TRANSCRIPTION FACTOR-DNA INTERACTIONS:

INVESTIGATING BINDING SPECIFICITIES IN BASE-RESOLUTION MODELS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Seungha Lee

December 2021

ABSTRACT

The epigenome is a multitude of all molecular interactions that governs genome regulatory activities. Sequence-specific transcription factors (TFs) are the key regulators that control gene expression by binding to specific DNA sequences at promoters or distal enhancers. Understanding how TFs recognize their DNA binding sites forms the basis for understanding mechanisms of transcriptional control. Technological developments have led to the identification of TF binding preferences and revealed how TF-DNA interactions, local DNA structure, and genomic features influence TF-DNA binding. Yet, a precise human epigenome remains largely undefined. Mechanistic insights into how transcription factors recognize their cognate sites across a genome may be best understood through a combination of *in vivo* and *in vitro* genome-wide binding measurements. To this end, I provided the first single base pair measurements of human stem cell factor KLF4 binding to a noncognate DNA sites *in vivo*. To further explore the mechanistic basis for such interactions, I initiated development of an *in vitro* biochemically defined system for measuring protein-DNA interactions genome-wide for human TFs. In the first stage of this process, I developed methods to express human TFs using *in vitro* transcription/ translation systems as well as an *E. coli* T7 expression system. By using His-GFP fusions to these TFs, I tracked their production and purification in real time using GFP's intrinsic fluorescence. I then explored whether His-GFP can be used to establish an *in vitro* genome-wide assay for site-specific DNA binding (PB-exo assay), where nickel resin is used to

retain His-tagged proteins, and GFP is used track binding in real time. Collectively, I demonstrated the feasibility of *in vitro* protein expression and purification as applied towards the PB-exo assay, including using nickel resin in place of antibodies to immunoprecipitated target proteins. The work here provides several concrete steps towards developing a genome-wide assay for in vitro protein-DNA interactions of human TFs.

BIOGRAPHICAL SKETCH

Seungha Alisa Lee was born in South Korea on December 14, 1994. After receiving her elementary education at the Hosu Elementary School in South Korea, she came to the United States by her age 15. Her secondary education was completed at Garrison Forest High school in Owings Mills, Maryland. In 2014, she entered Cornell University the College of Human Ecology, majoring in Human Biology, Health and Society from Cornell University. She continued her education at Cornell University Department of Molecular Biology for her master studies in Genetics.

ACKNOWLEDGEMENTS

This journey would not have been possible without support of many people. I first would like to express my special thanks of gratitude to my advisor, Dr. Frank Pugh, who gave me the opportunity to work on this wonderful project guiding me from time to time in making this project. During the time with him lessoned me a lot to explore a wide range of experiments and helped me to build solid foundational knowledge in biochemical to bioinformatic skills. Secondly, I would like to thank my special committee members, Dr. Andrew Grimson and Dr. Hojoong Kwak who were always willing to give sincere advice on every subject that I came up with. Any attempt at any level can't be satisfactorily completed without the support and guidance of my parents and friends. I especially thank my parents who always give me tremendous support on every step I made and endured this process with me by always offering support and love. Finally, I would like to thank God, for letting me through all the difficulties that I face. I have experienced your guidance day by day and will keep on trusting you for my future steps.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Sequence-specific transcription factors (TFs) are functional regulators that bind to transcriptional regulatory regions like promoters and enhancers to control the expression of their target genes. Each TF typically recognizes a collection of similar DNA sequences, which is represented as binding site motifs using models such as position weight matrices (PWM) [1]. The characterization of motif helps to understand the regulatory functions of TFs that consequently shape gene regulatory networks.

In recent years, many technological developments have characterized binding preferences of TFs to DNA by revealing thousands of TF motifs from wide range of organisms including humans [2-7]. Most of these large-scale studies have highlighted the evolutionary conservation of TF binding specificity, measuring the binding preferences of TFs and their mechanisms by which they find their site *in vivo* [4,6,7]. However, still the binding preferences remain unknown for over 40% of the approximately 1,400 sequence-specific TFs encoded in the human genome [3,7-12]. Recent high-throughput studies have highlighted that there is more to TF-DNA binding than primary nucleotide sequence preferences. In addition to the direct sequence readouts, interaction between cofactors and TFs is also discovered as a major factor that alters sequence preference [13]. Beyond direct DNA sequence readout, DNA shape is also discovered as a significant determinant that modulates sequence recognition. A particular DNA

shape can arise from multiple possible arrangements of nucleotide bases that create a distinct shape of the sugar phosphate backbone [14,15]. Standard motif discovery methods, such as MEME and HOMER [16,17], rely on position weight matrices but not DNA shape. Previous studies discovered that combing DNA shape analysis with traditional position weight matrices improved the predictions of *in vivo* binding [45-48]. Yet, the role of DNA shape itself had limited experimental investigation within the physiological context of an entire genome [13,14].

Different methods for characterizing different features that contributes to TF-DNA recognition help to enhance our understanding of what determines condition-specific TF binding, yet still more needs to be addressed. Here, I will discuss methods for identifying TF binding site motifs, emerging knowledge of additional features that impacts TF-DNA specificity, and novel approaches that our lab developed in characterizing TF-DNA binding *in vivo* and *in vitro*.

CHAPTER 2

ASSAYS FOR MEASURING SITE SPECIFIC BINDING ON GENOMIC SCALE

Methods to characterize TF-DNA binding preferences can be categorized into in vivo and in vitro approaches.

A widely used *in vivo* method to detect protein-DNA interactions is chromatin immunoprecipitation followed by high throughput DNA sequencing (ChIP-seq) [15]. ChIP-seq measures for genome wide profiling of DNA-binding proteins, histone modifications or nucleosomes. To briefly explain, genomic regions bound by a TF of interest are isolated via immunoprecipitation, and the bound sequences are identified through high-throughput sequencing. ChIP-seq signal peaks are typically inferred through peak calling algorithms and then analyzed with software such as MEME-ChIP or HOMER [16,17]. Hundreds of ChIP-seq dataset have been generated, particularly by the ENCODE Consortium [18], providing data on cell type specific TF binding events. However, ChIP-seq presents several key challenges for determining TF binding site motif. ChIP-seq peaks can span dozens to hundred bases, while the actual binding site motifs for most TFs are shorter than 10 bp [15]. Due to such inaccuracy, it is hard to perform precise mapping of binding sties, particularly when binding sites are clustered in proximity.

As alternatives to ChIP-seq, DNase-seq, ATAC-seq, and FAIRE-seq were developed which identifies regions of accessible chromatin using a non-specific

DNA nuclease, transposase, or formaldehyde crosslinking coupled with phenol-chloroform extraction, respectively [21-25]. DNase-seq, mapping DNase I hypersensitive (HS) sites, is a method of identifying the location of active gene regulatory elements. It captures DNase-digested fragments across the whole genome and sequencing them by high-throughput next generation sequencing [53]. Assay for Transposases Accessible Chromatin with high-throughput sequencing (ATAC-seq) is another method that probes DNA accessibility with hyperactive Tn5 transposases, which inserts sequencing adapters into accessible regions of chromatin. The sequencing reads can then be used to infer regions of increased accessibility, and map regions of transcription factor binding and nucleosome position [54]. Additionally, Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) is another assay that can simply isolate nucleosome-depleted DNA from chromatin, which DNA recovered in the aqueous phase is hybridized to a DNA microarray. This assay has utility as a positive selection for genomic regions associated with regulatory activity, including regions traditionally detected by nuclease hypersensitivity assays [55]. ATAC-seq and DNase-seq report on regions that are nucleosome-free, hence accessible to nucleases. Yet, FAIRE-seq exploits the fact that protein-free DNA is not extracted into the organic phase of phenol extraction, whereas protein-DNA crosslinks (from formaldehyde) are extractable. None of these existing assays are protein specific but are valuable for overall assessment of a protein-DNA landscape.

Current in vivo methods are constrained in their ability to identify motifs de novo.

In vivo methods to identify direct TF-DNA interactions sites in high resolution heavily relies on the quality of existing TF binding specificity models. Sequences of the bound regions are probed to or against to the results of *de novo* motif finding analysis. Furthermore, motifs have been identified using *in vitro* methods that can systematically assess many possible targets sequence to identify a distinct TF. Current existing *in vitro* methods include bacterial one-hybrid(B1H) selection [26], protein binding microarrays (PBMs) [27,28], PB-seq [29], or in vitro selection-based (eg. SELEX) approaches [30,31]. Particularly, SELEX assay has been able to characterize the binding of TF pairs and complexes [32,33,34] and genomic context PBMs (gcPBMs) have discovered the contributions of flanking genomic sequences to motif recognition [28]. Therefore, the combination of *in vivo* and *in vitro* approaches has led to insights on features that influence TF-DNA binding interactions. To have a better understanding of the complexities of TF binding determinants in high resolution, our lab has developed two ChIP-seq variant assays, ChIP-exo (*in vivo*) and PB-exo (*in vitro*), which help to compressively examine DNA sequence in binding site selectivity.

High throughput ChIP-exo assay

To overcome widely used ChIP-seq's limitations, our lab has developed an advanced version of ChIP-seq, ChIP-exo, which removes excess sequences with lambda exonucleases allowing nearly nucleotide-resolution mapping of binding sites [20]. ChIP-exo helps to map the binding location of genome-interacting proteins in near-bp resolution with less background noise [35]. Involvement of lambda exonuclease helps to achieve single bp resolution by hydrolyzing each strand of duplex DNA in the 5'-3' direction, stopping at about 6 bp from the point of crosslinking (**Fig 1.1**) [35]. DNA sequences downstream of the exonuclease stop-site remain intact and are sufficiently long to uniquely map to a reference genome. A pair of opposite-strand peaks is thus generated for each crosslink. They are separated in the 3' direction from each other by ~12 bp. This makes binding detection quite discriminative, meaning that in principle and, unlike ChIP-seq, some proteins bound a few bp apart are resolvable [35]. Protein complexes will produce multiple direct and indirect crosslinks, including some that occur when in the vicinity of DNA and not just through site-specific contacts. This

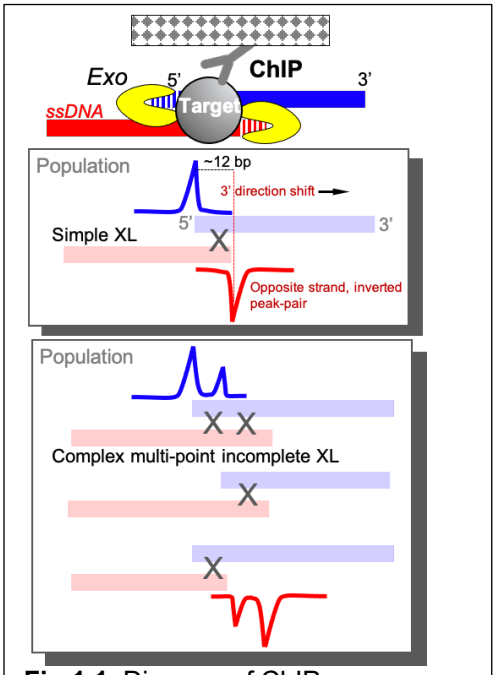

Fig 1.1. Diagram of ChIP-exo patterning. 5'-3' ssDNA exonucleases digest up to protein-DNA crosslinks. Sequencing tag 5' ends are generated, whose frequency is plotted relative to individual genome-wide strand-specific reference points.

results in pattern complexity in bulk measurements. Nonetheless, ChIP-exo produces an information-rich tag distribution pattern that is often characteristic (and diagnostic) of that protein. This is advantageous in discerning true binding from noise when ChIP'ing "difficult" proteins. By streamlining the assay to simplicity, we have mapped the precise positional organization of more than 400 different genome-interacting proteins in yeast and mammalian cell by ChIP-exo. Our lab has assayed over a thousand antibodies against putative transcription regulatory proteins in many cell lines including, K562, HepG2 and MCF7, and additional ChIP-exo datasets have been produced for human and mouse tissue organs. Most of the tested antibodies are low-cost and renewable DSHB-sourced antibodies that produced by The Protein Capture Reagent Project (PCRP) [36], whereas the others have commercial sources (Abcam, SCBT, etc.), making this assay more versatile. The utility of an antibody in ChIP-exo assay was evaluated by sequencing depth, library complexity, and the ability to generate significant peaks at sequence motifs, and enrichment at annotated regions (promoter, enhancer, insulators, TF binding sites, and transcription start sites) [36]. For example, CTCF and USF1 each have very strong ChIP-exo peaks around their cognate motifs (**Fig 1.2**).

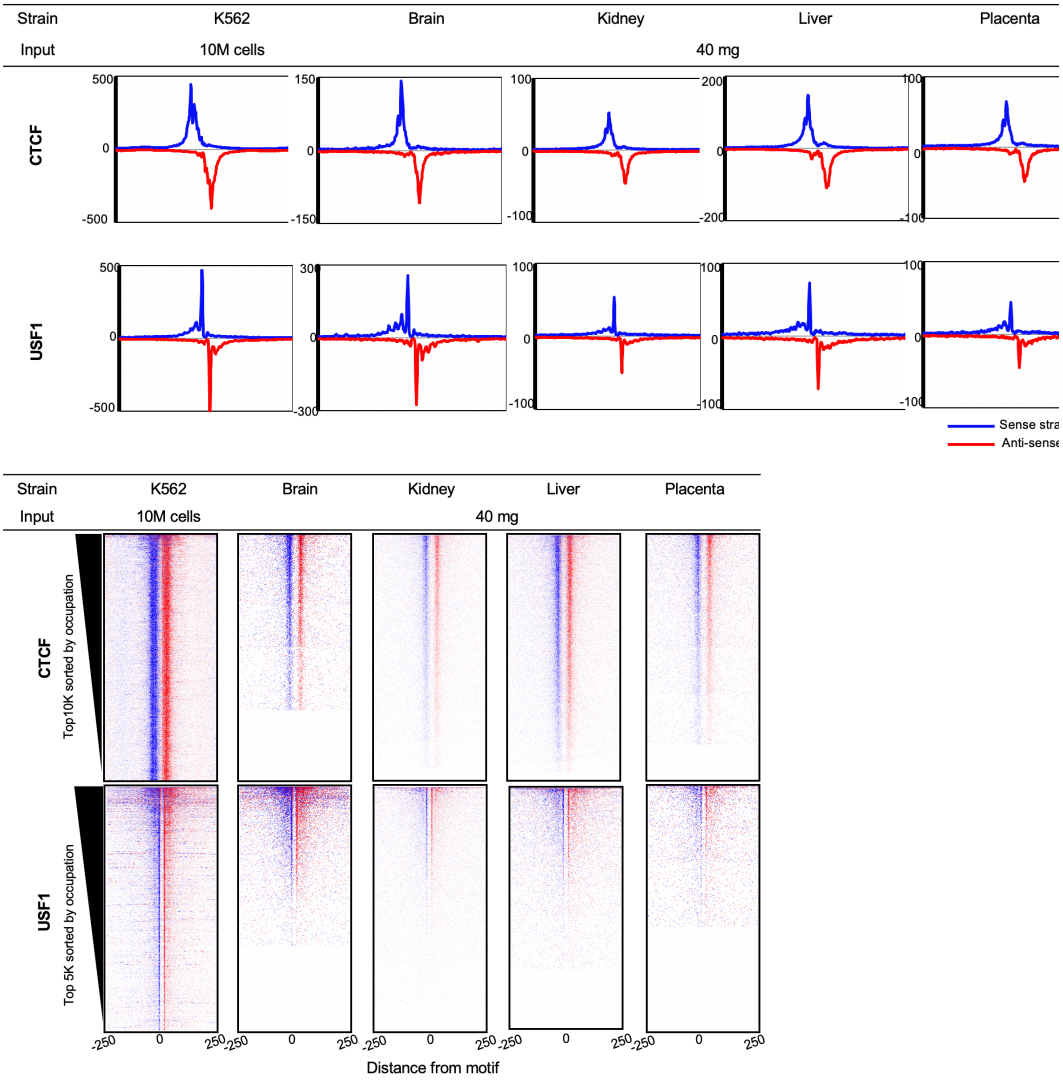

Fig 1.2 Demonstrating ChIP-exo applicability by showing composite plots and heatmaps for CTCF and USF1 in K562 cell and four different human organs by reproducing existing dataset. The above panel shows the average ChIP-exo signal at 10,000 cognate binding sites. The bottom panels show the same data but at individual binding sites [Data origin from Haining Chen].

CHAPTER 3

UNUSUAL SITE-SPECIFIC BINDING OF KLF4 IN VIVO

High throughput ChIP-exo of pluripotency factor, KLF4, in NCCIT cells: KLF4 has a unique binding region that surroundings of cognate binding sites which have unique sequence characteristics

Our lab has successfully produced yeast epigenome maps at single bp resolution [37], specifically defining the position of ~80 sequence-specific TFs, interacting cofactors, and active nucleosome depletion. This is also attainable for human epigenome by the same ChIP-exo workflows and data analysis. My project has focused in part on providing a more precise view of how the pluripotency factors are arranged and cooperate with one another at promoter or enhancers regions in human embryonic stem cells (hESCs). This project was done in collaboration with Dr. Cedric Feschotte's lab in the Department of Molecular Biology and Genetics at Cornell University. We aimed to map the architecture of the cis-regulome (site-specific protein-DNA interactions) of pluripotency factors in human embryonic stem cells using ChIP-exo. To gain a more precise view of how the pluripotency factors are arranged with one another at promoter/enhancers in hESCs, we used the pluripotent human embryonic carcinoma cell line (NCCIT) model, which are a teratoma derived cancer cell line that can act as an experimental model for ESC.

Methods

NCCIT chromatin preparation

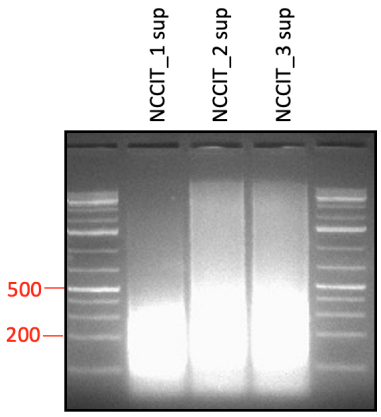

Fig 2.1. 2% agarose gel of the electrophoresed chromatin fragmentation check for NCCIT cells. Sonicated in Bioruptor (Diagenode) for 10 cycles with 30 s on/off intervals.

From Dr. Feschotte's lab, I received a of total 300 million of NCCIT cells. I cross-linked them with formaldehyde at a final concentration of 1 % for 10 min at room temperature, and quenched with 3 M Tris-HCl for 5 min. The supernatant was removed, and the cells were resuspended in 1 ml of PBS to wash. The washed cells were aliquoted to contain 50 million cells, centrifuged, removed supernant, and the pellet was flash frozen. From our hand, 50 million cell aliquots (for use in multiple ChIPs, typically 10 million cells/ ChIP) was lysed in 500 µl (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.5% NP40, and complete protease inhibitor (CPI, Roche)) by incubating on ice for 10 min. The lysate was microcentrifuge at 2500 rpm for 5 min at 4 °C. The supernatant was removed, the pellet resuspended in 1 ml (50 mM Tris-pH 8.0, 10 mM EDTA, 0.32% SDS, and CPI), and incubated on ice for 10 min to lyse the nuclei. The sample was diluted with 600 µl of immunoprecipitation dilution buffer (IP Dilution Buffer: 20 mM Tris-HCl, pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, and CPI) to a final concentration of (40 mM Tris-HCl, pH 8.0, 7 mM EDTA, 56 mM NaCl, 0.4% Triton-X 100, 0.2% SDS, and CPI), and sonicated with a Bioruptor (Diagenode) for 10 cycles with 30 s on/off intervals to obtain DNA fragments 100 to 500 bp in

size. (**Fig 2.1**). From this, it was determined that the fragmented chromatin was in a suitable size range for ChIP.

Chromatin immunoprecipitation (ChIP)

10 million cell equivalents of NCCIT chromatin were diluted to 200 µl with IP Dilution Buffer and incubated overnight at 4 °C with the appropriate antibody that were selected based on literature reports of successful antibodies **(Table 1**). A 10 µl bed volume of Dynabeads A/G was added to the to the NCCIT samples, and samples were incubated with mixing overnight.

Table 1. Source of used antibodies.

Antibody	Company	Catalog#	Reactivity	Source	Reference
NANOG	active motif	61419	Mouse, human	Rabbit,poly	https://www.activemotif.com/catalog/details/61419/nanog-antibody-pab
OCT4	active motif	39811	Mouse, human	Rabbit,poly	https://www.activemotif.com/catalog/details/39811/oct-4-antibody-pab
SOX2	active motif	39843	Mouse, human	Rabbit,poly	https://www.activemotif.com/catalog/details/39843.html
KLF4	proteintech	11880-1-AP	Mouse, human	Rabbit,poly	https://www.ptglab.com/products/KLF4-Antibody-11880-1-AP.htm?gclid=Cj0KCQiAhZT9BRDmARIsAN2E-J1Hd0kesccS_h8nSRSrMG9VyRH4lez55ke7ZxA5BGSLB_cZ3PP1j34aAkFPEALw_wcB#publications
ERG1	cell signaling	4153S	Mouse, human	Rabbit, mono	https://www.cellsignal.com/products/primary-antibodies/egr1-15f7-rabbit-mab/4153
SP1	santacruz	sc-17824X	Mouse, human, rat	Mouse, mono,	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4050696/

Chromatin elution from antibody beads

After immunoprecipitation, the following steps were carried out on the resin. The ChIP material on resin was washed sequentially with FA Lysis Buffer, NaCl Buffer, LiCl Buffer, and 10 mM Tris-HCl, pH 8.0 at 4 °C. For the A-tailing reaction (50 µl) containing: 15 U Klenow Fragment, -exo (NEB), 1 × NE Buffer 2, and 100 µM dATP was incubated for 30 min at 37 °C; then washed with 10 mM Tris-HCl, pH 8.0 at 4 °C. The first adapter ligation and kinase reactions (45 µl)

containing: 1200 U T4 DNA ligase, 10 U T4 PNK, 1 × NEB Next Quick Ligation Buffer, and 375 nM adapter (ExA2_iNN / ExA2B) was incubated for 1 h at 25 °C; then washed with 10 mM Tris-HCl, pH 8.0 at 4 °C. The fill-in reaction (40 µl) containing: 10 U phi29 polymerase, 1 × phi29 reaction buffer, 2 × BSA, and 180 µM dNTPs was incubated for 20 min at 30 °C; then washed with 10 mM Tris-HCl, pH 8.0 at 4 °C. The λ exonuclease digestion (50 µl) containing: 10 U λ exonuclease, 1 × λ exonuclease reaction buffer, 0.1% Triton-X 100, and 5% DMSO was incubated for 30 min at 37 °C; then washed with 10 mM Tris-HCl, pH 8.0 at 4 °C.

Reverse cross-linking and PCIA extraction

DNA was eluted from the resin, and reverse cross-linking and Proteinase K treatment were performed (40 µl) containing: 30 µg Proteinase K, 25 mM Tris-HCl, pH 7.5, 2 mM EDTA, 200 mM NaCl, and 0.5% SDS incubated for 16 h at 65 °C.

AMPure purification

The supernatant was then transferred to a new tube and purified with Agencourt AMPure magnetic beads (Beckman Coulter) following manufacturer's instructions (1.8 × volume). The sample was eluted from the AMPure beads in 20 µl of water, and the following enzymatic steps were carried out in solution.

2nd adaptor ligation

For second adapter ligation (total reaction volume 40 µl) was performed to the resuspended DNA was added 1200 U T4 DNA ligase, 1 × T4 DNA Ligase Buffer, 375 nM adapter (ExA1-58/ExA1-SSL_N5) and was incubated for 1 h at 25 °C.

The ligation reaction was then purified with AMPure beads (1.8 × volume) and resuspended in 15 µl of water.

PCR amplification

The sample was then amplified via PCR. For PCR amplification (total reaction volume 40 µl); to the resuspended DNA was added 2 U Phusion Hot Start polymerase (Thermo scientific), 1 × Phusion HF Buffer (Thermo scientific), 200 µM dNTPs, 500 nM each primer (P1.3 and P2.1) and amplified for 18 cycles (20 s at 98 °C denature, 1 min at 52 °C annealing, 1 min at 72 °C extension). A quarter of the reaction was amplified for an additional six cycles (24 total) and the presence of libraries was determined by electrophoresis on a 2% agarose gel (**Fig 2.2**). Libraries were in the size range of 200-500 bp, with minimal adapter dimers. We therefore concluded that ChIP-exo libraries were of the correct size and thus properly constructed.

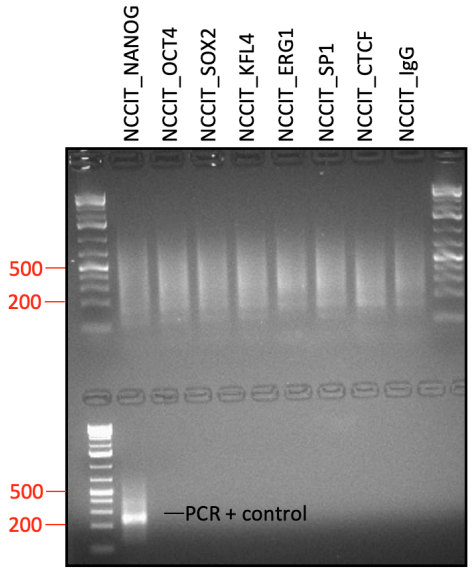

Fig 2.2. 2% agarose gel of the electrophoresed library following 18 cycles of PCR for NCCIT human pluripotency transcription factors (NANOG, OCT4, SOX2, KLF4, ERG1, SP1) and controls (CTCF, IgG) assayed by ChIP-exo in 10 M NCCIT cells each. Size markers (bp) are shown to the left. Empty lanes after PCR positive control in the second row.

Size selection

200 to 500 bp PCR products were gel-purified from a 2% agarose gel using the QIAquick Gel Extraction Kit (Qiagen).

DNA sequencing

High-throughput DNA sequencing was performed with a NextSeq 500 in paired-end mode producing 2 × 40 bp reads. Sequence reads were subsequently aligned to the human (hg19) genomes using bwa-mem (v0.7.9a) [49]. Aligned reads were filtered to remove non-unique alignments and PCR duplicates. PCR duplicates were defined as sequence reads possessing identical Read_1 and Read_2 sequences.

Data analysis

ChIP-exo libraries were sequenced on the NextSeq 500 (**Fig 2.3**). For the purposes of this study, I focused my analysis on KLF4, as its binding specificity was unusual and of significant interest compared to the control analysis for CTCF and IgG. I obtained 21 million paired-end sequencing reads, which was sufficiently deep coverage (12.7 M deduplicated reads) to make accurate peak-pair calls. From this, and with low stringency peak calling, I identified approximately 1 million peak pairs (low threshold binding events). This low threshold established the minimal number of detectable binding locations.

Mapping Statistics

Number of Tags | Percentage of Tags | Additional Information

Sample ID	Sequence Run	Genome	Read Count	Mapped Read Count	Uniquely Mapped Count	Deduplicated Count
29617 sal6030-201203-04S KLF4	509 (Run Name 846)	hg19	21,296,424	18,466,554	16,545,253	12,709,424

Downstream Analysis

Peak Statistics

Sample ID	Target	Sequence Run	Genome	Peak Calling (parameters)	All Peaks	Singletons	Peak-pairs (parameters)	# of Peak-pairs	Non-paired peaks
29617 sal6030-201203-04S KLF4	KLF4	509 (Run Name 846)	hg19	S5e10F0	12,751,133	12,478,016	u0d80b2	1,134,784	10,481,565

Fig 2.3 Mapping and peak calling statistics from Illumina sequencing.

Next, using the top-most 500 peak pairs (occupancy or density of sequencing reads), I conducted MEME for de novo motif discovery. What I discovered was an unexpected binding model for KLF4 that was distinct from the algorithmic predictions from previous ChIP-seq analysis. Many ChIP-seq studies represent sequence logos of KLF4 as GGGCGTGGT predicted by various *in vitro* affinity models, such as THerMos, Matrix REDUCE, Weeder, MEME, DREME and ChIPMunk [50]. In contrast, based our MEME based analysis of ChIP-exo peaks for KLF4, I discovered the following motif: TCTCGCGAGA (**Fig 2.4**).

Fig 2.4. Comparison with detected in vitro binding motif logos for KLF4 (left) and algorithmics predicted logo from a published ChIP-seq study (right) [50].

I next ascertained whether binding to the published cognate site could be detected by plotting the distribution of sequencing tag 5' ends (exonuclease stop sites) around the cognate KLF4 motif. To do this I used FIMO and a published position weight matrix of KLF4 sites [56] to find all instances through the genome. However, no enrichment of tags was found there (**Fig 2.5**), which demonstrates that KLF4 was not crosslinking and possibly not binding to its cognate motif.

Instead, it appeared to be crosslinking to a different site. I cannot rule out the possibility that the commercial KLF4 antibody is incorrect, despite receiving a properly labeled antibody tube.

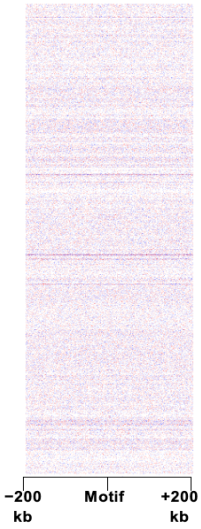

Fig 2.5. ChIP-exo tag distribution for KLF4 around motif instances of the published KLF4 site [56] in the human genome. Total number of tags were 54 K.

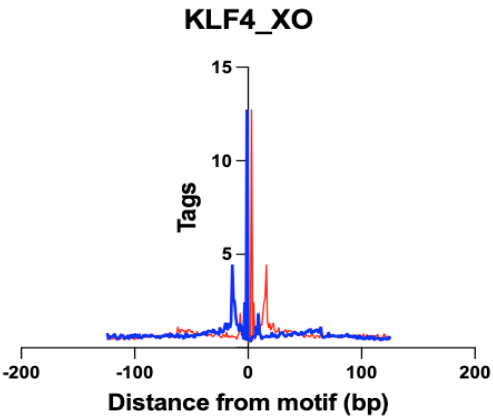

Fig 2.6. A composite plot for pluripotency factor, KLF4 in NCCIT cell; The total read counts was 21M reads, 18M reads were mapped, and among those total 16M was uniquely mapped.

To assess whether KLF4 binding was indeed site specific to the discovered motif, I plotted the distribution of exonuclease stop sites around the discovered motif (**Fig 2.6**). Clear ChIP-exo patterning was detected, with ~10 bp peak residing on either side of the motif midpoint. This indicates a single predominant point of crosslinking of KLF4 (or any interacting partner) to the motif. I therefore conclude that the interaction with this motif is site-specific, and a set of binding events have been discovered.

As the bound locations for KLF4 for ChIP-exo was different from the expected motif, this may be an indicator of having an additional factor that drives KLF4 binding interactions in site specific manner. This result can be interpretated as KLF4 not crosslinking efficiently to its true cognate motif, and instead crosslinking to other partner proteins that have their own cognate motif. As some TFs can recruit epigenetic factors such as chromatin remodelers or modifiers to alter chromatin states, I concluded that KLF4 may be interacting with other site-specific binding proteins that interact with the discovered motif. Such interplay between KLF4 and putative potential co-factor together can lead to dynamic gene expression regulation as binding to the novel KLF4 site.

How theses mechanisms play across a genome has not yet been fully understood because of the complexity of site-specific DNA binding mechanisms, which involves interactions with other proteins. To circumvent this limitation, I next choose to study factor binding on a genomic scale using an *in vitro* version of ChIP-seq assay that may probe distinct intrinsic aspects of DNA binding without other co factor present and provide the detailed investigation of ChIP cross-linking bias. By comparing *in vivo* with *in vitro* system, it will address why site-specific DNA recognition may differ from predicted binding in an isolated in vitro system.

CHAPTER 4

SYSTEMS FOR TF PRODUCTION AND PURIFICATION FOR IN VITRO GENOME-WIDE BINDING

The unusual DNA binding properties of KLF4, prompted me to explore the site specificity of KLF4 in a defined system without other proteins that might alter its specificity. To overcome such limitations, our lab has developed an *in vitro* version of ChIP-exo assay named Protein Binding with deep sequencing (PB-exo). PB-exo is a genome-wide approach to examine site-specific DNA binding in vitro without other co factor present. In developing PB-exo as an in vitro version of ChIP-exo, sheared genomic DNA is incubated with purified general regulatory factors to make the system as the "purest" condition, then treated to the standard ChIP-exo protocol (**Fig 3.1.**). This *in vitro* assay provides information about transcription binding across a genome in the absence of cellular impacts such as involvement of other binding co-factors or chromatin structural changes. This system is highly reproducible and adaptable as purified proteins can be replaced to whole-cell extract (WhIP-

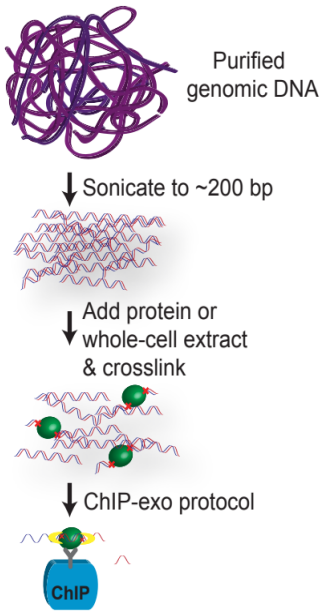

Fig 3.1. Schematic of PB-exo. Purified, sonicated genomic DNA is incubated with purified protein and cross-linked with formaldehyde. That material is then immunoprecipitated and treated as starting material in the ChIP-exo assay.

exo) to study DNA assembly of complex mixtures of proteins [35]. The source of DNA will be a human system to avoid any complications associated within the DNA. With this setup, PB-exo is expected to be applicable to various potential binding sites that can be screened with in a single reaction. Application of this assay will help to understand complex mechanisms of site-specific DNA interactions. As PB-exo is performed on purified protein and sheared gDNA crosslinked with formaldehyde, obtaining protein is the key step in this assay. Here, I introduced species independent translation initiation systems to express and characterize N-terminally GFP tagged human proteins of different sizes in *E. coli* and HeLa cell free systems. Using a combination of fluorescence light detection and SDS-PAGE analysis, I assessed the expression yields and the fraction of full-length translation product.

Methods

Cell-free expression system and In Vitro protein expression.

Cell-free protein expression is a rapid and high-throughput methodology for conversion of genetic information into protein-mediated biochemical activities. It relies on the self-directed nature of cellular protein translation machinery that retains sufficient biosynthetic activity upon cell disruption and fractionation [38]. Short time preparation and high expression yields enable it to be use in mammalian system to perform PB-exo.

TF-expressing plasmids

In this study, TF expression plasmid libraries were constructed that could easily manipulate the production of TF. Our three interests of proteins, GATA1, STAT3, and USF2 were selected based on their biological significance and the presence of ChIP-grade validated antibodies. They were also a prelude to KLF4, as their binding in vivo is well-characterized. GATA1 is a protein coding gene that plays an important role in erythroid development by regulating the switch of fetal hemoglobin to adult hemoglobin. STAT3 regulates host response to viral and bacterial infections, which mutations are associated with autoimmune diseases. USF2 is involved in regulating multiple cellular processes that are associated with renal dysplasia and Cowden syndrome. Lastly, sfGFP were expressed as a control set. These interest of TF plasmids, GATA1 (Addgene #67068), STAT3 (Addgene #67120), USF2 (Addgene #67069), sfGFP (Addgene #67045) were constructed in frame with pCell Free_G03 vector, T7 promoter with N-terminal His and EGFP tagged (**Fig 3.2**). The resulting plasmid

constructs were amplified by mini preparation with the Qiagen Spin Miniprep Kit (Qiagen #27104) and their proper construction confirmed by restriction digestion mapping [52] (**Fig 3.3**). I determined that the bp sizes of the bands migrating in the gel are as predicted from the schematic of the plasmid illustrated in (**Fig 3.3)**.

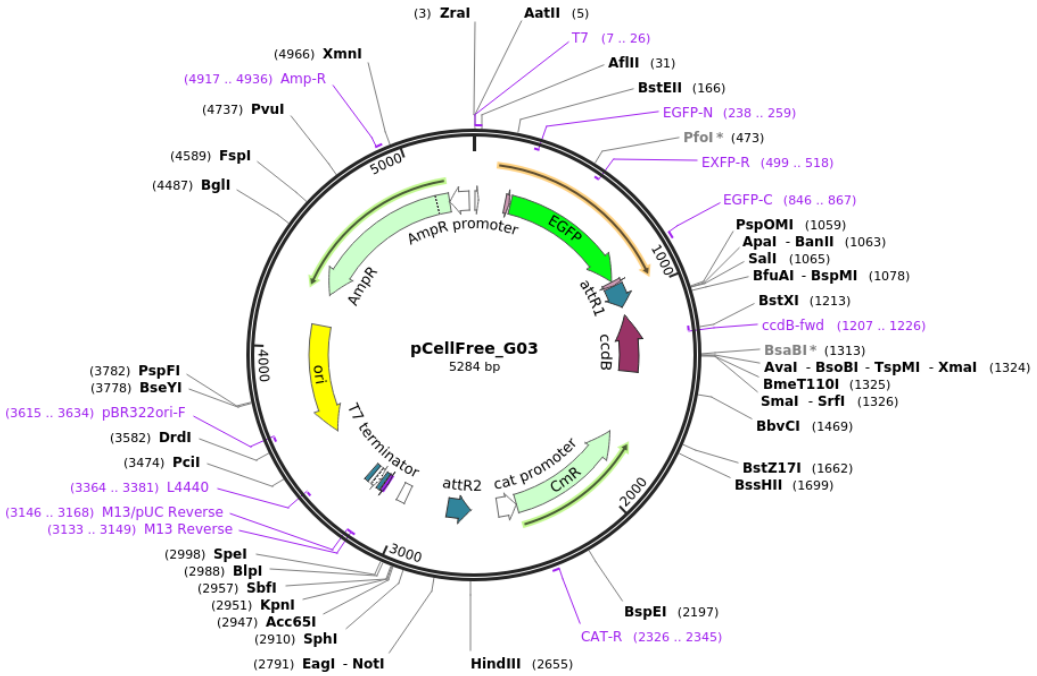

Fig 3.2. Schematic of expression plasmids**.** Empty backbone vector for cell-free expression construct with N-term 8xHis eGFP tag is shown.

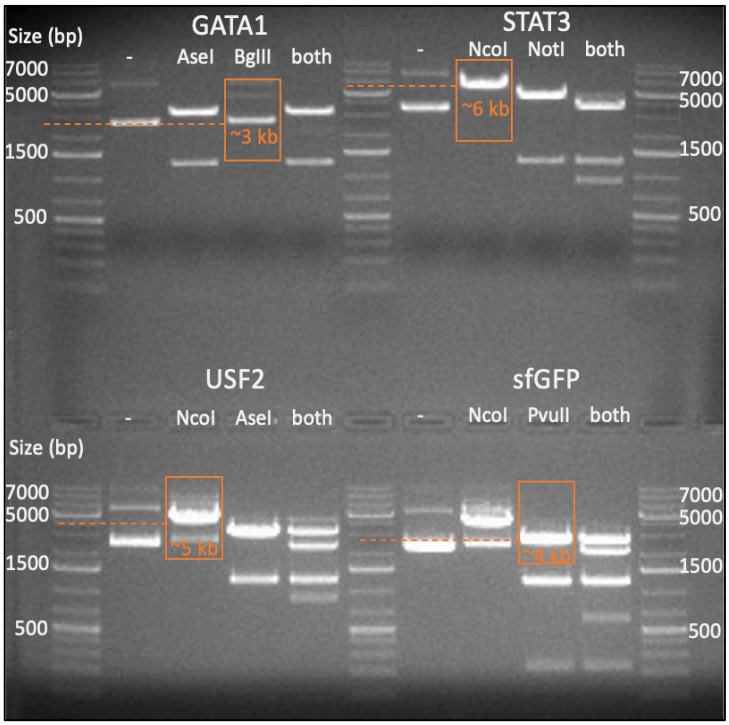

Fig 3.3. 2% agarose gel of the electrophoresed for restriction digestion mapping for extracted plasmid DNAs (GATA1, STAT3, USF2 and sfGFP).

Production and Detection of HeLa cell-based cell-free protein expression

The human coupled IVT (in vitro transcription/translation) protein expression system (Thermofisher, #88881) enables the translation and post-transcriptional modification of full-length proteins from plasmid templates. This system uses the human translational machinery to express active proteins by using HeLa cell lysate protein expression systems (**Fig 3.4**). HeLa cell free extracts can express proteins such as functional enzymes, phosphoproteins, glycoproteins, and membrane proteins with post-translational modification in single step that could immediately use in studying protein interactions *in vitro* [39]. To generate proteins of interest, 1-step Human Coupled IVT kit (Thermofisher, #88881) was used. 12.5 ul of HeLa lysate, 2.5 ul of accessory proteins, 5 ul of reaction mix, and 1 ug of plasmid DNA was thawed on ice. The reactions were prepared in a single tube (total volume of reaction 25 ul) at room temperature and incubated at 30 °C for 2 hrs. The protein expression level was quickly visualized by placing the tubes directly under the blue light were GFP protein could be visualized at ex/em: 482/512 nm. As shown in **Fig 3.5**, the positive control GFP produced bright fluorescence, whereas the negative control lacking the plasmid vector produced only low levels of background fluorescence. Both STAT3, USF2, and GATA1 (not shown) produced significant fluorescence, indicating that these TFs were successfully produced. I therefore conclude that in-tube fluorescence detection is a rapid means by which to detect production IVT proteins, in a rapid and semi-quantitative manner.

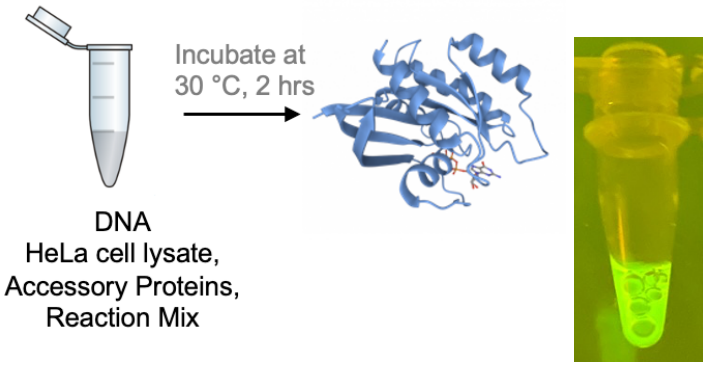

Fig 3.4. Scheme of HeLa cell lysate-based protein expression system (1-step human coupled IVT kit) for in vitro translation (left). The Fluorescent light detection after IVT reaction for USF2 protein; 30 °C, 2 hrs incubation at 30 °C; visualized under wv 465 nm (right).

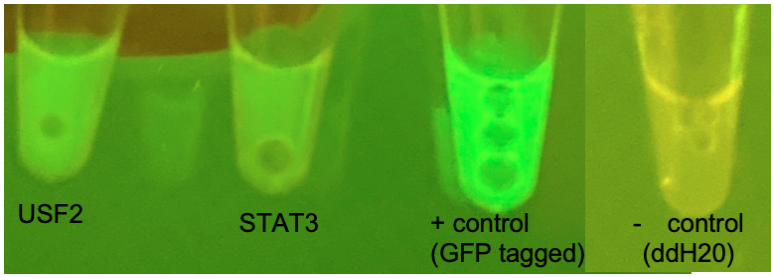

Fig 3.5. Fluorescent light detection after IVT reaction for USF2, STAT3 proteins with pT7CFE1 expression based GFP tagged vector as a positive control protein and water as a negative control; 30 °C, 2 hrs incubation at 30 °C; visualized under wv 465 nm (right).

E. coli-based cell-free protein expression system

The pET system is a powerful system for cloning and expression of large quantities of recombinant proteins in *E coli* (typically more than what is achievable by IVT). Here, I introduce the pET system to express the interest of proteins, STAT3, USF2, and GATA1, to produce them in large scale with cost effective manner to have enough purified proteins for preforming PB-exo from which information about site-specificity of binding can be determined and which overcomes limitation of in vivo assays. In the pET expression system, target

genes are cloned in pET plasmids under control of strong bacteriophageT7 transcription and translation signals. Expression is induced by providing a source of T7 RNA polymerase in the host cell. By using source of selective T7 RNA polymerase, ideally the desired product can generate proteins after few hours of induction. Here, I used BL21(DE3) as a competent E. *coli* cell (cat#:C2527H) to get the transformed bacterial cells (**Fig 3.6**). Transformation was successful as evidenced by many colonies on each plate, whereas no colonies were observed when a no DNA control was used.

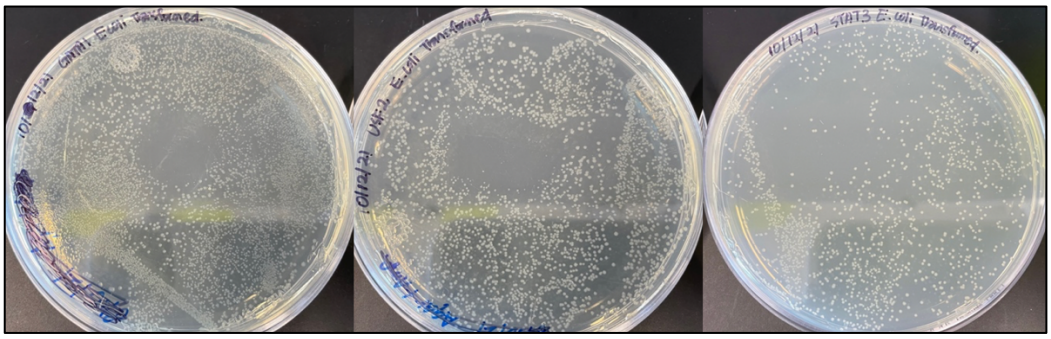

Fig 3.6. Bacterial transformation with target genes (from left: GATA1, USF2, STAT3). Negative control (no plasmid) gave no colonies.

A colony of BL21 Gold from an LB-agar plate with ampicillin (final concentration 100 ug/ml) was used to inoculate 5 mL of LB. The starting culture was grown overnight at 37 °C and used to inoculate 15 mL of LB medium culture. This culture was grown to an OD 600=4.5 at 37 °C. Cells were then harvested by centrifugation for 15 min at 2500 g and washed twice with PBS. The cell pellet was resuspended in PBS and the supernatant were saved from the cell debris by 30 min centrifugation at 30,000 g. The saved supernatant cells were later induced with 0.1mM of IPTG in 15 ml of LB media, and the expression was monitored by different induction times, 0,1,5 hrs. The samples were followed by

the SDS-PAGE analysis and assessed the expression yields and the fraction of full-length translation product.

Coomassie stained bands having the molecular weight expected of the expressed proteins was observed (**Fig 3.7**). However, bands of equivalent migration and relative intensity were not observed in other lanes where other TFs were expressed. Bands in other lanes that are consistent across all samples are background *E. coli* proteins. This indicates that the *E. coli* expression system was successfully over-expressing USF2, STAT3, and GATA1. Each protein molecular weight was confirmed as their sizes corresponded to what is reported in the literature [52].

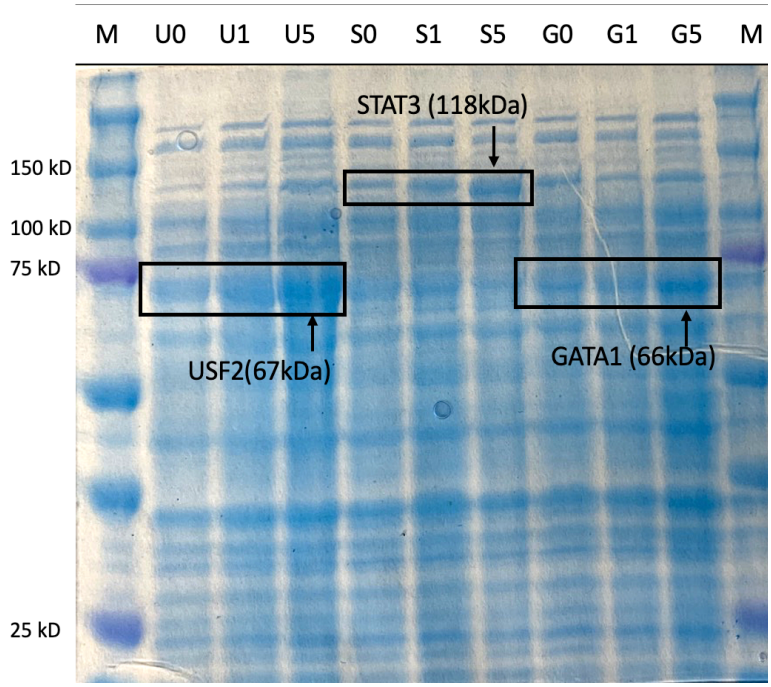

Fig 3.7. SDS page gel analysis for protein expression in E. coli cells with different IPTG induction periods (0,1,5hr); U: USF2, S: STAT3, G: GATA1.

CHAPTER 5

EXPLORATION OF THE USE OF NICKEL RESIN IN PB-EXO

Tracking protein localization throughout the PB-exo reaction steps

Obtaining a purified protein is one of the key components in performing successful PB-exo. To purify our recombinant His-tagged proteins, I used a metal affinity chromatography (IMAC) method, consisting of chelating resins charged with nickel ions and purify by coordinating with the histidine side chains of our recombinant His -tagged proteins (Thermofisher, #88221). Here, I have developed methodologies to visualize our N terminal His and GFP tagged protein of interest in real time during purification, using GATA1, USF2 and STAT3 plasmid constructs (**Fig 4.1**). Theoretically, a fully expressed GFP tagged protein should be optically detected at 510 nm wavelength, and under the appropriate wavelength, we can observe the protein's expression level immediately by detecting the emission of green, fluorescent light.

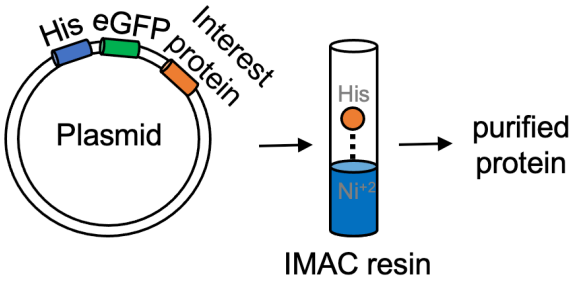

Fig 4.1. Scheme of obtaining protein purification using Nickel resin.

Potential use of Nickel resin in place of antibody resin.

The rationale for this approach is to exploit the high efficiency of nickel/His-tag to more efficiently capture TFs that are crosslinked to genomic DNA. While in principle GFP antibodies may also be used, in our experience GFP antibodies have low efficiencies at least in the presence of formaldehyde treated GFP.

Nickel resin that captures His tagged proteins are mixed with His- and GFP-tagged proteins of interest. If the GFP tagged protein binds to the resin in sufficient quantities, then it may be possible to detect the bind by enhanced fluorescence that is associated with the resin. To detect where proteins are localized throughout the PB-exo process, each of the following steps was subjected to detection under blue light. This process was explored using the TF USF2. After each step, the samples went under the blue light to capture the green light (**Fig 4.2**).

Methods

Nickel resin purification Steps:

1. IVT rxn: Expressed USF2-His-GFP and His-GFP *in vitro* using plasmid DNAs. This is the reference state of expected fluorescence. Essentially a positive control.
2. Binding rxn: Attached 8 ug of sheared human gDNA to expressed USF2-His-GFP that was extracted from K562 cells.
3. XL + quench: Formaldehyde crosslinking of USF2-His-GFP and His-GFP followed by quenching. The purpose was to evaluate whether formaldehyde, which will be used in PB-exo to covalent trap protein-DNA interactions, destroys GFP fluorescence. As shown, formaldehyde does not inhibit GFP fluorescence.
5. Ni resin: Attached the pre-equilibrated nickel-based beads to USF2-His-GFP and His-GFP (30 min of incubation). The purpose here was to evaluate

whether chelation with Nickel resin quenched the fluorescence signal, which did not.

6. Post SD: Centrifugation (“SD”) of the resin after incubation with crosslinked USF2-His-GFP and His-GFP. The purpose here was to evaluate whether the nickel resin could still interact USF2-His-GFP and His-GFP after they have been crosslinked. Surprisingly, USF2-His-GFP failed to interact with resin, as it remained in the supernate, whereas His-GFP was concentrated in the pellet. This suggests that while crosslinking is not intrinsically inhibitory to resin capture, the presence of crosslinked USF2 makes it inhibitory. Thus, whether a crosslinked protein can be captured on nickel resin may be protein specific.
6. Two washes were done, using 20 mM imidazole, to remove low affinity background. No fluorescence was detected in the supernate indicating that 20 mM imidazole does not strip off His-GFP proteins.
7. Two elutions were then performed using 250 mM imidazole. His-GFP was observed to be released into the supernate.

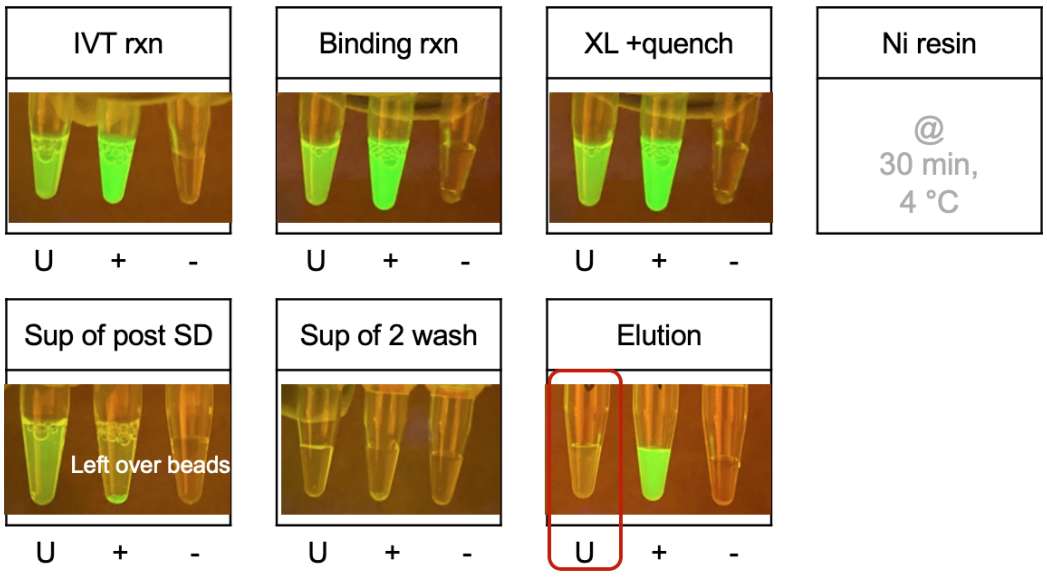

Fig 4.2. Visualizing protein localization by detecting green, fluorescent light under 465nm. U: USF2, *Positive control (+)*: pT7CFE1-GFP- CHis, which is 6x His tag at the C-terminus (provided within the Kit). *Negative control (-)*: ddh20.

In summary, I was able to detect the green, fluorescent light and could confirm the condition for IVT reactions. Also, by looking at the "Sup post SD" figure, the left-over beads were also emitting lights which showed that the proteins are well attached to the beads so that the beads themselves are emitting the lights. However, still some lights were seen during the wash steps. While visualizing the protein's localization throughout the generation process, we wanted to know if crosslinking with formaldehyde interferes the protein structure that might affect the expression level. I determined that it does interfere for USF2-His-GFP, but this may not be the general case because it does not interfere with His-GFP.

In the next phase of PB-exo development, the experiments in **Fig 4.2** should be repeated but in the presence of genomic DNA. The idea is that DNA bound USF2-His-GFP may be protected from inhibitory effects of formaldehyde crosslinking on nickel resin retention. To do this, I performed a procedure to

isolate genomic DNA. I first isolated high molecular weight DNA, which is insoluble. I then sheared the DNA to small fragments, which also solubilized it.

Purification of fragmented human genomic DNA

Human genomic DNA was obtained from 25 million K562 cells using the protocol for "Purification of Total DNA from Animal Blood or Cells (Spin Column)" with the optional RNAse treatment from a Qiagen Blood and Tissue Extraction Kit. The extracted genomic DNA was sonicated in Reaction Binding Buffer (20 mM HEPES- KOH, pH 7.5, 50 mM KCl, 5 mM MgCl2, 100 µg/ml BSA, 1 mM DTT) using Diagenode Pico (10 cycles of 30 sec on/ 30 sec off pulses) to produce DNA fragments. When visualized on a 2% agarose gel, most of the DNA molecules ranged from 200 to 500 bps (**Fig 4.3**). Extracted DNAs were quantified with Qubit and flash freeze into 8 µg aliquots.

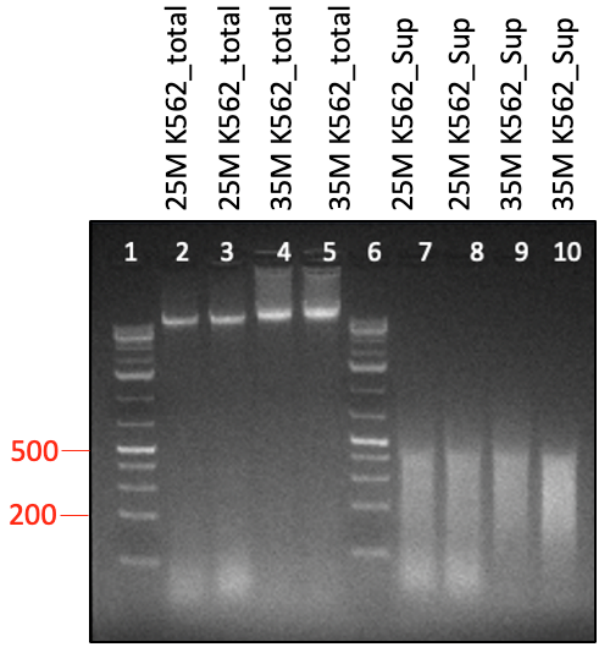

Fig 4.3. 2% agarose gel of the electrophoresed for gDNA extraction from 25 M and 35 M K562 cells; total: total extract, sup: supernatant.

CHAPTER 6

DISCUSSION

Design of PB-exo experiment

Genome-wide binding of USF2 or any other TF throughout the human genome in a purified defined system has two main requirements including a source of purified protein and purified genomic DNA (**Fig 5)**. To date I have expressed multiple TFs in vitro using two different expression systems: IVT using Hela cell extracts, and E. coli expression systems. Both are purifiable through nickel resin due the presence of a His tag, although such purification has not yet been demonstrated in my hands. Genomic DNA has now been purified and is ready for the PB-exo assay.

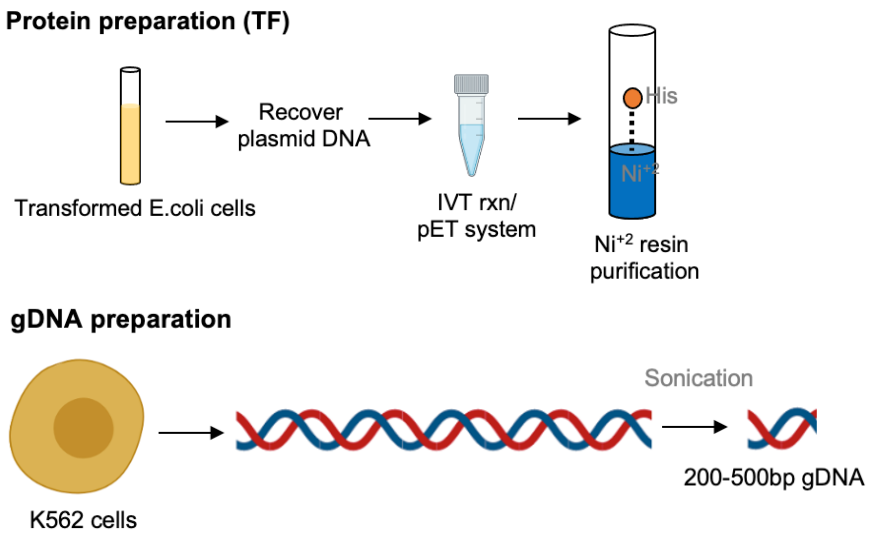

Fig 5. Summary of schematic view of PB-exo preparation steps. The way to obtain human gDNA (upper) and purify interest of proteins (lower).

Possible interpretation: Comparing In vitro assay to in vitro assay for the same TF

If the TF factor was detected both *in vivo* and *in vitro*, then this factor is likely to have an intrinsic ability to bind to its cognate site, without use of other cofactors. But if site-specific binding of a TF factor was only detected in vivo but not in in vitro, then a cofactor would be required to obtain site specificity. Yet, there may be a region where a factor is only detected in vitro, we can predict that this site can be blocked by other proteins which explains why there was not the binding events detected. Lastly, if a factor was bound nowhere, we can predict that MEME or FIMO, which are commonly used motif detection programs that we also use to analyze, are limited as they do not capture all the info of DNA. And in this is the case where we can predict that intrinsic factors may be significantly involved in this motif, which DNA structure changes can be involved.

Determining how binding specificity is achieved

DNA binding specificity arises where proteins recognize nonuniform properties of DNA. Such nonuniformity originates from physical and chemical properties of bases and sequential order. While nonuniformity is distinctly established through direct base readout (i.e., hydrogen bonding of protein amino acid side chains with an ordered arrangement of DNA bases), it is now appreciated that nonuniformity is further exhibited indirectly by base stacking and composition of the shape of a chemically uniform sugar-phosphate backbone. Base stacking and backbone conformation may be manifested through a wide range of parameters such as roll, propeller twist, helical twist, minor groove width, and

sugar pucker, etc., of which some may be computationally predicted through DNA base pentamer sequences [14].

While a specific DNA sequence is expected to produce a single principal intrinsic shape, the reverse is not always necessarily true. Rather, a specific DNA shape may arise from many different sequence mixtures. Also, neighboring sequences may contribute to DNA shape. Neighboring effects are not well-captured in MEME because each nucleotide position is compiled independently of the identity of neighboring nucleotides and MEME logos do not clearly capture shape information. Based on our previous findings in yeast organism, it suggests how DNA shape is contained within motif regions having low sequence definition. Shape specificity is likely important within regions of direct sequence readout, where the motif base sequence is well-defined. However, we conclude that separating direct from indirect readout within those regions is not feasible using genomic data alone, but these distinctions are illuminated when combined with atomic-level structural information of protein-DNA interactions [14]. We have found in yeast that when bound by a GRF (general regulatory factor), the DNA is forced to adopt a unique and specific conformation change of DNA. These twists, bends, and duplex deformations occur at specific base pairs and in a specific direction within the motif and correspond to positions with the greatest differences in DNA shape between bound and unbound motif occurrences. Our genomic data show that true GRF binding sites are comprised of a combination of DNA sequence and DNA shape elements and that deviations in either will modulate the affinity of the GRF for

DNA [14]. No evidence currently exists that shape alone can drive site-specific regulatory protein binding, as with most biological processes, the regulation of protein binding to DNA is a continuum and many possibilities remain to be tested.

CHAPTER 7

CONCLUSION

Concluding remarks

The human genome contains approximately 25,000 genes that encode all the proteins and other molecules that make up an individual cell. The epigenome is a compilation of chemical modifications that affect how cells in different parts of the body use the human genome to form different types of cells and tissues. As the epigenomic modifications do not affect the DNA sequence itself, it influences specific genes expression, turning on or off, which can be altered in response to diseases. Many studies were done to made progress in epigenomics field, such as Human Genome Project, which decoded the complete sequence of the human genome, or DNA methylation studies [10] which researchers have found different approaches for detecting the changes in methylation across the entire human genome [40-42]. Beyond methylation of individual DNA nucleotides, finding patterns of histone modification- another type of epigenetic modification- was actively studied as well [13,14,15]. To gain a comprehensive view of how epigenomics contributes to human biology and disease, further extensive studies are needed. Although much development was made, we currently do not have a complete or a high-resolution understanding of the molecular architecture of genes and genome regulatory proteins that assemble across a human genome. Precisely defining where all genome regulatory proteins are bound across the genome at sufficient resolution will be the key element to

unravel the positional organizations of within the complex human system. What is specifically challenging is that genome regulatory architecture differs from one organ, tissue, and cell type to another. Until now, there have been large NIH supported projects like ENCODE, or GTEx to generate a comprehensive mapping of genome regulation. ChIP-seq is a commonly used approach to address such questions but have relatively low resolution and large signal to noise ratio. For this reason, the use of ChIP-exo and its *in vitro* cousin, PB-exo, may be of use in defining the human epigenome more completely and at higher resolution.

REFERENCES

1. Stormo, Gary D. "Modeling the specificity of protein-DNA interactions." *Quantitative biology* 1.2 (2013): 115-130.
2. Franco-Zorrilla, José M., et al. "DNA-binding specificities of plant transcription factors and their potential to define target genes." *Proceedings of the National Academy of Sciences* 111.6 (2014): 2367-2372.
3. Hume, Maxwell A., et al. "UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions." *Nucleic acids research* 43.D1 (2015): D117-D122.
4. Jolma, Arttu, et al. "DNA-binding specificities of human transcription factors." *Cell* 152.1-2 (2013): 327-339.
5. Narasimhan, Kamesh, et al. "Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities." *Elife* 4 (2015): e06967.
6. Nitta, Kazuhiro R., et al. "Conservation of transcription factor binding specificities across 600 million years of bilateria evolution." *elife* 4 (2015): e04837.
7. Weirauch, Matthew T., et al. "Determination and inference of eukaryotic transcription factor sequence specificity." *Cell* 158.6 (2014): 1431-1443.
8. Vaquerizas, Juan M., et al. "A census of human transcription factors: function, expression and evolution." *Nature Reviews Genetics* 10.4 (2009): 252-263.
9. Wang, Jie, et al. "Factorbook. org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium." *Nucleic acids research* 41.D1 (2012): D171-D176.
10. Kulakovskiy, Ivan V., et al. "HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models." *Nucleic acids research* 44.D1 (2016): D116-D125.
11. Mathelier, Anthony, et al. "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles." *Nucleic acids research* 44.D1 (2016): D110-D115.
12. Inukai, Sachi, Kian Hong Kock, and Martha L. Bulyk. "Transcription factor–DNA binding: beyond binding site motifs." *Current opinion in genetics & development* 43 (2017): 110-119.
13. Siggers, Trevor, and Raluca Gordân. "Protein–DNA binding: complexities and multi-protein codes." *Nucleic acids research* 42.4 (2014): 2099-2111.
14. Rossi, Matthew J., William KM Lai, and B. Franklin Pugh. "Genome-wide determinants of sequence-specific DNA binding of general regulatory factors." *Genome research* 28.4 (2018): 497-508.
15. Furey, Terrence S. "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions." *Nature Reviews Genetics* 13.12 (2012): 840-852.
16. Ma, Wenxiu, William S. Noble, and Timothy L. Bailey. "Motif-based analysis of large nucleotide data sets using MEME-ChIP." *Nature protocols* 9.6 (2014): 1428-1450.

17. Zhou, Tianyin, et al. "Quantitative modeling of transcription factor binding specificities using DNA shape." *Proceedings of the National Academy of Sciences* 112.15 (2015): 4654-4659.
18. ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (2012): 57.
19. Rhee, Ho Sung, and B. Franklin Pugh. "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution." *Cell* 147.6 (2011): 1408-1419.
20. Boyle, Alan P., et al. "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells." *Genome research* 21.3 (2011): 456-464.
21. Buenrostro, Jason D., et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature methods* 10.12 (2013): 1213-1218.
22. Hesselberth, Jay R., et al. "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting." *Nature methods* 6.4 (2009): 283-289.
23. Song, Lingyun, et al. "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity." *Genome research* 21.10 (2011): 1757-1767.
24. Giresi, Paul G., et al. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin." *Genome research* 17.6 (2007): 877-885.
25. Meng, Xiangdong, Michael H. Brodsky, and Scot A. Wolfe. "A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors." *Nature biotechnology* 23.8 (2005): 988-994.
26. Berger, Michael F., et al. "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities." *Nature biotechnology* 24.11 (2006): 1429-1435.
27. Berger, Michael F., and Martha L. Bulyk. "Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins." *Gene mapping, discovery, and expression*. Humana Press, 2006. 245-260.
28. Gordân, Raluca, et al. "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape." *Cell reports* 3.4 (2013): 1093-1104.
29. Guertin, Michael J., et al. "Accurate prediction of inducible transcription factor binding intensities in vivo." *PLoS genetics* 8.3 (2012): e1002610.
30. Jolma, Arttu, et al. "Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities." *Genome research* 20.6 (2010): 861-873.
31. Roulet, Emmanuelle, et al. "High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites." *Nature biotechnology* 20.8 (2002): 831-835.
32. Slattery, Matthew, et al. "Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins." *Cell* 147.6 (2011): 1270-1282.

33. Jolma, Arttu, et al. "DNA-dependent formation of transcription factor pairs alters their binding specificity." *Nature* 527.7578 (2015): 384-388.
34. Sefah, Kwame, et al. "Development of DNA aptamers using Cell-SELEX." *Nature protocols* 5.6 (2010): 1169-1185.
35. Rossi, Matthew J., William KM Lai, and B. Franklin Pugh. "Simplified ChIP-exo assays." *Nature communications* 9.1 (2018): 1-13.
36. Blackshaw, Seth, et al. "The NIH Protein Capture Reagents Program (PCRP): a standardized protein affinity reagent toolbox." *Nature methods* 13.10 (2016): 805-806.
37. Lai, William KM, et al. "Screening of PCRP transcription factor antibodies in biochemical assays." *bioRxiv* (2020).
38. Rossi, Matthew J., et al. "A high-resolution protein architecture of the budding yeast genome." *Nature* 592.7853 (2021): 309-314.
39. Rosenblum, Gabriel, and Barry S. Cooperman. "Engine out of the chassis: cell-free protein synthesis and its uses." *FEBS letters* 588.2 (2014): 261-268.
40. Weber, Michael, et al. "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." *Nature genetics* 37.8 (2005): 853-862.
41. Guertin, Michael J., and John T. Lis. "Mechanisms by which transcription factors gain access to target sequence elements in chromatin." *Current opinion in genetics & development* 23.2 (2013): 116-123.
42. Guertin, Michael J., et al. "Accurate prediction of inducible transcription factor binding intensities in vivo." *PLoS genetics* 8.3 (2012): e1002610.
43. Rohs, Remo, et al. "The role of DNA shape in protein–DNA recognition." *Nature* 461.7268 (2009): 1248-1253.
44. Slattery, Matthew, et al. "Absence of a simple code: how transcription factors read the genome." *Trends in biochemical sciences* 39.9 (2014): 381-399.
45. Slattery, Matthew, et al. "Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins." *Cell* 147.6 (2011): 1270-1282.
46. Gordân, Raluca, et al. "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape." *Cell reports* 3.4 (2013): 1093-1104.
47. Zhou, Tianyin, et al. "Quantitative modeling of transcription factor binding specificities using DNA shape." *Proceedings of the National Academy of Sciences* 112.15 (2015): 4654-4659.
48. Agius, Phaedra, et al. "High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions." *PLoS computational biology* 6.9 (2010): e1000916.
49. Wilson, Mandy L., et al. "Sequence verification of synthetic DNA by assembly of sequencing reads." *Nucleic acids research* 41.1 (2013): e25-e25.
50. Sun, Wenjie, et al. "TherMos: estimating protein–DNA binding energies from in vivo binding profiles." *Nucleic acids research* 41.11 (2013): 5555-5568.
51. Dror, Iris, et al. "A widespread role of the motif environment in transcription factor binding across diverse protein families." *Genome research* 25.9 (2015): 1268-1280.

52. Gagoski, Dejan, et al. "Performance benchmarking of four cell-free protein expression systems." *Biotechnology and bioengineering* 113.2 (2016): 292-300.
53. Song, Lingyun, and Gregory E. Crawford. "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." *Cold Spring Harbor Protocols* 2010.2 (2010): pdb-prot5384.
54. Buenrostro, Jason D., et al. "ATAC-seq: a method for assaying chromatin accessibility genome-wide." *Current protocols in molecular biology* 109.1 (2015): 21-29.
55. Giresi, Paul G., et al. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin." *Genome research* 17.6 (2007): 877-885.
56. Kulakovskiy, Ivan V., et al. "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis." *Nucleic acids research* 46.D1 (2018): D252-D259.