAA
1	TP63	1365	1382	ACATGTTGGGACATGTC
1	ZEB1	1407	1415	CTCACCTG
1	ZIC4	1431	1445	ACCCCCCGCTGTGC
1	ATF4	1467	1479	GATGATGCAATA
1	ATF7	1499	1512	GATGACGTCATCG
1	BARHL2	1533	1542	CTAAACGGT
1	BARX1	1559	1566	CAATTAG
1	BATF::JUN	1581	1591	AAATGACTCA
1	BCL6B	1609	1625	GCTTTCTAGGAATTCA

1	BHLHE22::OLIG3	1649	1658	CCATATGTT
1	BHLHE23	1675	1686	AACATATGTTT
1	BHLHE40	1705	1714	TCACGTGAC
1	BHLHE41	1731	1740	TCACGTGAC
1	CDX1	1757	1765	CAATAAAA
1	CDX2	1781	1791	AGCCATAAAA
1	CEBPA	1809	1819	TTGCACAATA
1	CENPB	1837	1851	CCGCATACAACGAA
1	CLOCK	1873	1882	ACACGTGTT
1	CREB1	1899	1906	GACGTCA
1	CREB3	1921	1934	TGCCACGTCATCA
1	CTCF	1955	1973	GGCCACCAGGGGGCGCTA
1	CUX::CUX2	1999	2008	AATCGATAA
1	DBP::TEF	2025	2036	ATTACGTAACA
1	DLX6	2055	2062	CAATTAC
1	DMRT3	2077	2087	ATGTATCAAT
1	DUX4	2105	2115	AATTTAATCA
1	E2F1	2133	2144	TTGGCGCCAAA
1	E2F2	2163	2178	AAATGGCGCCATTTT
1	E2F4::E2F6	2201	2211	GGCGGGAAGG
1	E2F7	2229	2242	TTTCCCGCCAAAA
1	EBF1	2263	2276	TTCCCAAGGGAAT
1	EGR1	2297	2310	CCCCGCCCCCGCC
1	EGR2	2331	2341	CGCCCACGCA
1	ELF1::ELF4	2359	2370	ACCCGGAAGTG
1	ELK1	2389	2398	CCGGAAGTG
1	ELK4	2415	2425	CACTTCCGGC
1	EMX1:HOXA2	2443	2452	CTAATTACC
1	EN1	2469	2476	TAATTAG
1	EOMES	2491	2503	AGGTGTGAAAAT
1	ESR1	2523	2539	AGGTCACGGTGACCTG
1	ESRRB	2563	2573	CAAGGTCATA
1	ESX1	2591	2600	CCAATTAAC
1	ETV2	2617	2627	ACCGGAAATA
1	ETV6	2645	2654	GCGGAAGTG
1	EVX1::EVX2	2671	2680	GTAATTAGC
1	EWSR1-FLI1	2697	2714	GAAGGAAGGAAGGAAGG
1	FIGLA	2739	2748	CCACCTGTT
1	FOS::JUN	2765	2771	GACTCA
1	FOSL2::JUNB	2785	2795	GATGACTCAT
1	FOXA1	2813	2827	CCATGTTTACTTTG
1	FOXC1::FOXB1	2849	2859	ATGTAAATAT
1	FOXF2	2877	2890	AAACGTAAACAAT
1	FOXH1	2911	2921	CCAATCCACA

1	FOXO3:FOCG1	2939	2946	TAAACAA
1	GATA1::TAL1	2961	2978	TTATCTGTGAGGAGCAG
1	GATA3::GATA5	3003	3013	GATAAGATCT
1	GBX2	3031	3040	CCAATTAGC
1	GCM1	3057	3067	ATGCGGGTAC
1	GLI2	3085	3096	CGACCACACTG
1	GLIS1	3115	3130	GACCCCCCACGAAGC
1	GLIS2	3153	3166	ACCCCCCGCGAAG
1	GMEB2	3187	3194	TACGTAA
1	GRHL1	3209	3220	AAACCGGTTTT
1	GSC	3239	3248	CTAATCCCC
1	GSC2	3265	3274	CTAATCCGC
1	GSX1	3291	3300	CTAATTAAA
1	HES5	3317	3328	GGCACGTGCCA
1	HESX1	3347	3356	CTAATTGGC
1	HEY1-2	3373	3382	ACACGTGCC
1	HIC2	3399	3407	TGCCCACC
1	HINFP	3423	3434	AACGTCCGCGG
1	HMBOX1	3453	3462	CTAGTTAAC
1	HNF1A	3479	3493	GTTAATGATTAACT
1	HNF4A	3515	3529	TGGACTTTGGACTC
1	HNF4A	3551	3565	GGGTCAAAGTCCAA
1	HOXA10	3587	3597	GTAATAAAAA
1	HOXA13-HOXD13	3615	3624	CAATAAAAA
1	ZNF740	3641	3650	CCCCCCCAC
1	HOXA5	3667	3674	ACTAATT
2	HOXB13	475	484	CAATAAAAC
2	E2F3	501	518	AAAATGGCGCCATTTT
2	CREB3L1	543	556	TGCCACGTCATCA
2	HES7	577	588	GGCACGTGCCA
2	HNF1B	607	619	TTAATGATTAAC
2	DUXA	639	651	TAATTTAATCAA
2	BATF3	671	684	GATGACGTCATCA
2	GSX2	705	714	CTAATTAAA
2	EGR4	731	746	TACGCCCACGCATTT
2	HNF4G	769	783	GAGTCCAAAGTCCA
2	ELF3	805	817	ACCGGAAGTAA
2	HLF	837	848	ATTACGTAACC
2	ESRRB	866	879	GGTCAAGGTCATA
2	ESRRB	883	895	GTCAAGGTCATA
2	ZIC3	899	913	ACCCCCCGCTGCGC
2	TEAD3	935	942	CATTCCA
2	TFAP2B/C	957	967	GCCTCAGGCA
2	TBX19	985	1004	TTCACACCTAGGTGTGAAA

84

2	POU3F1	1031	1042	TATGCAAATTA
2	POU4F2-3	1061	1076	TGCATAATTAATGAG
2	RELA	1099	1108	GGAATTTCC
2	RORA(Var.2)	1125	1138	ATAAGTAGGTCAA
2	RUNX3	1159	1168	AACCGCAAA
2	RXRB-RXRG	1185	1198	GGGTCAAAGGTCA
2	FOXC2	1219	1230	AAGTAAACAAA
2	FOXL1	1249	1255	TAAACA
2	SCRT2	1269	1281	TGCAACAGGTGG
2	SOX10	1301	1306	TTTGT
2	SP2	1319	1333	CCCCGCCCCCTCCC
2	SP8	1355	1366	CCACGCCCACT
2	SPIC	1385	1398	AAAAGAGGAAGTA
2	FOS	1419	1429	GTGACTCATT
2	ZBTB7B/C	1447	1458	CGACCACCGAA
2	GCM2	1477	1486	ATGCGGGTA
2	GLIS3	1503	1516	ACCCCCCACGAAG
2	VSX1	1537	1544	TAATTAT
2	VSX1	1548	1554	AATTAT
2	ZFX	1558	1570	CCTAGGCCTCGG
2	HOXC10	1590	1599	TCGTAAAAT
2	HOXC11	1616	1626	GTCGTAAAAT
2	HOXD11	1644	1653	TCGTAAAAA
2	HSF1-2-4	1670	1682	TCTAGAACGTTC
2	ID4	1702	1711	ACACCTGTC
2	INSM1	1728	1739	GTCAGGGGGCG
2	IRF1	1758	1778	TTTACTTTCACTTTCACTTT
2	IRF7	1806	1819	CGAAAGCGAAAGT
2	IRF8	1840	1853	CGAAACCGAAACT
2	ISL2	1874	1881	CACTTAA
2	ISX	1896	1903	TAATTAA
2	TBX21	1918	1927	AGGTGTGAA
2	TCF3	1944	1953	ACACCTGCT
2	JUN	1970	1982	AGATGATGTCAT
2	JUND	2002	2012	GTGACTCATC
2	KLF4	2030	2040	GGGTGTGGCC
2	KLF13	2058	2075	TGCCACGCCCCTTTTG
2	KLF5	2100	2109	CCCtGCCCC
2	LBX2	2126	2135	CCAATTAGC
2	LEF1	2152	2166	AAGATCAAAGGGTT
2	LHX2	2188	2197	CTAATTAAC
2	LMX1A-LMX1B	2214	2221	TAATTAA
2	MAF::NFE2	2236	2250	TGACTCAGCAATTT
2	MAFF	2272	2289	CTGAGTCAGCAATTTTT

2	MAX::MYC	2314	2324	AGCACGTGGT
2	MEF2A	2342	2353	CTAAAAATAGA
2	MEF2B	2372	2383	CTATAAATAGC
2	MEIS1	2402	2408	TGACAG
2	MEOX2-HOXB2	2422	2431	GTAATTAAC
2	MGA	2448	2455	GGTGTGA
2	MIXL1	2470	2480	TCTAATTAAC
2	MIXL1	2484	2493	CTAATTAAC
2	MLX-MLXIPL	2497	2506	TCACGTGAT
2	MNX1	2523	2532	GTAATTAAA
2	MSC-MYF6	2549	2558	ACAGCTGTT
2	MSX1	2575	2582	CAATTAG
2	MTF1	2597	2610	TTGCACACGGCAC
2	MYB	2631	2673	cattaTAACGGTCTTTAACGGTCTTTAACGGTCTTttagcgc
2	MYBL1	2677	2688	CCGTTAACGGT
2	MZF1(Var.2)	2707	2716	GAGGGGGAA
2	Nanog	2733	2740	ACAATGG
2	NEUROD2	2755	2764	CCATATGGT
2	NFATC2	2781	2787	TTTCCA
2	NFE2	2801	2811	ATGACTCATC
2	NFIA	2829	2838	GTGCCAAGT
2	NFIC	2855	2860	TGGCA
2	NFIC::TLX1	2873	2886	GGCACCATGCCAA
2	NFIL3	2907	2917	TATGTAACGT
2	NFIX	2935	2943	GTGCCAAG
2	NFkb1	2959	2971	GGGGAATCCCCT
2	NFYA	2991	3008	GAGTGCTGATTGGTCCA
2	NHLH1	3033	3042	GCAGCTGCG
2	NKX2-3	3059	3068	CCACTTGAA
2	NKX6-1	3085	3092	TAATTAA
2	NR1H2::RXRA	3107	3123	AAGGTCAAAGGTCAAC
2	NR2C2	3147	3161	GGGGTCAGAGGTCA
2	NR3C1	3183	3199	GGTACATAATGTTCCT
2	NR3C2	3223	3239	GGAACACAATGTTCCC
2	NR4A2	3263	3270	AGGTCAC
2	NRF1	3285	3295	CGCCTGCGCA
2	NRL	3313	3323	ATTTGCTGAC
2	OLIG1	3341	3350	ACATATGTT
2	OLIG2	3367	3376	CCATATGGT
2	ONECUT1-2	3393	3406	AAAAATCGATAAT
2	OTX1	3427	3434	TAATCCG
2	PAX1-9	3449	3465	GTCACGCATGACTGCA
2	PAX5	3489	3507	AGGGCAGCCAAGCGTGAC
2	PAX7-3	3533	3542	AATCGATTA

2	PBX1	3559	3571	CCATCAATCAAA
2	PBX1	3575	3586	CATCAATCAAA
2	PHOX2A-PROP1	3590	3600	AATTTAATTA
2	PITX3	3618	3626	TTAATCCC
3	PKNOX1-2	475	486	GACAGGTGTCA
3	E2F8	505	516	TTCCCGCCAAA
3	EGR3	535	549	TACGCCCACGCACT
3	FOXP1	571	585	AAAAGTAAACAAAG
3	ELF5	607	617	CCCGGAAGTA
3	MAFG	635	655	AATTGCTGAGTCAGCATATT
3	MEF2D	683	694	CTATAAATAGA
3	HOXC13	713	723	CTCGTAAAAA
3	MEIS2	741	748	TGACAGC
3	JDP2(Var.2)	763	774	ATGACGTCATC
3	IRF2	793	810	GAAAGTGAAAGCAAAAC
3	MAX	835	844	CCACGTGCT
3	EHF	861	872	ACCCGGAAGTA
3	FOXP2	891	901	AGTAAACAAA
3	CEBPB	919	928	TTGCGCAAT
3	ELK3	945	954	CCGGAAGTA
3	FOSL1	971	981	GTGACTCATG
3	FOXD1	999	1006	TAAACAT
3	PLAG1	1021	1034	GGGCCCAAGGGGG
3	POU1F1	1055	1068	ATATGCAAATTAG
3	POU2F2	1089	1101	TCATTTGCATAT
3	POU3F3	1121	1133	TTATGCTAATTT
3	POU3F4-POU5F1B	1153	1161	ATGCAAAT
3	JUN(Var.2)	1177	1190	GGAGATGACTCAT
3	LHX9	1211	1218	CAATTAA
3	POU4F1	1233	1246	TGAATAATTAATG
3	POU6F1	1267	1276	TTAATTAAT
3	POU6F2	1293	1302	GCTCATTAT
3	EN2	1319	1328	CCAATTAGC
3	NKX2-8	1345	1353	CACTTGAA
3	ESR2	1369	1383	GGTCACCCTGACCT
3	PPARG	1405	1424	TAGGTCACGGTGACCTACT
3	GATA2	1428	1441	GATTCTTATCTGT
3	PRDM1	1485	1499	GAAAGTGAAAGTGA
3	PROX1	1521	1532	AAGACGCCTTA
3	RARA	1551	1568	AGGTCAAAAGGTCAATG
3	RARA(var.2)	1593	1609	GGTCATGCAAAGGTCA
3	RARA::RXRA	1633	1649	GGTCATGGAGAGGTCA
3	RAX	1673	1682	CCAATTAAC
3	REL	1699	1708	GGGATTTCC

3	REST	1725	1745	TCAGCACCATGGACAGCGCC
3	RFX2,3,4,5	1773	1788	GTTGCCATGGCAACG
3	RHOXF1	1811	1818	TAATCCC
3	RORA	1833	1842	TCAAGGTCA
3	RREB1	1859	1878	CCCAAACCACCCCCCCCCA
3	RUNX2	1905	1913	AACCGCAA
3	RXRA::VDR	1929	1943	GGTCAACGGGTTCA
3	SCRT1	1965	1979	AGCAACAGGTGGTT
3	SHOX	2001	2008	TAATTGG
3	SMAD2::SMAD3::SMAD4	2023	2035	TGTCTGTCACCT
3	SMAD3	2055	2064	GTCTAGACA
3	SNAI2	2081	2089	ACAGGTGT
3	SOX21	2105	2119	ACAATGGTAGTGTT
3	SOX4	2141	2156	AACAATTGCAGTGTT
3	SOX8	2179	2194	ACAATGTGCAGTGTT
3	SOX9	2217	2225	CATTGTTC
3	SOX2	2241	2255	CTTTGTTATAGAAA
3	SP1	2277	2287	CCCCGCCCCC
3	SP4	2305	2321	AAGCCACGCCCCCTTT
3	SPDEF	2345	2355	CCCGGATGTA
3	SPI1	2373	2386	AAAAGCGGAAGTA
3	SPIB	2407	2413	GAGGAA
3	SPIC	2427	2440	AAAAGAGGAAGTA
3	SREBF1	2461	2470	TCACCCCAC
3	SREBF2	2487	2496	TGGGGTGAT
3	SRF	2513	2528	GACCATATATGGTCA
3	SRY	2551	2559	TAAACAAT
3	STAT1	2575	2585	TTCCAGGAAA
3	STAT1::STAT2	2603	2617	CAGTTTCATTTTCC
3	STAT3	2639	2649	TTCTGGGAAA
3	T	2667	2682	CACACCTAGGTGTGA
3	TAL1::TCF3	2705	2716	GACCATCTGTT
3	TBP	2735	2744	TATAAAAAG
3	TBR1	2761	2770	GGTGTGAAA
3	TBX20	2787	2797	AGGTGTGAAG
3	TCF4	2815	2824	GCACCTGCT
3	TCF7L2	2841	2854	AAGATCAAAGGAA
3	TEAD1-4	2875	2884	ACATTCCAT
3	TFAP2A	2901	2911	GCCTCAGGCA
3	TFAP2A(var.2)	2929	2940	GCCCCCGGGCA
3	TFAP2B/C/A(var.3)	2959	2971	GCCCTGAGGGCA
3	TFAP2C-TFAP2B	2991	3002	GCCCCAGGGCA
3	TFAP4	3021	3030	ACAGCTGAT
3	TFCP2	3047	3056	AACCGGTTT

3	TGIF1-2	3073	3084	GACAGCTGTCA
3	THAP1	3103	3111	TGCCCGCA
3	TP53	3127	3144	ACATGCCCGGGCATGTC
3	TP73	3169	3186	ACATGTCTGGACATGTC
3	USF1	3211	3221	CCACGTGACC
3	USF2	3239	3249	TCATGTGACC
3	VAX1-2	3267	3274	TAATTAC
3	VENTX	3289	3297	CCGATTAG
3	XBP1	3313	3326	ATGCCACGTCATC
3	YY1	3347	3358	AAGATGGCGGC
3	YY2	3377	3387	TCCGCCATTA
3	ZBED1	3405	3417	TATCGCGACATA
3	ZBTB18	3437	3449	ATCCAGATGTTC
3	ZBTB33	3469	3483	TCTCGCGAGATCTG
3	ZBTB7A	3505	3516	GCGACCACCGA
3	ZNF143	3535	3550	ACCCACAATGCATTG
3	ZNF263	3573	3593	GAGGAGGAGGGGGAGGAGGA
3	ZNF354C	3621	3626	TCCAC
3	ZNF410	3639	3655	CCATCCCATAATACTC

Supplementary Table 6.2.2: List of TF binding sites identified in each catTFRE clone by CiiiDER software with default settings.

The number of binding sites is represented as * for 1 to 10 binding sites, ** for10 to 20 binding sites, *** for 20 to 50 binding sites and **** for > 50 binding sites

TFs	Clone 1	Clone_2	Clone_3	Control
Ahr::Arnt	**	**	*	*
Alx1	***	***	**	Na
ALX3	****	****	***	*
Alx4	***	***	**	Na
Ar	*	*	Na	Na
ARGFX	****	****	***	*
Arid3a	****	****	****	***
Arid3b	**	***	***	*
Arid5a	**	*	*	*
Arnt	***	**	*	*
ARNT::HIF1A	***	***	***	***
ARNT2	***	**	*	Na
Arntl	***	**	*	Na
Arx	***	***	***	Na
ASCL1	***	***	****	*
ASCL1(var.2)	*	**	**	*
Ascl2	*	*	***	*
Atf1	***	***	**	*
ATF2	**	*	*	Na
ATF3	*	*	*	Na
ATF4	*	*	*	Na
ATF6	*	*	*	Na
ATF7	*	*	*	Na
Atoh1	**	**	*	*
ATOH1(var.2)	Na	*	*	*
ATOH7	**	**	**	*
BACH1	*	**	*	Na
Bach1::Mafk	*	*	*	Na
BACH2	**	**	*	Na
BACH2(var.2)	*	*	*	Na
BARHL1	***	***	***	*
BARHL2	***	***	***	*
BARX1	****	****	****	**
BARX2	**	***	*	*
BATF	**	**	**	*
BATF::JUN	**	**	**	*
BATF3	**	**	**	*

BCL6	*	Na	*	Na
BCL6B	*	Na	Na	Na
Bhlha15	**	**	**	*
BHLHA15(var.2)	**	***	***	*
BHLHE22	**	**	*	*
BHLHE22(var.2)	Na	*	*	*
BHLHE23	**	**	*	Na
BHLHE40	***	**	*	Na
BHLHE41	**	**	*	Na
BSX	****	****	****	**
CDX1	*	*	*	*
CDX2	*	*	Na	*
CDX4	*	*	*	*
CEBPA	*	*	*	*
CEBPB	**	*	*	*
CEBPD	*	*	*	*
CEBPE	**	*	*	*
CEBPG	*	*	*	Na
CEBPG(var.2)	*	*	*	Na
CENPB	*	Na	*	*
CLOCK	***	**	*	Na
CREB1	**	*	*	Na
CREB3	*	*	*	Na
CREB3L1	*	*	*	Na
Creb3l2	***	**	*	Na
CREB3L4	**	*	*	Na
CREB3L4(var.2)	***	**	*	Na
Creb5	**	*	*	Na
CREM	**	*	*	Na
Crx	**	*	*	Na
CTCF	*	Na	Na	Na
CTCFL	*	*	*	Na
CUX1	**	**	*	*
CUX2	**	**	*	*
DBP	*	*	Na	Na
Ddit3::Cebpa	*	*	*	*
Dlx1	****	****	***	*
Dlx2	****	****	***	*
Dlx3	****	****	****	**
Dlx4	****	****	***	*
DLX5	****	****	***	*
DLX6	****	****	***	*

Dmbx1	*	*	*	Na
Dmrt1	*	*	*	*
DMRT3	*	**	*	*
DMRTA2	***	***	***	***
DMRTC2	*	*	*	Na
DPRX	***	**	**	*
DRGX	****	****	***	*
Dux	**	***	**	Na
DUX4	*	*	*	Na
DUXA	**	*	*	Na
E2F1	*	*	Na	*
E2F2	**	*	*	*
E2F3	*	Na	Na	Na
E2F4	*	*	Na	*
E2F6	*	*	*	*
E2F7	*	Na	*	Na
E2F8	*	Na	*	Na
EBF1	Na	Na	*	Na
Ebf2	*	Na	*	*
EBF3	*	Na	*	*
EGR1	*	*	*	Na
EGR2	*	*	*	Na
EGR3	*	*	*	Na
EGR4	*	*	*	Na
EHF	*	*	*	*
ELF1	*	*	*	Na
ELF2	*	*	*	Na
ELF3	*	*	*	Na
ELF4	*	*	*	Na
ELF5	***	**	***	*
ELK1	**	*	*	Na
ELK3	**	*	**	Na
ELK4	**	*	*	*
EMX1	****	****	***	*
EMX2	****	****	***	*
EN1	****	****	***	*
EN2	****	****	****	**
EOMES	*	**	**	*
ERF	**	*	*	Na
ERG	**	*	**	Na
ESR1	*	Na	*	Na
ESR2	*	*	***	Na

92

ESRRA	*	**	**	Na
ESRRB	*	**	*	Na
Esrrg	**	**	**	*
ESX1	****	****	****	**
ETS1	**	*	**	Na
ETS2	**	*	*	Na
ETV1	*	*	*	Na
ETV2	*	Na	*	Na
ETV3	**	*	*	Na
ETV4	*	*	*	Na
ETV5	**	*	*	*
ETV6	**	*	**	*
EVX1	****	****	****	**
EVX2	****	****	****	**
EWSR1-FLI1	*	Na	Na	Na
FERD3L	Na	*	*	Na
FEV	**	*	**	Na
FIGLA	***	***	***	*
FLI1	**	*	*	Na
FOS	***	***	**	*
FOS::JUN	***	***	**	*
FOS::JUN(var.2)	*	*	*	Na
FOS::JUNB	***	***	**	*
FOS::JUND	**	***	**	*
FOSB::JUN	*	*	*	Na
FOSB::JUNB	***	***	**	*
FOSB::JUNB(var.2)	**	**	*	Na
FOSL1	**	**	**	Na
FOSL1::JUN	**	***	**	*
FOSL1::JUN(var.2)	**	**	*	Na
FOSL1::JUNB	***	***	**	*
FOSL1::JUND	***	***	***	*
FOSL1::JUND(var.2)	**	**	**	*
FOSL2	**	**	**	Na
FOSL2::JUN	***	***	**	*
FOSL2::JUN(var.2)	**	*	*	Na
FOSL2::JUNB	***	***	**	*
FOSL2::JUNB(var.2)	**	**	*	Na
FOSL2::JUND	***	***	**	*
FOSL2::JUND(var.2)	**	*	*	Na
FOXA1	**	*	*	Na
FOXA2	**	*	**	*

FOXA3	**	*	**	Na
FOXB1	*	*	*	Na
FOXC1	**	*	**	*
FOXC2	**	*	**	*
FOXD1	**	*	**	*
FOXD2	*	*	*	Na
Foxd3	*	*	*	*
FOXE1	*	*	*	*
Foxf1	*	*	*	Na
FOXF2	*	*	*	Na
FOXG1	*	*	*	Na
FOXH1	*	*	*	*
FOXI1	**	*	**	Na
Foxj2	**	**	**	*
Foxj3	*	*	**	Na
FOXK1	*	*	*	Na
FOXK2	**	*	**	Na
FOXL1	***	**	***	*
Foxl2	**	*	*	Na
FOXN3	*	*	*	Na
Foxo1	**	*	**	*
FOXO3	*	*	*	Na
FOXO4	***	**	**	*
FOXO6	**	*	**	*
FOXP1	**	*	**	Na
FOXP2	*	*	*	Na
FOXP3	***	**	***	*
Foxq1	*	*	*	Na
GABPA	Na	*	*	Na
GATA1	**	*	*	*
GATA1::TAL1	*	Na	*	Na
GATA2	**	*	*	*
GATA3	***	***	***	**
GATA4	*	*	*	*
GATA5	**	*	*	*
GATA6	*	*	*	*
GBX1	****	****	***	*
GBX2	****	****	***	*
GCM1	*	*	*	Na
GCM2	*	*	*	*
GFI1	*	*	*	Na
Gfi1b	*	*	*	*

94

GLI3	*	*	Na	Na
GLIS1	*	*	Na	Na
GLIS2	*	*	*	Na
GLIS3	*	*	Na	Na
Gmeb1	****	****	***	***
GMEB2	***	***	**	*
GRHL1	*	Na	*	Na
GRHL2	*	*	*	Na
GSC	****	****	****	*
GSC2	****	****	****	*
GSX1	****	****	****	**
GSX2	****	****	****	**
Hand1::Tcf3	*	*	**	**
HAND2	*	*	*	*
HES1	***	**	*	*
HES2	***	**	*	*
HES5	**	*	*	Na
HES6	*	*	Na	Na
HES7	*	*	Na	Na
HESX1	****	****	***	*
HEY1	***	**	*	*
HEY2	***	**	*	Na
Hic1	**	***	**	**
HIC2	***	***	***	***
HIF1A	***	***	*	*
HINFP	*	*	*	*
HLF	Na	*	Na	Na
HLTF	****	****	****	****
HMBOX1	**	*	*	*
Hmx1	*	*	*	Na
Hmx2	*	Na	*	Na
Hmx3	*	*	*	Na
HNF1A	*	*	*	Na
HNF1B	*	*	*	Na
HNF4A	*	*	Na	*
HNF4A(var.2)	*	*	Na	Na
HNF4G	*	*	Na	*
HOXA1	****	****	****	***
HOXA10	***	***	***	*
Hoxa11	*	**	*	*
HOXA13	*	*	Na	*
HOXA2	****	****	***	*

95

HOXA4	****	****	***	*
HOXA5	****	****	****	**
HOXA6	****	****	****	**
HOXA7	****	****	****	***
HOXA9	****	****	***	**
HOXB13	*	*	Na	*
HOXB2	****	****	****	**
HOXB3	****	****	****	**
HOXB4	****	****	***	*
HOXB5	****	****	***	*
HOXB6	****	****	****	*
HOXB7	**	***	**	*
HOXB8	****	****	****	**
HOXB9	*	*	*	Na
HOXC10	**	**	*	*
HOXC11	*	*	*	Na
HOXC12	*	*	*	Na
HOXC13	*	*	*	Na
HOXC4	****	****	***	*
HOXC8	****	****	****	**
HOXC9	*	*	*	Na
HOXD10	*	*	*	Na
HOXD11	*	*	*	Na
HOXD12	*	*	*	Na
HOXD13	*	*	*	*
HOXD3	***	****	***	*
HOXD4	****	****	***	*
HOXD8	****	****	****	**
HOXD9	*	**	*	*
HSF1	Na	*	Na	Na
HSF2	Na	*	Na	Na
HSF4	Na	*	Na	Na
IKZF1	Na	*	*	Na
INSM1	*	*	*	Na
IRF1	*	*	*	*
IRF2	*	*	*	Na
IRF3	*	*	*	Na
IRF4	*	*	Na	Na
IRF5	*	*	Na	Na
IRF6	*	*	Na	*
IRF7	*	*	*	Na
IRF8	*	*	Na	Na

IRF9	*	*	Na	Na
Isl1	**	***	***	*
ISL2	***	**	***	*
ISX	****	****	***	*
JDP2	**	**	*	Na
JDP2(var.2)	*	*	*	Na
JUN	**	*	*	*
JUN(var.2)	***	***	**	*
JUN::JUNB	***	***	***	*
JUN::JUNB(var.2)	*	*	*	Na
JUNB	**	**	**	Na
JUNB(var.2)	*	*	*	Na
JUND	**	**	**	Na
JUND(var.2)	**	**	*	Na
Klf1	*	*	*	*
KLF10	*	*	*	*
KLF11	*	**	*	*
Klf12	*	*	*	*
KLF13	*	*	*	Na
KLF14	*	*	*	Na
KLF15	*	*	*	Na
KLF16	**	*	*	*
KLF17	*	Na	*	Na
KLF2	***	**	***	*
KLF3	*	*	*	Na
KLF4	*	*	*	Na
KLF5	**	**	**	*
KLF6	**	**	*	*
KLF9	*	*	*	*
LBX1	****	****	***	*
LBX2	****	****	***	**
LEF1	*	*	*	Na
LHX1	****	****	****	****
LHX2	****	****	***	*
Lhx3	*	**	*	Na
Lhx4	****	****	***	*
LHX5	****	****	***	*
LHX6	****	****	***	*
Lhx8	****	****	****	*
LHX9	****	****	***	*
LIN54	*	**	*	*
LMX1A	***	****	***	Na

LMX1B	***	***	**	Na
MAF	Na	*	*	Na
MAF::NFE2	*	*	*	Na
MAFA	*	*	*	Na
Mafb	*	**	**	*
MAFF	*	*	*	*
MAFG	*	*	*	*
MAFK	*	*	*	Na
MAX	***	**	*	Na
MAX::MYC	***	**	*	Na
MAZ	Na	*	*	*
MEF2A	*	*	*	*
MEF2B	*	*	*	Na
MEF2C	*	*	*	*
MEF2D	*	*	*	Na
MEIS1	***	***	****	***
MEIS1(var.2)	*	*	*	Na
MEIS2	**	*	**	*
MEIS2(var.2)	*	*	*	Na
MEIS3	**	*	***	**
MEOX1	****	****	***	*
MEOX2	****	****	****	***
MGA	*	**	**	*
MITF	*	*	*	*
mix-a	***	****	***	Na
MIXL1	****	****	****	**
MLX	***	**	*	Na
Mlxip	***	**	*	Na
MLXIPL	***	**	*	Na
MNT	***	**	*	Na
MNX1	****	****	***	**
MSANTD3	**	**	*	*
MSC	*	*	**	Na
MSGN1	****	****	****	**
MSX1	****	****	***	*
MSX2	****	****	***	*
Msx3	****	****	***	*
MTF1	Na	*	Na	Na
MXI1	Na	Na	*	*
MYB	*	*	*	*
MYBL1	*	*	*	*
MYBL2	*	Na	*	Na

MYC	***	**	*	Na
MYCN	***	**	*	Na
MYF5	Na	*	*	*
MYF6	**	**	**	Na
MYOD1	*	**	***	*
MYOG	Na	*	*	*
MZF1	*	*	*	*
MZF1(var.2)	**	***	**	*
NEUROD1	*	*	*	*
NEUROD2	**	**	*	*
NEUROG1	**	**	*	Na
NEUROG2	**	**	Na	Na
NEUROG2(var.2)	*	*	*	*
NFAT5	**	**	**	*
NFATC1	***	***	***	**
NFATC2	**	**	**	*
NFATC3	***	***	***	**
NFATC4	***	**	***	**
NFE2	**	**	*	Na
NFE2L1	*	*	*	Na
Nfe2l2	*	*	*	Na
NFIA	*	*	*	*
NFIB	Na	*	Na	Na
NFIC	*	**	*	*
NFIC(var.2)	Na	*	Na	Na
NFIC::TLX1	Na	*	Na	Na
NFIL3	Na	*	Na	Na
NFIX	***	***	***	***
NFIX(var.2)	Na	*	Na	*
NFKB1	*	*	Na	Na
NFKB2	*	*	Na	Na
NFYA	**	**	*	Na
NFYB	*	*	*	Na
NFYC	**	**	*	Na
NHLH1	Na	*	**	*
NHLH2	Na	*	*	Na
NKX2-2	Na	*	*	*
NKX2-3	**	*	*	*
NKX2-5	*	*	*	*
Nkx2-5(var.2)	*	*	*	*
NKX2-8	**	**	**	**
Nkx3-1	**	*	*	*

Nkx3-2	*	*	Na	Na
NKX6-1	****	****	***	*
NKX6-2	****	****	****	***
NKX6-3	***	****	***	*
Nobox	****	****	***	*
NOTO	****	****	****	*
Npas2	***	**	*	*
NR1D1	Na	Na	*	Na
NR1D2	*	*	*	Na
NR1H2::RXRA	Na	*	Na	Na
Nr1h3::Rxra	*	*	*	Na
NR1H4	*	*	**	*
NR1H4::RXRA	Na	Na	*	Na
NR1I2	Na	Na	*	Na
NR1I3	Na	*	*	*
NR2C1	***	***	***	*
NR2C2	Na	*	Na	Na
NR2C2(var.2)	***	***	****	**
Nr2e1	*	*	*	*
Nr2e3	Na	Na	Na	*
NR2F1	**	***	***	*
NR2F1(var.2)	Na	*	Na	Na
NR2F1(var.3)	*	Na	*	Na
NR2F2	**	**	***	Na
Nr2f6	*	*	Na	Na
Nr2f6(var.2)	Na	Na	*	Na
NR2F6(var.3)	*	Na	Na	Na
NR3C1	*	*	Na	Na
NR3C2	*	*	Na	Na
NR4A1	*	*	*	*
NR4A2	**	***	***	**
NR4A2::RXRA	*	*	*	Na
NR5A1	**	**	**	*
Nr5a2	*	*	*	*
NR6A1	*	*	*	Na
NRF1	Na	*	Na	Na
NRL	*	*	*	*
OLIG1	**	**	Na	Na
OLIG2	**	**	*	*
OLIG3	**	**	*	*
ONECUT1	*	*	*	Na
ONECUT2	*	*	Na	*

100

ONECUT3	*	*	Na	Na
OSR1	Na	*	Na	*
OSR2	Na	Na	*	*
OTX1	**	***	**	*
OTX2	*	*	*	Na
OVOL1	*	*	Na	Na
OVOL2	*	**	Na	*
PAX1	*	*	Na	Na
Pax2	****	****	****	***
PAX3	*	*	Na	Na
PAX4	****	****	***	*
PAX5	*	*	*	*
PAX6	*	*	*	Na
PAX7	*	*	Na	Na
PAX9	Na	*	Na	Na
PBX1	*	*	Na	Na
PBX2	*	*	*	Na
PBX3	Na	*	*	Na
PDX1	****	****	****	**
PHOX2A	*	**	*	*
PHOX2B	*	*	*	Na
PITX1	**	***	**	*
PITX2	***	***	**	*
PITX3	***	***	**	*
PKNOX1	Na	*	*	Na
PKNOX2	Na	*	*	Na
PLAG1	Na	Na	*	Na
Plagl1	*	Na	*	*
PLAGL2	*	Na	*	*
POU1F1	*	*	*	*
POU2F1	*	*	*	*
POU2F2	*	*	*	*
POU2F3	*	*	*	Na
POU3F1	*	*	**	*
POU3F2	*	*	*	Na
POU3F3	*	*	*	Na
POU3F4	*	*	*	Na
POU4F1	*	*	*	Na
POU4F2	Na	*	*	Na
POU4F3	Na	*	*	Na
POU5F1	*	*	*	*
Pou5f1::Sox2	*	*	Na	Na

POU5F1B	*	*	**	*
POU6F1	***	***	**	Na
POU6F1(var.2)	***	***	**	*
POU6F2	***	****	**	*
PPARA::RXRA	*	*	Na	Na
PPARD	*	*	Na	Na
PPARG	Na	Na	*	Na
Pparg::Rxra	*	*	*	Na
PRDM1	Na	*	*	*
Prdm15	Na	*	Na	Na
PRDM4	*	*	*	*
PROP1	*	*	Na	Na
PROX1	Na	Na	*	Na
PRRX1	****	****	****	**
PRRX2	****	****	***	*
Ptf1a	*	**	***	*
Ptf1a(var.2)	Na	*	*	*
Ptf1a(var.3)	*	**	***	*
RARA	Na	Na	*	Na
RARA(var.2)	Na	*	*	Na
RARA::RXRA	Na	*	*	*
RARA::RXRG	Na	*	*	*
Rarb	Na	Na	*	Na
Rarb(var.2)	Na	*	*	Na
Rarg	Na	Na	*	Na
Rarg(var.2)	Na	*	Na	Na
RAX	****	****	***	*
RAX2	****	****	****	**
RBPJ	*	*	*	*
Rbpjl	*	**	***	*
REL	*	*	*	Na
RELA	*	*	*	Na
RELB	*	*	*	*
REST	Na	Na	*	Na
RFX1	*	Na	*	Na
RFX2	Na	Na	*	Na
RFX3	Na	Na	*	Na
RFX4	Na	Na	*	Na
RFX5	Na	Na	*	Na
RFX7	*	**	*	*
Rhox11	*	**	**	*
RHOXF1	****	***	***	***

RORA	**	***	***	*
RORA(var.2)	*	*	*	Na
RORB	**	**	***	*
RORC	*	*	**	*
RREB1	*	Na	*	Na
RUNX1	*	*	*	*
RUNX2	*	*	**	*
RUNX3	*	*	**	*
Rxra	*	*	Na	Na
RXRA::VDR	Na	Na	*	Na
RXRB	*	*	Na	Na
RXRB(var.2)	*	*	Na	Na
RXRG	*	*	Na	Na
RXRG(var.2)	Na	*	Na	Na
SCRT1	*	*	*	Na
SCRT2	*	*	*	Na
SHOX	****	****	****	**
Shox2	****	****	***	*
SIX1	Na	*	Na	Na
Six3	*	*	*	*
Smad2::Smad3	**	**	**	*
SMAD2::SMAD3::SMAD4	*	Na	*	Na
SMAD3	*	Na	*	Na
Smad4	*	*	*	*
SMAD5	*	Na	*	Na
SNAI1	***	***	***	*
SNAI2	*	*	**	*
SNAI3	*	*	**	*
SOHLH2	***	**	*	*
Sox1	*	Na	*	Na
SOX10	*	*	*	*
Sox11	Na	Na	*	Na
SOX12	Na	*	*	Na
SOX13	*	*	**	*
SOX14	*	*	*	Na
SOX15	*	**	***	***
Sox17	*	*	**	*
SOX18	***	***	****	***
SOX2	*	*	**	Na
SOX21	Na	Na	*	Na
Sox3	*	*	*	*
SOX4	*	**	***	**

103

Sox5	*	*	**	*
Sox6	*	*	**	*
SOX8	*	*	***	*
SOX9	*	*	**	*
SP1	*	*	*	Na
SP2	*	*	*	Na
SP3	*	*	*	Na
SP4	*	*	*	Na
SP8	*	*	**	*
SP9	**	**	**	*
SPDEF	*	*	*	Na
SPI1	*	*	*	Na
SPIB	*	*	*	Na
SPIC	*	*	*	*
Spz1	*	*	*	Na
SREBF1	**	*	*	*
SREBF1(var.2)	*	*	*	*
SREBF2	*	*	*	*
SREBF2(var.2)	*	*	*	Na
SRF	Na	Na	*	Na
SRY	*	**	***	***
STAT1	*	Na	*	Na
STAT1::STAT2	*	*	*	*
Stat2	*	*	*	*
STAT3	*	*	*	Na
Stat4	*	*	*	Na
Stat5a	*	Na	*	*
Stat5a::Stat5b	*	Na	*	Na
Stat5b	*	Na	*	Na
Stat6	Na	Na	*	Na
TAL1::TCF3	**	**	**	Na
TBP	**	*	*	*
TBR1	*	**	**	*
TBX1	*	*	*	*
TBX15	*	*	**	Na
TBX18	*	*	**	*
TBX19	Na	*	*	Na
TBX2	*	**	**	*
TBX20	*	*	*	*
TBX21	*	*	**	*
TBX3	***	****	****	***
TBX4	*	**	**	*

TBX5	*	***	**	*
TBX6	***	***	***	***
TBXT	Na	*	*	Na
Tcf12	*	**	***	*
TCF12(var.2)	*	*	*	*
Tcf21	Na	*	*	Na
TCF21(var.2)	**	**	*	Na
TCF3	*	*	**	*
TCF4	*	*	**	*
TCF7	*	*	*	*
TCF7L1	*	*	*	Na
TCF7L2	*	*	*	*
TCFL5	*	*	*	*
TEAD1	Na	*	*	Na
TEAD2	*	*	*	*
TEAD3	**	**	**	*
TEAD4	*	*	*	*
TEF	*	*	*	Na
TFAP2A	Na	*	*	*
TFAP2A(var.2)	*	*	**	*
TFAP2A(var.3)	*	*	*	*
TFAP2B	*	*	**	**
TFAP2B(var.2)	Na	*	*	*
TFAP2B(var.3)	*	Na	*	*
TFAP2C	*	Na	**	*
TFAP2C(var.2)	Na	*	*	*
TFAP2C(var.3)	Na	Na	*	*
TFAP2E	*	**	***	**
TFAP4	Na	*	*	*
TFAP4(var.2)	****	****	****	**
TFCP2	*	Na	*	*
TFDP1	*	*	*	*
TFE3	***	**	**	*
TFEB	***	***	**	*
TFEC	**	*	*	Na
TGIF1	Na	*	*	Na
TGIF2	Na	*	*	Na
TGIF2LX	Na	*	**	Na
TGIF2LY	*	*	**	Na
THAP1	****	****	****	****
THRB	*	*	Na	Na
THRB(var.2)	*	Na	*	Na

THRB(var.3)	*	*	Na	Na
TLX2	****	****	***	*
TP53	*	Na	*	Na
TP63	*	Na	*	Na
TP73	*	Na	*	Na
TWIST1	*	*	*	*
Twist2	**	**	**	*
UNCX	****	****	****	**
USF1	**	*	*	*
USF2	**	*	*	*
VAX1	****	****	***	*
VAX2	****	****	***	*
VDR	**	*	**	*
VENTX	***	***	***	*
VEZF1	***	**	***	*
VSX1	****	****	***	*
VSX2	****	****	***	*
Wt1	Na	Na	*	Na
XBP1	*	*	*	Na
YY1	*	*	*	*
YY2	*	*	*	*
ZBED1	Na	Na	*	Na
ZBTB12	*	*	*	*
ZBTB14	Na	Na	Na	*
ZBTB18	*	*	*	*
ZBTB26	*	*	*	Na
ZBTB32	Na	*	Na	Na
ZBTB33	Na	*	*	*
ZBTB6	Na	*	Na	Na
ZBTB7A	**	*	*	*
ZBTB7B	*	*	*	*
ZBTB7C	**	*	*	*
ZEB1	*	*	**	*
ZFP42	*	*	*	Na
ZFP57	*	Na	*	*
Zfx	*	*	Na	*
ZIC1	*	*	Na	Na
Zic1::Zic2	Na	*	Na	*
Zic2	Na	*	Na	*
ZIC3	*	*	Na	Na
ZIC4	*	*	Na	Na
ZIC5	*	*	Na	Na

ZKSCAN1	*	*	*	*
ZKSCAN5	*	Na	Na	Na
ZNF135	Na	*	Na	Na
ZNF143	*	Na	*	Na
ZNF148	Na	*	*	Na
ZNF263	Na	*	*	*
ZNF274	Na	Na	*	*
Znf281	*	Na	*	Na
ZNF317	*	Na	*	Na
ZNF341	*	*	*	*
ZNF354C	***	***	***	***
ZNF384	*	*	*	*
ZNF410	Na	Na	*	Na
Znf423	*	Na	*	*
ZNF449	Na	Na	Na	*
ZNF652	Na	*	Na	Na
ZNF682	Na	*	Na	Na
ZNF740	*	Na	*	Na
ZNF75D	*	*	*	Na
ZSCAN29	*	Na	*	*
ZSCAN4	Na	Na	Na	*

6.3 Appendix 3: Supplementary figures
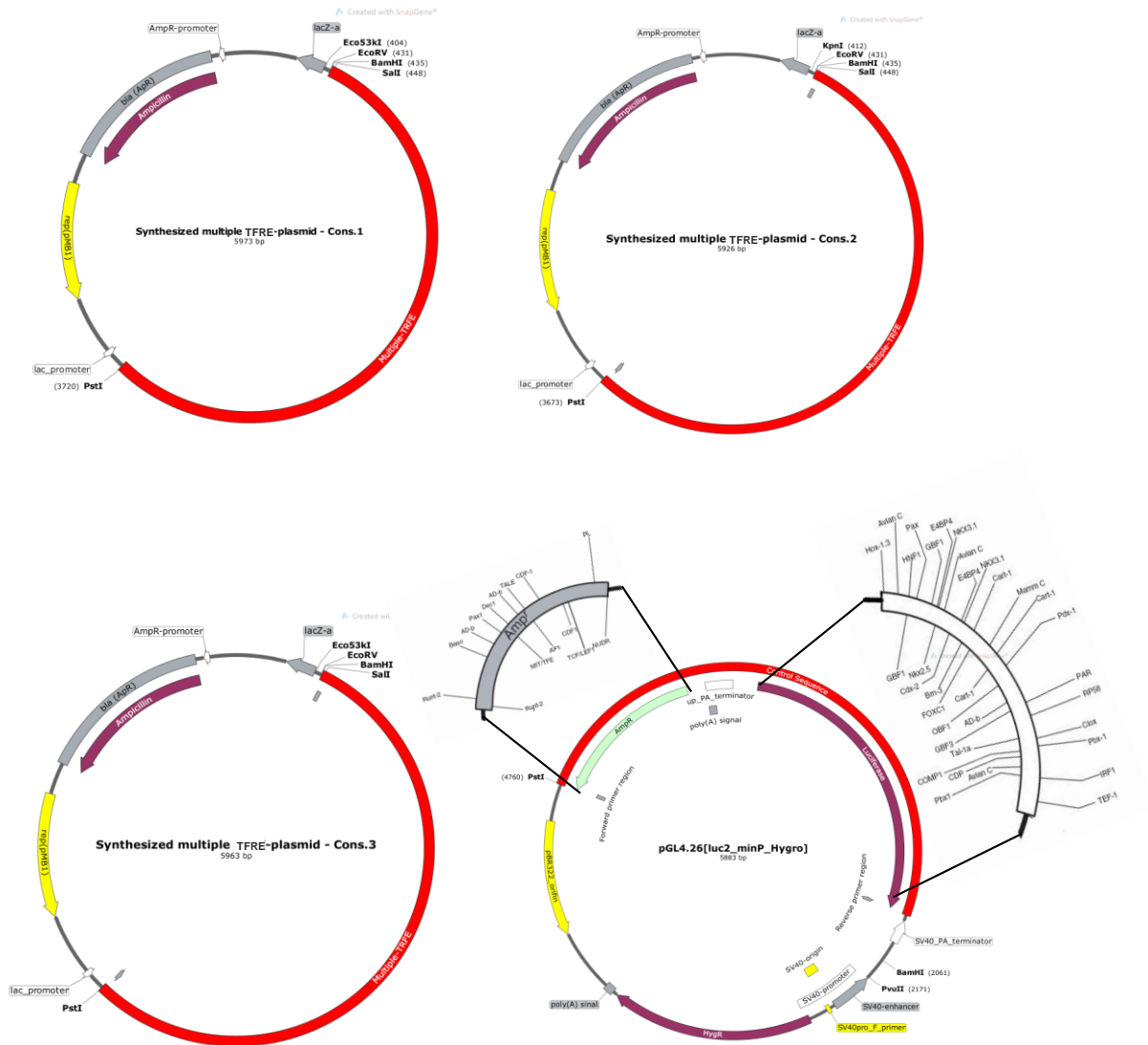
Supplementary Figure 6.3.1: **Construct map of the three synthesized multiple TRFE-plasmids and control pGL4.26PvuII- plasmid (designed by T. Tran and R. Eskeland).** The number of transcription factor binding sites in the control pGL4 plasmid backbone has been considerably reduced by mutation of known binding sites as illustrated over the plasmid map.
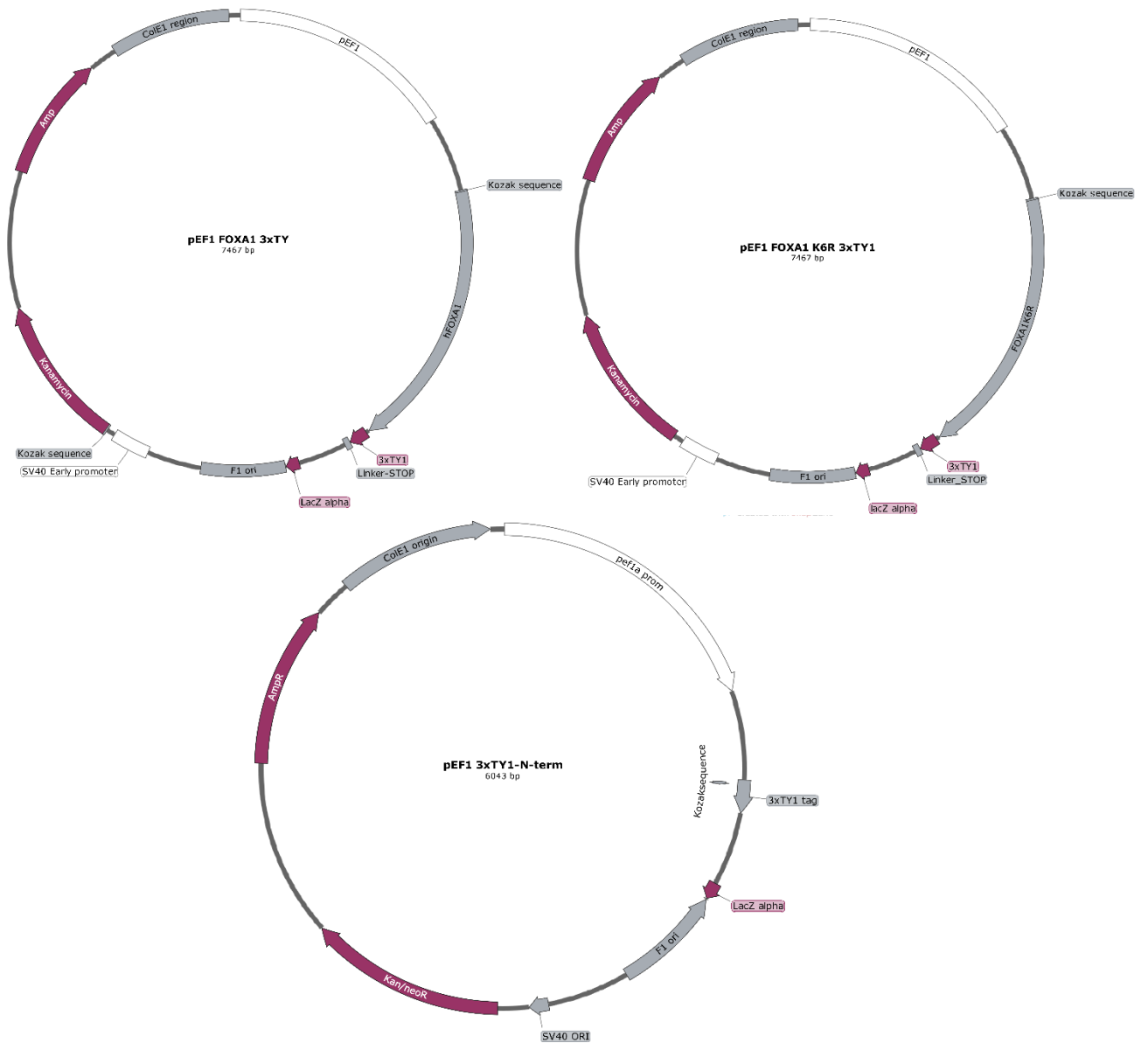
Supplementary Figure 6.3.2: **Construct map of the three plasmids used to overexpress FOXA1 3xTy, FOXA1 K6R 3xTy (designed by I. Cuervo), and 3xTy (designed by M. Fosslie) in the DU145 cell line.**

A

EcoRV (431) SalI (448)

Plasmid backbone

Synthesized multiple TFRE-plasmid (~5963 bp)

Multiple-TFRE

PstI (~3701)

PvuII (2171) BamHI (2061)

Plasmid backbone

Control – plasmid pGL4.26 (5882 bp)

Control Sequence

PstI (4759)

B

	1		2		3		Control	
Mw	Uncut	EcoRV	Uncut	EcoRV	Uncut	EcoRV	Uncut	PvuII	Mw

kbp: 12, 5, 3, 2, 1,6

Lanes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14

C

EcoRV + SalI   Control PvuII + BamHI

Mw	1	2	3		

kbp: 12, 5, 3

Lanes: 1 2 3 4 5 6 7 8 9

D

	1			2			Control		

EcoRV+SalI, EcoRV+SalI +PstI, Uncut, EcoRV+SalI, EcoRV+SalI +PstI, Uncut, PvuII+BamHI, PvuII+BamHI + PstI, Uncut

kbp: 12, 5, 3, 2, 1,6

← Biotinylated-TFRE
← Plasmid backbone

Lanes: 1 2 3 4 5 6 7 8 9 10 11

E

	1				2				Control			

Input, Supernatant, Wash, Boiled beads, Input, Supernatant, Wash, Boiled beads, Input, Supernatant, Wash, Boiled beads

kbp: 12, 5, 3, 2, 1,6

← Biotinylated-TFRE
← Plasmid backbone

Lanes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14

Supplementary Figure 6.3.3: Binding of biotinylated TFRE DNA to Dynabeads.

(A) Schematic illustration of the synthesized multiple TFRE-plasmid and the control plasmid (pGL4.26). The restriction enzyme cutting site is represented in black arrows. (B) The first RE digestion for each construct in comparison to the uncut construct. First, each construct was linearized by RE digestion with a single cutter RE to produce blunt ends (EcoRV for TFRE constructs and PvuII for the control plasmid. (C) The second RE digestion for each construct. Next, each construct was RE digested with a single cutter (SalI for the TFRE-plasmids and BamHI for the control plasmid) to produce sticky ends that enable downstream biotinylation. (D) The third RE digestion for construct 1, 2 and control, in comparison to the second RE digestion and the uncut construct. After biotinylation, DNA was digested with PstI to separate biotinylated-TRFE from the plasmid backbone. (E) Biotinylated DNA construct immobilized onto Dynabeads™. Biotinylated DNA (1 ug) from construct 1, 2 and 3 was immobilized onto Dynabeads using the kilobaseBINDER™ Kit to ensure specific binding of the biotinylated-TFRE construct to the beads. 2 ul of the supernatant and the washing after DNA immobilization was loaded to the gel. After washing DNA was boiled in SDS at 95 C for 5 min and 10 ul was loaded to the gel. 1 ul from the input DNA from construct 1, 2 and control were also loaded to the gel.
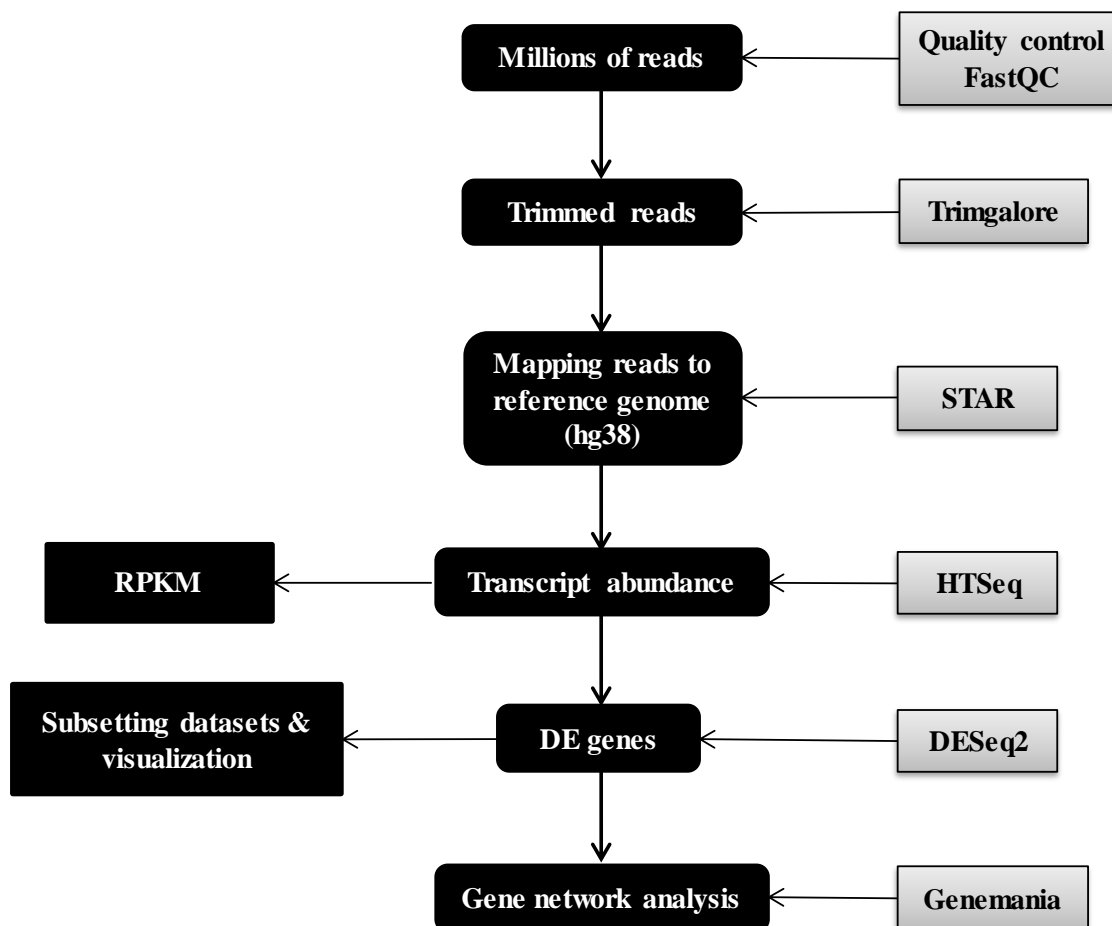
Supplementary Figure 6.3.2: Expression analysis workflow implemented in the current study.
The main steps in the analysis are indicted with black boxes, while software that was used to implement each step is indicted with grey boxes.
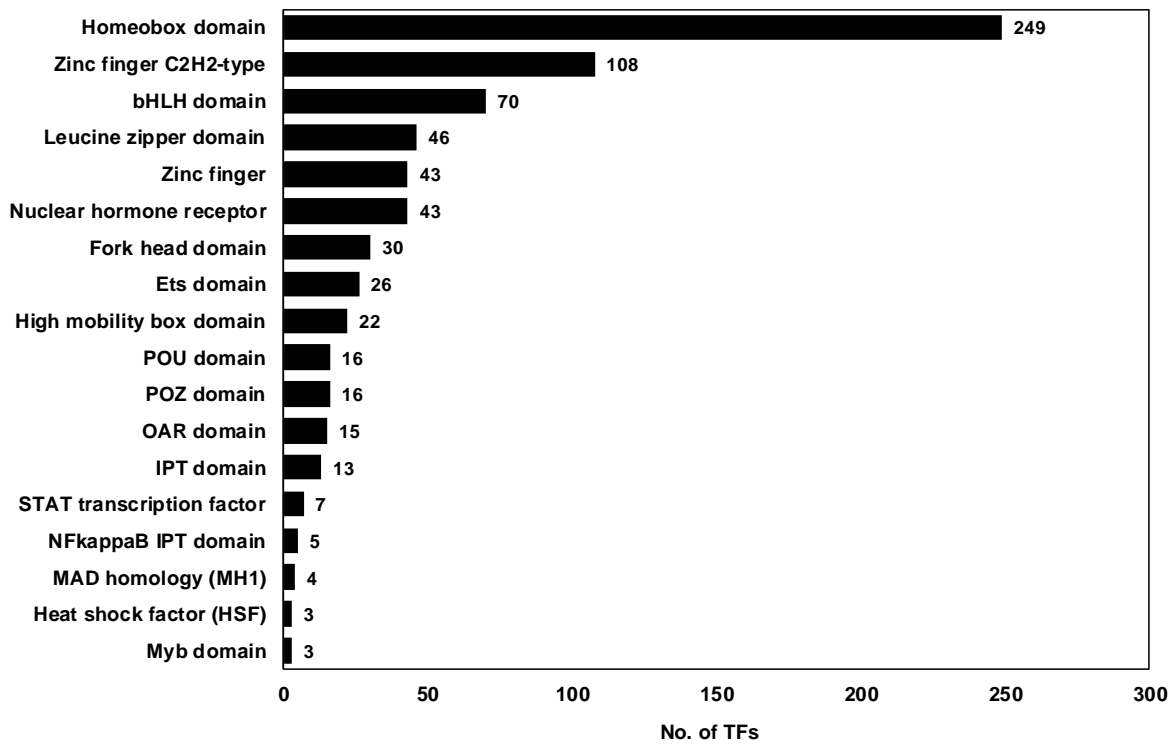
Supplementary Figure 6.3.4: The protein domains found most frequently in the transcription factors included in the catTFRE system. Functional analysis of TFs by classifying them into families based on predicted proteins domains using the InterProScan 5.41 software, with default parameters.

6.4 Appendix 4: Recipes

General:

5 M NaCl:

• 146.1 g NaCl (58.44 g/mol) in 500 ml dH2O

• The solution was filtered

1 M Tris-HCl (pH 8.0):

• 121.1 g Tris (121.14 g/mol) in 800 ml milliQ

• The pH was adjusted to 8.0 with 37% HCl

• The final volume was made to 1 L with dH2O

• The solution was filtered

0.5 M Tris-HCl (pH 6.8):

• 30.28 Tris (121.1 g/mol)

• 400 ml of dH2O dissolved

• The pH was adjusted to 6.8 with 37% HCl

• The final volume was made to 500 ml with dH2O

• The solution was autoclaved

0.5 M EDTA (pH 8.0):

• 93.1 g Na2EDTA.H2O (372.24 g/mol)

• 400 ml milliQ

• The pH was adjusted to 8.0 with NaOH tablets

• The final volume was made to 1 L with dH2O

• The solution was filtered

TE buffer:

• 10 ml of 1 M Tris-HCl pH 8.0

• 2 ml of 0.5 M EDTA

• The volume was filled up to 1 L with dH2O

5x Orange G:

• 25 ml 100% glycerol

• 250 µl of 0.5 M EDTA

• 0.075 g of Orange G (452.38 g/mol)

• The solution was sterile filtered before use

Solutions for bacterial cell culture:

50% Glycerol:

• 50 ml 100% glycerol

• 50 ml dH2O

• The solution was autoclaved

LB medium:

• 10 g tryptone

• 5 g yeast extract

• 10 g NaCl (58.44 g/mol)

• 1000 ml dH2O

• The pH was adjusted to 7.2 with NaOH tablets

• The solution was autoclaved

LB agar plates:

• 6 g agar was added to 400 ml LB medium

• The solution was autoclaved

• The solution was placed in a water bath set to 55°C

• When the solution was approximately 55°C, antibiotics were added if required

• 20 ml of the LB agar was added for one petri dish (enough for 20 plates)

• The plates were dried in a laminar flow hood

SOB medium:

• 20 g tryptone

• 5 g yeast extract

• 0.5 g NaCl (58.44 g/mol)

• 800 ml dH2O

• 1 ml of 2.5M KCl

• The pH was adjusted to 7.0

• The volume was filled up to 1 L with dH2O

• The solution was autoclaved

• 10 ml of sterile filtered 2 M MgSO4 was added before use

Transformation buffer:

• 3 g Pipes (302.4 g/mol)

• 2.2 g CaCl2 (147.02 g/mol)

• 18.6 g KCl (74.56 g/mol)

• 800 ml dH2O

• The pH was adjusted to 6.7 with 10 M KOH

• 10.9 g MnCl2 (197.9 g/mol)

• The final volume was adjusted to 1 L with dH2O

• The solution was sterile filtered before use

Solutions for mammalian cell culture:

DU145 culturing medium:

• 450 mL DMEM (Gibco)

• 500 units P/S (Gibco)

• 50 mL FBS (Capricorn Scientific)

• 400 µg/ml G418 (Geneticin)

0.5 mM EDTA:

• 0.5 mL 0.5 M EDTA

• 500 mL 1x dPBS

• The solution was sterile filtered before use

Lysis buffer for gDNA extraction:

• 5 ml of 1 M Tris-HCl (pH 8.0)

• 2 ml of 5 M NaCl

• 0.5 ml of 0.5 M EDTA (pH 8.0)

• 0.5 ml of 20% SDS

• The solution was filled up to 50 ml with dH2O

• Proteinase K was freshly added before use to a final concentration of 0.1 mg/ml

RIPA Lysis Buffer

• 1.5 ml of 1 M NaCl
• 0.1 ml of 1% Nonidet P-40

• 0.05 ml of 0.5% Sodium deoxycholate

• 0.01 ml of 0.1% SDS

• 5 ml of 50 mM Tris (pH 7.4)

• Add ddH2O to 10 ml

Solutions for Nuclear protein extraction:

Hypotonic Buffer N

• 10 mM Hepes pH 7.5.

• 2 mM MgCl2

• 25 mM KCl

• 1 mM DTT

• 1 mM PMSF

Buffer N

• 10 mM Hepes pH 7.5.

• 2 mM MgCl2

• 25 mM KCl

• 250 mM Sucrose

• 1 mM DTT

• 1 mM PMSF

• 1x Protease inhibitor cocktail

IP Buffer C (Nuclear extract)

• 25 mM Hepes pH 7.6

• 150 mM NaCl

• 12.5 mM MgCl2

• 0.1 mM EDTA

• 10% Glycerol

• 0.2 mM PMSF

• 1mM DTT

• 1x Protease inhibitor cocktail

Solutions for TF pulldown with catTFRE

BC100 /300 / 500

• 25 mM Hepes pH 7.6

• 100 / 300 / 500 mM NaCl

• 1 mM MgCl2

• 0.5 mM EGTA

• 0.1 mM EDTA

• 10% Glycerol

• 0.05% NP40

• 0.2 mM PMSF

• 1mM DTT

• 1x Protease inhibitor cocktail

Table 6.4.1 different washing buffered tested for TF pulldown (Section 2.4.3)

Buf. BC 150 + TTN	Buf. BC 150 ++ TTN	Buf. BC 250 + TTN	Buf. BC 250 ++TTN	Buf. BC 150 -TTN
25 mM Hepes pH 7.6 150 mM NaCl 1 mM MgCl2 0.5 mM EGTA 0.5 mM EDTA 10% Glycerol 0.5% NP40 0.5 % Tween 20 0.1 % Triton 100 X 0.2 mM PMSF 1mM DTT 1x Pro. inh. cocktail	25 mM Hepes pH 7.6 150 mM NaCl 1 mM MgCl2 0.5 mM EGTA 0.5 mM EDTA 10% Glycerol 0.5% NP40 1 % Tween 20 0.5 % Triton 100 X 0.2 mM PMSF 1mM DTT 1x Pro. inh. cocktail	25 mM Hepes pH 7.6 250 mM NaCl 1 mM MgCl2 0.5 mM EGTA 0.5 mM EDTA 10% Glycerol 0.5% NP40 0.5 % Tween 20 0.1 % Triton 100 X 0.2 mM PMSF 1mM DTT 1x Pro. inh. cocktail	25 mM Hepes pH 7.6 250 mM NaCl 1 mM MgCl2 0.5 mM EGTA 0.5 mM EDTA 10% Glycerol 0.5% NP40 1 % Tween 20 0.5 % Triton 100 X 0.2 mM PMSF 1mM DTT 1x Pro. inh. cocktail	25 mM Hepes pH 7.6 150 mM NaCl 1 mM MgCl2 0.5 mM EGTA 0.5 mM EDTA 10% Glycerol 0.5% NP40 0.2 mM PMSF 1mM DTT 1x Pro. inh. cocktail

Solutions for SDS-PAGE:

1 M DTT

• 7.7 g DTT (154.3 g/mol)

• 50 ml dH2O

20% SDS

• 100 g SDS (228.37 g/mol)

• 500 ml dH2O

4x SDS loading dye:

• 2.5 ml of 1 M Tris-HCl pH 8.0

• 4 g SDS (228.37 g/mol)

• 20 ml 100% glycerol

• 0.3 mg bromophenol blue (669.96 g/mol)

• The final volume was made to 50 ml with dH2O

• DTT was freshly added to an end concentration of 10% prior to use.

10x SDS running buffer:

• 150.1 g glycine (75.07 g/mol)

• 30.28 g Tris-(hydroxymethyl) aminomethane (121.14 g/mol)

• 10 g SDS (228.37 g/mol)

• 1 L dH2O

Solutions for Western blotting:

10x TBS-T washing buffer:

• 50 ml of 1 M Tris-HCl pH 8.0

• 150 ml of 5 M NaCl

• 2.5 ml Tween 20

• The volume was filled up to 500 ml with dH2O

Coomassie blue solution:

• 0.25 g Coomassie Brilliant Blue

• 4.5 ml dH2O

• 4.5 ml methanol

• 10 ml acetic acid

6.5 Appendix 5: Materials

Table 6.5.1: An overview of the materials that were used in this thesis.

Material	Company	Catalogue number
Antibiotics		
Ampicillin	Sigma	A0166
Kanamycin	Sigma	K4000
Antibodies		
Goat-anti-GAPDH	Sigma	PLA0302
Mouse-anti-Ty	In house hybridoma from Keith Gull Lab	-
Rabbit-anti-CTCF	Sigma	07-729
Donkey-anti-mouse-IR680	LI-COR	926-68072
Donkey-anti-mouse-IR800	LI-COR	926-32212
Donkey-anti-rabbit-IR680	LI-COR	926-68073
Donkey-anti-rabbit-IR800	LI-COR	926-32213
Donkey-anti- goat -IR800	LI-COR	925-68024
Buffers		
10x NEB buffer 3	New England BioLabs®	B7203
10x NEB CutSmart buffer	New England BioLabs®	B7204
10x NEB buffer 1	New England BioLabs®	B7201
10x NEB buffer 2	New England BioLabs®	B7202
Cell culture		
DMEM	Gibco®Invitrogen	41965-039
1x DPBS	Gibco®Invitrogen	14200-067
FBS advanced	Capricorn Scientific	FBS-11A
P/S	Gibco®Invitrogen	15140-122
Trypsin-EDTA	Sigma	T4174
DMSO	Sigma	D2650
Tryphan blue	Life Technologies	T10282
Countess™ automated cell counter	Invitrogen	C10227
EVE™ cell counting slides	Nano EnTek	EVS-050
Imaging		
Odyssey® CLx	LI-COR	-
LSM880 airyscan confocal	ZEISS	-
Ladders		
1 Kb Plus DNA Ladder	Invitrogen	10787018
Precision Plus Protein™ Dual Color Standard	Bio-Rad	161-0374
Kits		
NucleoSpin® Gel and PCR Clean-up	Machery-Nagel	740609.250
NucleoSpin® Plasmid	Machery-Nagel	740588.250
NucleoSnap® Plasmid Midi	Machery-Nagel	740494.50
NucleoBond® Xtra Maxi	Machery-Nagel	740414.50
QIAGEN® Plasmid Giga prep	QIAGEN	12991

kilobaseBINDER™	Thermo Fisher	60101
Plasmids		
Synthesized multiple TRFE-plasmids 1 to 3	Designed by T. Tran and R. Eskeland	-
Control pGL4.26PvuII-plasmid	Designed by T. Tran and R. Eskeland	-
pEF1 FOXA1 3xTy	Designed by I. Cuervo	-
pEF1 FOXA1 K6R 3xTy	Designed by I. Cuervo	-
pEF1 3xTy	Designed by M. Fosslie	-
Diverse		
Biotinylated dUTP	Sigma	136632-31-0
Biotinylated dATP	Sigma	19524016
dCTP	Thermo Fisher	R0151
dGTP	Thermo Fisher	R0161
Dynabeads™ M-280 Streptavidin	Thermo Fisher	11205D
DTT	Sigma	D0632
Enzymes		
RNase A	Sigma	R6513
Klenow Fragment (3´→ 5´ exo-)	New England BioLabs®	M0212S
PvuII	New England BioLabs®	R0151S
BamHI	New England BioLabs®	R0136S
PstI	New England BioLabs®	R0140S
Eco53Ki	New England BioLabs®	R0116S
KpnI	New England BioLabs®	R0142S
SalI	New England BioLabs®	R0138T
SDS-PAGE and Western blotting		
4–20% Criterion™ TGX™ Precast Midi Protein Gel	Bio-Rad	5671094
Trans-Blot Turbo System	BioRad	1704156
Filter paper Whatman™	GE Healthcare Life Technologies	3017-915
PVDF membrane Amersham™ Hybond™	GE Healthcare Life Technologies	10600023
Software		
SnapGene® 4.1.7	SnapGene®	-
NanoDrop 2000	Thermo Scientific™	-
Image Studio™	LI-COR	-
fastqc (Version 0.11.5)	Andrews 2010	-
Trim Galore (Version 0.6.4)	Krueger 2015	-
STAR (Version 2.7.1)	Dobin et al. 2013	-
DESeq2 (Version 1.26.0)	Love et al. 2014	-
Genemania App of Cytoscape	Warde-Farley et al. 2010	-
R (Version 3.6.3)	ggplot2 and pheatmap packages	-

6.6 Appendix 6: Mycoplasma test results

eurofins | Genomics

Table 1: Sample and production details

| Job no. | Identification | | Results | | | |
	Barcode	Cell line name	Testing date	PCR inhib. *	Mycoplasma	Summary
	87928242	87928242	01/10/2019	absent	absent	clean
	87928259	87928259	01/10/2019	absent	absent	clean
	87928266	87928266	01/10/2019	absent	absent	clean
	87928273	87928273	01/10/2019	absent	absent	clean
	87928341	87928341	01/10/2019	absent	absent	clean
	87928372	87928372	01/10/2019	absent	absent	clean
	87928389	87928389	01/10/2019	absent	absent	clean
	87928396	87928396	01/10/2019	absent	absent	clean
	87928402	87928402	01/10/2019	absent	absent	clean
	87928419	87928419	01/10/2019	absent	absent	clean
	87928426	87928426	01/10/2019	absent	absent	clean

* PCR inhibition: present, will result in an invalid mycoplasma test

121