TESIS O PROYECTO DE CREACIÓN

APROBADO COMO REQUISITO PARCIAL DEL
PROGRAMA DE ESTUDIOS DE HONOR

COMITÉ DE TESIS O PROYECTO DE CREACIÓN	NOMBRE	FIRMA
Mentor(a)	José A. Rodríguez Martínez	
Director(a) de estudios	Ivelisse Rubio Canabal, Ph.D.	
Lector(a)	Esther A. Peterson-Peguero	
Lector(a)	Anabel Puig Ramos, Ph.D	
Lector(a)		
Visto Bueno	Dra. Elaine Alfonso Cabiya Director(a) PREH o su Representante	

Fecha

14 Ave. Universidad, Ste. 1401, San Juan, PR 00925
Tel. 787-764-0000 Ext. 85060
programa.honor@upr.edu

Uncovering the DNA-binding properties of the GATA4 and TBX5 cardiac transcription factor complex

A Senior Thesis Presented
By

EMILI PATRICIA ROSADO RODRÍGUEZ

Submitted to the Honors Program of the

University of Puerto Rico Río Piedras Campus

December 2020

Molecular and Cellular Biology

Uncovering the DNA-binding properties of the GATA4 and TBX5 cardiac transcription factor complex

Emili P. Rosado Rodríguez[1], Jessica Rodríguez Ríos[1], José Rodríguez Martínez[1]
[1] University of Puerto Rico, Río Piedras Campus, San Juan, P.R.

Abstract

Transcription factors (TFs) are essential gene regulators of cellular differentiation in human development. Eukaryotic TFs are notorious for often binding DNA as multimeric protein complexes to regulate gene expression. GATA4 and TBX5 are transcription factors that are central components of the gene regulatory network of human heart development and function. Recent studies have determined the binding specificity of these transcription factors as monomers. However, the DNA-binding properties of the cooperative complex between GATA4 and TBX5 remain undetermined. Based on this, we wanted to know the intrinsic DNA-binding preferences of the cooperative complex formed by GATA4 and TBX5. We determined the *in vitro* DNA-binding specificity of the GATA4:TBX5 complex using Systematic Evolution of Ligands by Exponential Enrichment (SELEX-seq). Our preliminary results demonstrate that the TF monomeric binding motifs differ from the DNA-binding sequences recognized by the heteromeric complex. The GATA4:TBX5 cooperative complex also showed spacing and orientation preferences. We are still analyzing the SELEX-seq data using Autoseed and R programming. The findings of this study will help to understand the DNA-recognition rules of the GATA4:TBX5 complex and its potential roles in normal heart development.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Background and Justification

Transcription factors (TFs) are sequence specific DNA-binding proteins that activate or suppress the transcription of genes. TFs are determinants of cellular state and have been shown to control cell differentiation. These proteins recognize specific DNA sequences through what is known as the DNA-binding domain (DBD). This structure is responsible for the classification of these proteins into several families depending on their DBD. Usually, the DNA-binding proteins belonging to a certain family, recognize similar consensus DNA sequences or motifs. Still, the DNA recognition properties of individual proteins within a single family can vary due to spatial and temporal arrangements (Lambert et al., 2018; Stormo, 2013; Stormo and Zhao, 2010). Therefore, evaluating the DNA-binding specificity of TFs is a challenge that we must overcome to decipher gene regulatory networks controlled by a given transcription factor.

It has been demonstrated that eukaryotic TFs can work as monomers but usually form heteromeric complexes with other TFs to control gene expression (Jolma et al., 2015; Wilkinson et al., 2017; Luna-Zurita et al., 2016; Morgunva and Taipale, 2017; Siggers and Gordân, 2014). These cooperative complexes allow the distinction of unique DNA sequences that are different from the individual binding motifs (Jolma et al., 2010, 2013, 2015). More importantly, it has been demonstrated that TF interdependence prevents ectopic binding and activation of incorrect genes (Luna-Zurita et al., 2016; Ang et al., 2016). To accomplish specificity, TF complexes have unique spatial and temporal patterns of arrangements that do not correspond to the addition of the individual properties of monomeric TFs. Recognizing the lack of systemic characterization of transcriptional cooperation, scientists are beginning to focus on its study. The importance and intricacy of transcription explain why mutations on TFs are linked to multiple diseases (Bass et al., 2015). Not surprisingly, it has also been shown that cardiac transcription factors often coregulate genes during heart development and some of their mutations cause congenital heart defects.

Congenital Heart Defects (CHD) are a group of malformations present at birth that affect heart structure. In the United States and Puerto Rico (PR), CHD occurs in approximately 1% of live births, with similar prevalence worldwide; it is also the leading cause of mortality after birth (Triedman and Newburger, 2016; Departamento de Salud PR, 2014). More interesting, of the total CHD cases in PR, approximately 54.8% correspond to ventricular septal defects (VSD) and atrial septal defects (ASD) (Puerto Rico Health Department, 2017).

These types of CHDs have been associated with a regulatory network that includes transcription factors such as GATA4 and TBX5 (Maitra et al., 2010).

1.2 Relevance and Innovation

Studying the synergistic properties of complex formation during transcription is a challenge that must be overcome to completely characterize their regulation mechanism, understand their role during cellular development and explain their connection to diseases. Our research will establish a new model of cardiac TF complexes and provide greater insight into their DNA-binding specificity. This investigation provides data and analyses that will support future research to expand our knowledge of gene regulation. Our data will allow us to make better predictions of gene regulatory networks driving heart development and will be fundamental for biomedical applications to treat CHD in the future. This is why all the data collected in this research will be shared with the scientific community.

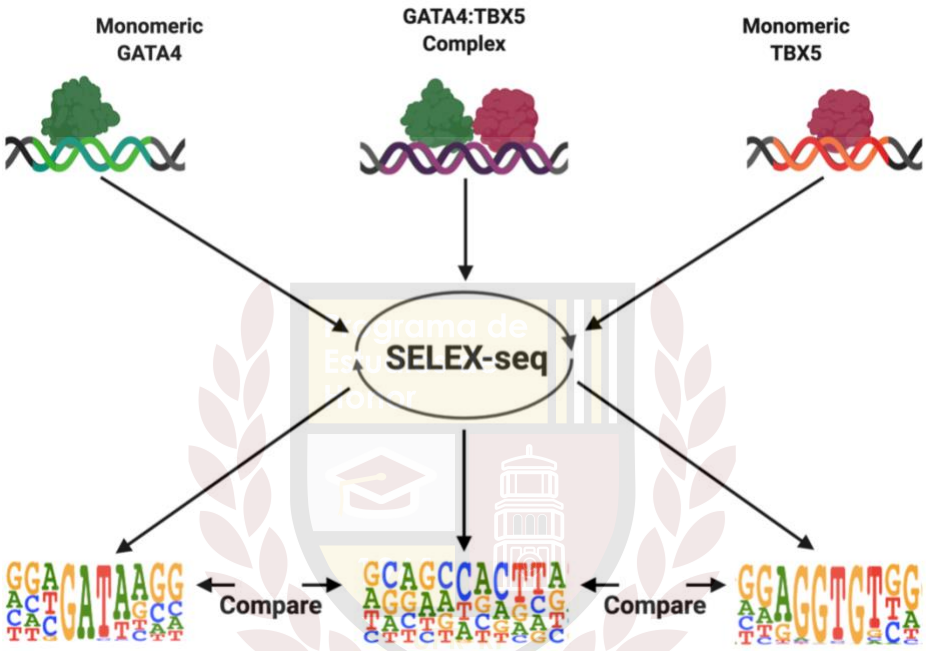
1.3 Problem and Hypothesis

Although there is evidence about the cooperative interactions between GATA4 and TBX5, there is a lack of research focused on determining the complex's DNA recognition grammar rules such as the spacing and orientation between each protein's binding sequence. Consequently, our main objective was to uncover the DNA-binding sequences of this cooperative complex to better understand the grammar rules (spacing and orientation) that govern its specificity. Our **central hypothesis** stated that the DNA-binding sequences recognized by the GATA4:TBX5 complex differ from the specific DNA sequences preferred by the monomeric TFs. Additionally, we hypothesized that the GATA4:TBX5 complex has strong spacing and orientation preferences.

1.4 Research Questions and Specific Aims

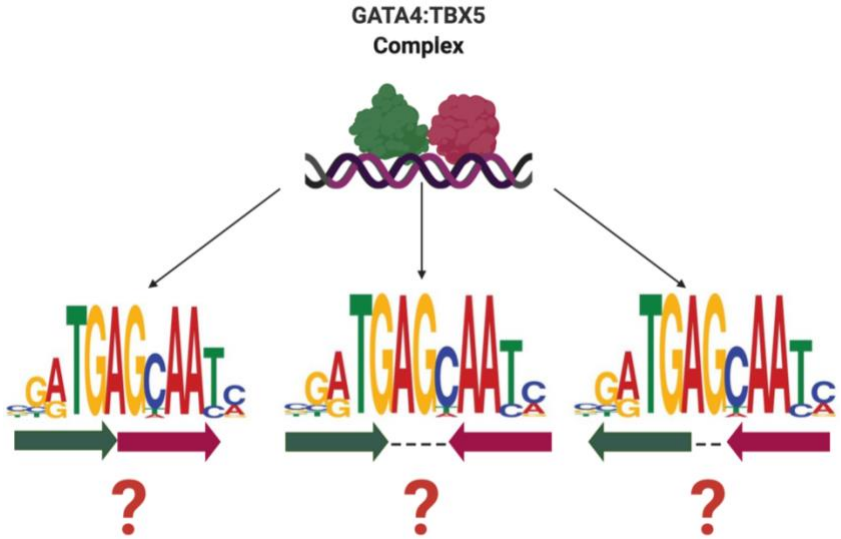
<u>Research question 1</u>: Do the specific DNA-binding sequences of the GATA4:TBX5 complex differ from the individual TF motifs?

<u>Aim 1</u>: Determine and compare the specific DNA-binding sequences of: monomeric GATA4, monomeric TBX5, and the heteromeric GATA4:TBX5 complex using SELEX-seq.



<u>Research question 2</u>: Does the GATA4:TBX5 complex show strong spacing and orientation preferences?

<u>Aim 2</u>: Determine if the GATA4 and TBX5 DNA-binding motifs have strong orientation and spacing preferences in complex.



2. REVISED LITERATURE

2.1 Transcription Factors Regulate Gene Expression

The central dogma of molecular biology states that the genetic information encoded by DNA is transcribed into messenger RNA molecules that will ultimately be translated into polymers of amino acids (Lodish et al., 2000). These amino acids are the building blocks of proteins, which are the molecules that perform essential functions within cells. Similar to many biological fundamentals, the central dogma is a dynamic concept influenced by many exceptions. For example, retroviruses have the ability to retro-transcribe RNA into DNA, a process known as reverse transcription. Despite its variations, gene transcription is one of the most regulated biological processes. Concerted gene expression is responsible for cell-type specification during development and adult tissue homeostasis (Wilkinson et al., 2017). Transcription factors (TFs) are sequence-specific DNA-binding proteins that bind to promoters and enhancers to activate or suppress gene expression. They can recruit co-activators or co-repressors, and displace histones to control gene transcription. They are so important for cellular differentiation and function that it has been estimated that the human genome encodes ~1,700-1,900 potential TFs (Lambert et al., 2018; Vaquerizas et al. 2009). Some of the cellular programs regulated by TFs include: metabolism, immunity, reproduction, cellular proliferation and organ development (Marson et al., 2007; Jia et al., 2016; Zhang et al., 2019; Kumar-Yadav et al., 2018).

Transcription factors recognize specific DNA sequences through their DNA binding domain (DBD). Based on their DBDs, these proteins can be classified into several structural families. A few common protein families are: zinc fingers, helix-loop-helix, homeodomains, and basic leucine zipper proteins (Luscombe et al., 2000). Despite this general classification, “DBDs are rarely identical, indicating the possibility that small differences in protein sequence could lead to significant differences in binding specificity” (Slattery et al., 2011). The DNA-binding specificities of a particular TF can be represented as “motifs”, which are models representing the short DNA sequences preferred by a protein (Lambert et al., 2018). TF’s specificity depends on DNA base and DNA shape readout. The first term (‘base readout’) describes the hydrogen bonds and hydrophobic molecular interactions that drive the nucleotide-protein interactions. On the other hand, transcription factors can also differentiate among multiple DNA structural features such as bending and winding of DNA strands, a quality referred to as ‘shape readout’. (Slattery et al., 2014; Kribelbauer et al., 2020). As briefly discussed in this section, gene transcription is a sophisticated biological program which continues to be rigorously studied.

2.2 Transcription Factors Bind DNA as Multiprotein Complexes

Scientists have discovered that ubiquitous TFs can control general cellular machineries as monomers but need to combine with other proteins to regulate tissue-specific genes. According to Luna-Zurita et al. (2016), interdependent binding serves not only to co-regulate gene expression, but also to prevent TFs from distributing to ectopic loci and activating lineage-incorrect genes. Transcription factor cooperativity allows the recognition of composite sites that are markedly different from the monomeric TF's motifs. The cooperativity can be the result of various mechanisms. For example, TFs could weakly bind to each other in solution and then strengthen their interaction after binding to DNA (Sánchez et al., 1997). Another possible mechanism is that TFs cooperate without any direct and specific protein-protein interaction (Vashee et al., 1998). A single TF dimer can bind to multiple DNA motifs given that the recognition sites of its individual TFs can occur in different orientations and/or spacings relative to each other. Recent large-scale studies have revealed that the dimeric mode of binding is more common than previously appreciated. Even proteins that were known to bind DNA as monomers, can also form dimers with specific orientation and spacing preferences (Jolma et al., 2013; Siggers and Gordân, 2014).

In their 2015 study, Jolma et al. identified that only 5% of 3,630 TF pair interactions appeared to bind to DNA independently of each other. However, 95% of those TF pairs did bind DNA in a co-dependent manner, as indicated by the presence of both expected motifs with strong orientation and spacing preferences. Besides, their results demonstrated that most of the TF dimer bound sites had a large overlap between the individual TF recognition motifs. In contrast, if the most enriched motif pair had a gap, two or more spacings were more commonly observed. This suggests that for our bioinformatics analysis we must identify if the GATA4:TBX5 complex shows strong orientation and spacing preferences. If so, it would mean that their binding to DNA is co-dependent.

Systematic destabilization of combinatorial TF binding is commonly altered if one of the co-bound TFs is mutated (Stefflova et al., 2013). In a recently published study, Kribelbauer et al. (2020) showed that experimenting with mutated TFs in a complex-specific manner can provide insights into the genome-wide binding and function of heteromeric TF complexes. Accordingly, understanding how multiple transcription factors regulate gene expression is essential to characterize how different mutations alter their synergy and give rise to human diseases (Bass et al., 2015; Wilkinson et al., 2017; Jiménez-Sánchez et al., 2001).

2.3 Structure and Function of GATA4

GATA4 is a zinc-finger DNA-binding protein, which means that it is coordinates zinc ions that help stabilize its structure. This particular protein recognizes the consensus DNA motif: 5'-WGATAR-3' (Ang et al., 2016). GATA 4 is central for cardiomyocyte (CM) proliferation and septal development in a dose-dependent fashion (Ang et al., 2016). If this TF is deleted in embryonic cells, myocardial thinning is observed (Pu et al., 2004; Zeisberg et al., 2005). GATA4 is also active during postnatal heart development, not just in embryonic stages. It appears that many Gata4-dependent heart tissues cannot be rescued by Gata6 (Borok et al., 2015). There are three well characterized human GATA4 mutations identified by direct sequencing of different family members with Congenital Heart Defects (CHD) (Garg et al., 2003). The first is a glycine to serine substitution at position 296 (G296S). Secondly, a deletion in the nucleotide position 359 causes a frameshift mutation (E359del) predicted to result in a truncated GATA4 protein or a degradation of the GATA4 mRNA before translation. The third mutation is a deletion of the terminal end of chromosome 8p (8pter). GATA4 protein expression and activity are modulated through diverse mechanisms such as phosphorylation, acetylation and sumoylation (Suzuki, 2011). Besides its role in heart development, GATA4 regulates other processes like cell survival and proliferation (Nemer and Nemer, 2010; Suzuki, 2011).

2.4 Structure and Function of TBX5

TBX5 is part of the well-studied group of T-box genes. It binds to the consensus motif: 5'-AGGTGTGA-3'. Luna-Zurita et al. (2016) discovered a second significant TBX5 motif (5'-GAGGTG-3'). Mutated TBX5 is the main cause of the Holt-Oram syndrome, a disorder characterized by forelimb and cardiac abnormalities (Zhu et al., 2017). Bruneau et al. (2001) have shown that TBX5 binds multiple T-box binding elements (TBEs) in both the ANF and cx40 promoters. Similar to GATA4, TBX5's regulation during heart development does not have compensatory mechanisms such as feedback or genetic redundancy. Two of the TBX5 mutations that cause CHD include: nonsense mutations that produce a truncated protein and an aspartate to tyrosine substitution at codon 61 (D61Y) that severely affects the aorta and mitral valve (Al-Qattan et al., 2015). TBX5 is also correlated with atrial fibrillation and hypertension, indicating that it can function during postnatal states (Zhu et al., 2017). As expected, TBX5 usually interacts with other proteins to regulate transcription.

2.6 GATA4 and TBX5 Form a Cooperative Complex

GATA4 and TBX5 are known for participating in protein-protein interactions during cellular differentiation and regulation. Scientific data has revealed that GATA4 and TBX5 co-bind to gene targets and that their co-occupancy *in vivo* facilitates activation of cardiogenic transcription (Ang et al., 2016). Both TFs physically interact in co-immunoprecipitation assays and cooperatively activate the Atrial Natriuretic Factor (ANF), which encodes a peptide hormone secreted by the cardiac atria (Nemer and Nemer, 2010). Another identified transcriptional target of Gata4 and Tbx5 is Myh6 (myosin heavy chain), which helps generate the mechanical force for cardiac contraction (Ang et al., 2016; Dixon et al., 2011)). Additionally, Tbx5 promotes cardiomyocyte proliferation in cooperation with Gata4 by regulating Cdk4, a kinase important for cell proliferation (Misra et al., 2013).

Dixon et al. (2011) confirmed that GATA4 and TBX5 were the minimum elements needed to activate cardiac gene expression in human embryonic stem cells. Besides, it has been proven that the trimeric complex formed by GATA4, NKX2-5 and TBX5 is necessary to generate contracting cardiomyocytes (Luna-Zurita et al., 2016). Consistent with this fact, “mice doubly heterozygous for null alleles of Tbx5 and Nkx2-5 or Tbx5 and Gata4 have defects in heart formation that are more severe than those caused by each individual mutation” (Luna-Zurita et al., 2016). The results of this latter experiment also suggested the existence of a preferential motif distribution facilitating heterotypic TF interactions between GATA4, TBX5 and NKX2-5.

To support this, the glycine-to-serine missense mutation (G296S) on GATA4 affects its interaction with TBX5, disrupting their binding at enhancer elements associated with genes for heart development and muscle contraction (Ang et al., 2016; Garg et al., 2003). Ang’s results showed that when GATA4 was mutated, TBX5 was bound to mislocalized “lost sites” that resulted in aberrant activation of endothelial genes. This evidence supports Luna-Zurita’s claim that interdependent binding is essential for preventing ectopic gene expression. Equally important, haploinsufficiency of both Gata4 and Tbx5 can cause defects in atrioventricular septum formation and myocardial development. Related to their vital role, mice heterozygous for both Gata4 and Tbx5 mutations exhibited nearly 100% lethality by postnatal day 7 (Maitra et al., 2010). Most of the research cited here was conducted *in vivo*. Hence, we need detailed *in vitro* analyses of the DNA-binding properties governing the regulatory activity of the GATA4:TBX5 complex.

2.7 SELEX-seq to Determine the DNA-binding Sequences of Transcription Complexes

Classical experiments for studying novel protein-DNA interactions include *in vitro* and *in vivo* methods. The limitation of *in vivo* technologies such as chromatin immunoprecipitation (ChIP-seq), relies on the fact that under cellular conditions, the genome can be modified and shaped in a way that alters protein binding (Orenstein and Shamir, 2017). As a consequence, it is impossible to find all the protein's possible DNA binding sequences and its specificity. Therefore, we cannot fully decipher the grammar rules governing DNA specificity and protein binding with the available *in vivo* methods. In contrast, *in vitro* approaches can produce higher resolution data that facilitates our analysis. For example, using a synthetic DNA library excludes the limitations of genomic remodeling, cofactor presence and competing proteins. However, both individual approaches cannot provide all the necessary information. Ideally, *in vivo* and *in vitro* results should be compared to better predict protein-DNA interactions in a certain molecular context.

Among the available *in vitro* technologies to study the binding motifs of transcription factors, the most prominent are protein-binding microarrays (PBMs) and SELEX-based methods (Systematic Evolution of Ligands through Exponential Enrichment). PBMs consist of adding the protein of interest to a double-stranded DNA microarray and then measuring the amount of DNA-bound protein using a fluorescent antibody (Andrilenas et al., 2015). Many TF binding profiles have been deciphered using PBMs (Newburger and Bulyk, 2009). One of its limitations is that it only allows the identification of 8-10 bp long DNA sites. Therefore, longer sites may be missed during analysis. Consistently with this, the coverage of PBM models is very low for TFs that prefer longer than 10-mer DNA-motifs, families that bind to DNA as dimers, and heteromeric protein complexes (Jolma et al., 2010, 2013). In many cases, the reported PBM models describe partial specificity or half-sites. Since our research aim is to study a TF cooperative complex, we decided to use a SELEX-based method because it allows the analysis of multimeric binding sites spanning 20 bp or more. More importantly is that compared to PBMs, high-throughput SELEX can better predict *in vivo* binding (Orenstein and Shamir, 2014).

To test our hypothesis we used SELEX-seq, a novel approach developed by Slattery et al. (2011). It is a SELEX-based method combined with massively parallel sequencing. This *in vitro* selection experiment started with a 200 nM double-stranded DNA library. Each double strand contained 16 randomized base pairs flanked by sequences needed for PCR amplification and sequencing on a Illumina platform. They added the proteins to the library in a binding reaction and ran an EMSA gel (Electrophoretic Mobility Shift Assay). After the

reaction, they cut out, PCR-amplified, and purified the DNA bound by the cooperative complexes. These steps were repeated for three rounds to enhance sequence enrichment. Lastly, they sequenced all the samples and made PWMs (Position Weight Matrices) to analyze the proteins' specificity. We adapted their methods to incorporate a DNA library with double strands containing 20 randomized base pairs instead of 16. Although we used an *in vitro* approach to study our TF cooperative complex, we want to compare our data to existing *in vivo* ChIP-seq datasets. It is important to recognize that SELEX-seq is best for identifying enriched sequences. (Jolma et al., 2010; Kribelbauer et al., 2019).

3. MATERIALS AND METHODS

3.1 *TBX5* and *GATA4* DNA cloning

We bought DNA plasmids encoding the genes of human *GATA4* (Clone ID FHC23192) from Promega Corp. (Madison, WI) and *TBX5* (Clone ID HsCD00079979) from DNASU Plasmid Repository (Tempe, AZ). In addition, we bought the pEU-E01-GST-TEV-MCS-N1 vector from CellFree Sciences Co., Ltd. (Kanagawa, Japan). Then, we performed a Phusion High Fidelity PCR reaction to amplify and linearize the pEU plasmid using a forward primer [TTGTATAGAATTTACGGCTAGCGC] and a reverse primer [GCCCTGAAAATACAGGTTTTCG]. To confirm the vector length, we made a 1% agarose gel and purified it using the DNA Extraction Kit from Qiagen (Germantown, MD). Lastly, we measured the DNA concentration using the Nanodrop One Spectrophotometer from Thermo Fisher Scientific Inc. (Madison, WI). Different primers were used to amplify and create the overlapping ends between the amplicons of interest and the pEU plasmid (**Table 1**). The cloning step consisted of mixing: the NEB Impact Kit Gibson Assembly master mix from New England Biolabs, Inc. (Ipswich, MA), nuclease free water, the gene insert and the linearized pEU plasmid in a 4:1 ratio (insert:vector). The reaction was incubated at 50 °C for an hour, and the reaction products were transformed in DH-5 alpha *E. coli*. We conducted PCR colony screens to confirm the length of the inserts, did a MiniPrep DNA plasmid purification of confirmed plasmids using the kit from Qiagen (Germantown, MD), and measured the DNA concentration. The final step was to sequence the clones to verify that no mutations were incorporated. See **Table 2** to find the calculations for the Gibson Assembly.

Table 1: Primers used to amplify the target genes and pEU vector with the appropriate overlaps to conduct the Gibson Assembly Cloning

Name of the primer	**Primer sequence**	**T_m** (°C)	**GC content**
pEU14_GST_N_Fw	TTGTATAGAATTTACGGCTAGCGC	57	42%
pEU14_GST_N_Rv	GCCCTGAAAATACAGGTTTTCG	56	45%

pEU14_GA_TBXFL_Fw	CTGTATTTTCAG GGCATGGCCG ACGCAGAC	68	57%
pEU14_GA_TBXFL_Rv	CGTAAATTCTAT ACAACTACAAG CTATTGTCGC	60	36%
pEU14_GA_GT4FL_Fw	CTGTATTTTCAG GGCATGTATCA GAGCTTG	61	43%
pEU14_GA_GT4FL_Rv	CCGTAAATTCTA TACAACTACGC AGTGATTATG	59	36%

Table 2: Calculations made to do the Gibson Assembly cloning with a 4:1 ratio of insert to vector. The NEB Ligation calculator was used for this step.

Name	**Final Concentration (ng)**	**Volume used (µl)**
pEU vector	30	0.9
TBX5	47.2	0.6
GATA4	40.36	0.8
Positive Control (pEU plasmid)	----------	5
Negative Control	----------	4 H_2O + 1 of pEU

3.2 Protein Expression

3.2.1 Transcription Reaction

After the cloned plasmids were confirmed by Sanger sequencing, we used a wheat germ cell free protein expression system from the CellFree Sciences Co., Ltd. (Kanagawa, Japan). First, we ran a 20 µl transcription reaction for each clone (TBX5 and GATA4). We used the Green Fluorescent Protein clone as our control. The reaction contained nuclease free water, 5X transcription buffer LM, 25 mM NTP mix, 80 U/µl RNase inhibitor, 80U/ul SP6 RNA polymerase and 300 ng of DNA plasmid (**Tables 3 and 4**). These reactions were incubated at 37 °C for 6 hours.

Table 3: Calculations for the transcription reaction following the protocol of CellFree Sciences Co. The DNA plasmid volume was calculated to have a final concentration of 300ng.

DNA plasmid	**Concentration (ng/µl)**	**Volume (µl)**	**H_2O volume (µl)**
TBX5	219.0	1.37	11.88
GATA4	227.5	1.32	11.93
GFP	126.2	2.38	11.12

Table 4: Reagents used to make a master mix reaction for the transcription step following the protocol of CellFree Sciences Co. Each volume was multiplied by the number of samples (3).

Reagent	**Volume (µl)**	**Final Concentration**
Nuclease free water	Depends on sample	---------

5X transcription buffer LM	4	1X
NTP mix (25mM)	2	2.5 mM
RNase inhibitor (80U/ul)	0.25	1 U/µl
SP6 RNA polymerase (80U/ul)	0.25	1 U/µl
DNA Plasmid	Depends on sample	300 ng/µl
Total volume of reaction	20	

3.2.2 Translation Reaction

After, a fresh 1X SUB-AMIX SGC feeding buffer was prepared following the company's protocol. Approximately 2 ml of buffer was transferred per well into a 24 deep-well plate. The 50 µl translation reaction was performed using 20 mg/ml creatine kinase, WEPRO 7240 reagent, 1X SUB-AMIX SGC and the mRNA produced in the first step (**Table 5**). Mini dialysis cups (Thermo Scientific, Rockford, IL) with the translation reactions were placed in the well plate, covered with plastic foil and incubated at 15 °C for 24 hours.

Table 5 : Reagents used to make a master mix reaction for the translation step following the protocol of CellFree Sciences Co.

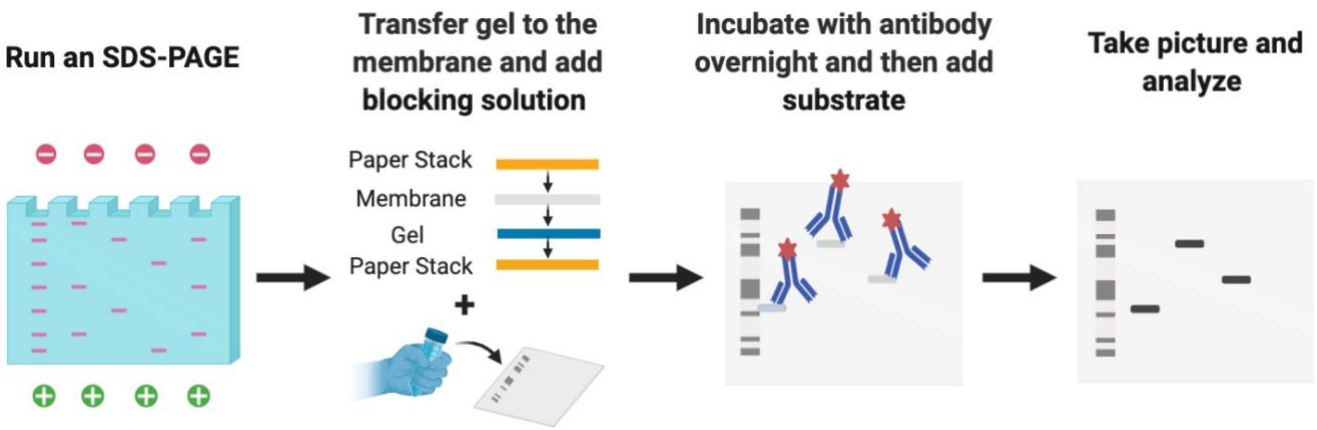
Reagent	**Volume (µl)**	**Final Concentration**
mRNA	16.7	1/3 vol.
1X SUB-AMIX SGC	24	----------
WEPRO 7240/7240H/7240G (240 OD)	8.3	40 OD

Creatine kinase (20mg/ml)	1	40 ug/ml
Total	50	

3.2.3 SDS-PAGE and Western Blot

After the protein extract was collected, we did a SDS-PAGE followed by Western Blot to confirm our proteins were effectively produced and corresponded to the expected molecular weight. We mixed 5ul of protein with 2 µl of Beta-mercaptoethanol loading dye, heated the samples at 95 °C for 5 minutes and then loaded them into a pre-made 12% polyacrylamide gel from Bio-Rad Laboratories, Inc. (Hercules, CA). After the gel ran for 1.5 hours at 120V, the PVDF western blot membrane was submerged in methanol for 60 seconds and 2 paper stacks were submerged in the transfer buffer (50 ml of 5X transfer buffer, 50 ml of ethanol and 150 ml of purified water) for 60 seconds. The transfer sandwich (paper stack, membrane, gel and paper stack) was prepared and placed in the trans-blot turbo transfer system machine from Bio-Rad Laboratories, Inc. (Hercules, CA) for 7 minutes. The membrane was then blocked with 0.25 g of milk resuspended in 10 ml of 1X TBST buffer (450 ml of purified water, 50 ml of 10X TBS, 500 µl of 20% Tween and pH 7.4) for 1 hour. After blocking time, the membrane was incubated overnight at 4°C with an 1:10,000 dilution of the anti-GST HRP-conjugated antibody. The membrane was imaged using the Azure Sapphire Biomolecular Imager from Azure Biosystems (Dublin, CA). See **Figure 1** for the overview of a western blot experiment.

Figure 1: Overview of how a western blot is made.

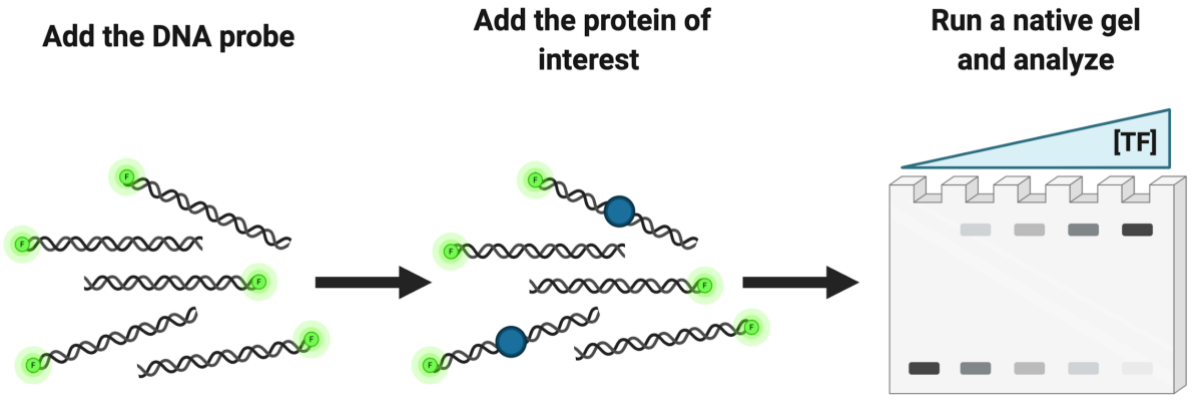


3.3 Electrophoretic Mobility Shift Assay

We used electrophoretic mobility shift assay (EMSA) to test the DNA-binding activity of the TFs produced via the wheat germ cell-free protein expression system. This experiment depends on fluorescence measures and native gel electrophoresis to assess the properties already mentioned, binding and specificity. We designed a DNA sequence made of 60 nucleotides, containing 20 bases of the Atrial Natriuretic Factor (ANF) gene promoter [GTAATATCACACCTGTACAT] flanked by 20 constant bases on the 5' end [CTCGCCTGGGCAGAAGTGTC] and 20 constant bases on the 3' end [GACACTTCTGCCCAGGCGAG]. Then, the 5' end of the probe was modified with the dye IR700 [CTCGCCTGGGCAGAAGTGTC] from Integrated DNA Technologies, Inc. (Coralville, IA). The double-stranded DNA probe was synthesized with an extension reaction containing: 2 µl of the ANF sequence, 6 µl of 50.9 µM IR700, 25 µl of EconoTaq Master Mix 2X from Lucigen Corp. (Middleton, WI), and 17 µl of nuclease free water. The protocol for the thermocycler was: 95 °C for 2 min., 55 °C for 1 min., 72 °C for 5min., 10 °C on hold.

We mixed 15 ul binding reactions containing 10 nM of the labeled DNA probe in 1.3X Trevor binding buffer (50 mM Tris pH 7.5 and 250 mM NaCl), 33.3 ng pdl-dC, 33.3 ng BSA, .07% Tween-20, 13 mM DTT and nuclease free water. To make sure we chose a protein concentration that would allow us to see a significant DNA-binding band that could work for the SELEX-seq experiments, we tested 3 different dilutions of both GATA4 and TBX5. For each separate binding reaction we added 5 ul of one of three dilutions: 1/25, 1/5 or 1/1. The samples were incubated at 30 °C for 30 min. and then at room temperature for 30 min. Native 5% polyacrylamide gels were pre-ran at 74V for 15 min. We loaded 15 µl of each sample at 61V and the gel ran at 121V for 2.5 hours. Once finished, the gels were imaged using the Azure Sapphire Biomolecular Imager (Azure Biosystems, Dublin, CA).

Figure 2: Fundamentals of an electrophoretic mobility shift assay.

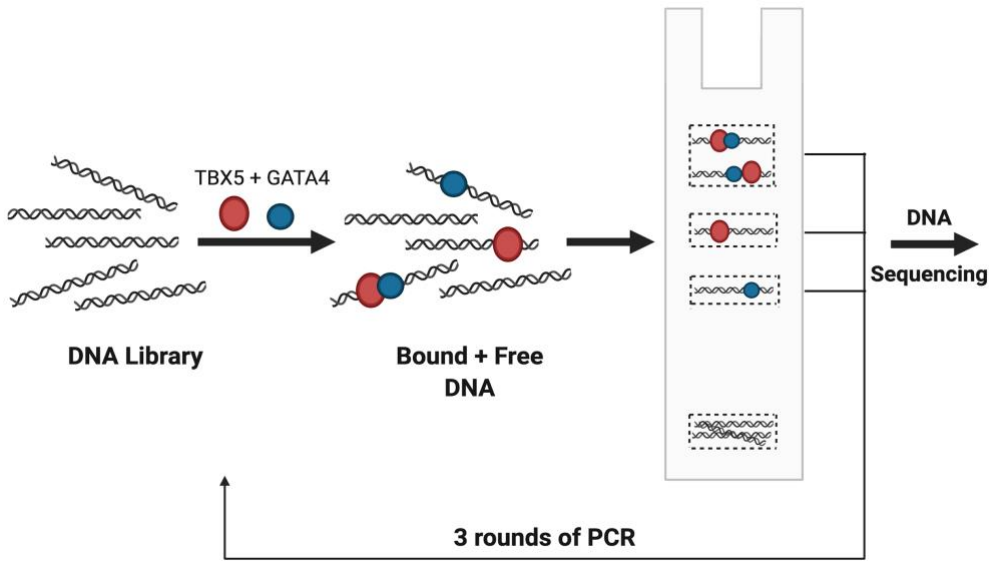


3.4 Systematic Evolution of Ligands by Exponential Enrichment (SELEX-seq)

The DNA-binding sites of GATA4, TBX5 and the GATA4:TBX5 complex were determined by SELEX-seq. We used a 200 nM biotinylated DNA library from Integrated DNA Technologies, Inc. (Coralville, IA) with a central randomized 20bp sequence flanked by 20 constant nucleotides on each side. The binding reactions contained 10nM of the labeled DNA (ANF probe or the DNA library) and a 1/5 protein dilution. The reaction master mixes also contained: 1X Trevor binding buffer (50 mM Tris pH 7.5 and 250 mM NaCl), 25 ng pdI-dC, 25 ng BSA, .05% Tween-20 and 10 mM DTT. The samples were incubated at 30 °C for 30 min. and then at room temperature for 30 min. Native 5% polyacrylamide gels were pre-ran at 74V for 15 min. We loaded 20 µl of each sample at 61V and the gel ran at 121V for 2.5 hours. The gel was imaged using the Azure Sapphire Imager.

The bands corresponding to the protein:DNA complexes were cut, and the DNA eluded in 500 µl of the elution buffer from Qiagen (Germantown, MD) overnight at 30°C and thermoshaking at 1,200 RPM (Thermoshaker BIOGRANT). Bound DNA was enriched with Dynabeads M280 Streptavidin magnetic beads using the protocol from Invitrogen (Carlsbad, CA). After washing the beads with the elution buffer 3 times, the pulled-down DNA was resuspended in a PCR master mix with EconoTaq and amplified for 20 cycles. We purified the DNA samples using the Qiagen kit, measured their concentration with the Nanodrop One Spectrophotometer, and used them for subsequent SELEX rounds. After the three rounds of selection were performed, we incorporated unique 6-bp barcodes and Illumina sequencing adapters to the DNA sequences of all the rounds, including the original library.

Figure 3: Overview of the binding specificity determination using SELEX-seq



3.5 SELEX-seq Data Analysis

Raw sequencing data were binned in a table according to the barcoding number for each sample. The scripts used for the Autoseed analysis were recovered from the Supplementary File 3 published by Nitta et al. (2015). The Readme.txt document in the supplemental file contained all the instructions to run the Autoseed program and obtain different files with the following information: logos, heatmaps and the matrices to create the PWMs. The following command is an example of the ones used to generate all the files we needed (Letters in bold change depending on the name of the file):

./totalautoseed -20N **R44_888**.txt **R44_042**.txt 1 8 10 0.35 - 50 40 > **GATA4TBX5R3_042_R44_888**.txt; cp Kmer_summary8to10.svg **GATA4TBX5R3_042_R44_888**.svg

We used this particular command only to generate the files for Round 3 of SELEX-seq because we expected those sequences to be more enriched than previous rounds. Then, we manually curated the data to choose 5-7 *seeds* of each sample (GATA4, TBX5 and GATA4:TBX5 complex). *Seed* is the term describing the enriched bound sites identified by Autoseed. In the future we will use these seeds to create Position Weight Matrices (PWMs) and other analyses to interpret our data in depth.

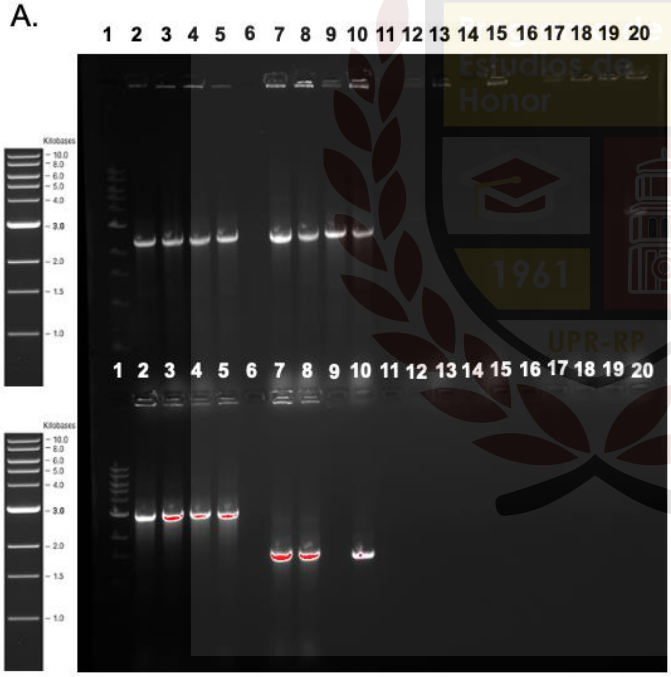
4. RESULTS

4.1 DNA Cloning

After the Gibson Assembly reaction products were transformed in DH-5 alpha *E. coli*, we conducted PCR colony screens to confirm the presence of the correct inserts in the plasmid. We ran a 1% agarose gel electrophoresis to compare the insert length with the expected ones. In **Figure 4.a**, lanes 2 through 5 of the second row showed the PCR products in the expected length of TBX5 (2,984bp). Although lanes 12-20 contained GATA4, only lane 20 exhibited the correct amplification of GATA4 (2,753bp). Consequently, we ran another colony screen for GATA4 and confirmed the length.

Figure 4: Colony screens of the inserts (GATA5, TBX5 and GFP) to confirm their lengths

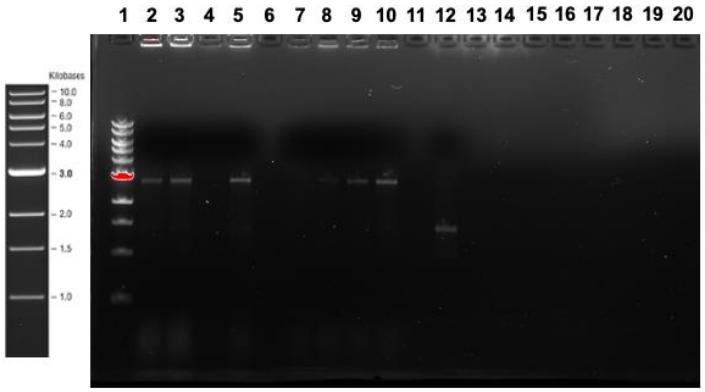
A.



B.

Lane	Sample	Size (bp)
2-5	**NKX2-5** Clone 1	2402
7-10	**NKX2-5** Clone 2	2402
12-15	**GATA4** Clone 1	2753
17-20	**GATA4** Clone 2	2753
2-5	**TBX5** Clone 1	2984
7-8	**GFP N1**	1500
10	**pEU-GST**	1463

C.



D.

Lane	Sample	Size (bp)
2-5	**GATA4** Clone	2753
7-10	**GATA4** Clone 4	2753
12	**pEU-GST**	1463

After the colony screens, we did a MiniPrep DNA plasmid purification of the pEU plasmids with the inserts, to confirm their lengths and measure the DNA concentration. The DNA concentrations had to be sufficient for the transcription step of the wheat germ cell free protein synthesis system. The concentrations were: TBX5 219.0 ng/µl, GATA4 227.5 ng/µl and GFP control 126.2 ng/µl (**Table 6**). Afterwards, these purified plasmids were verified by Sanger sequencing.

Table 6: DNA plasmid concentration of the clones obtained with the Miniprep that were chosen to conduct the subsequent protein expression experiments.

Clone	**Concentration (ng/µl)**
TBX5 Clone 1	219.0
GATA4 Clone 2	227.5
GFP Clone 2	126.2

4.2 Sanger Sequencing

GATA4 and TBX5 plasmids were verified by Sanger sequencing using SP6 forward and MCS reverse primers. We did an alignment of our experimental clones with the theoretical DNA sequences of both transcription factors using the BLAST tool from NCBI (**Figures 5 and 6**).

Figure 5: BLASTn Sequence Alignment of the plasmid containing **GATA4**

```
Query  661   TACGACATCCCAACGACCGAAAACCTGTATTTTCAGGGCATGTATCAGAGCTTGGCCATG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   TACGACATCCCAACGACCGAAAACCTGTATTTTCAGGGCATGTATCAGAGCTTGGCCATG

Query  721   GCCGCCAACCACGGGCCGCCCCCCGGTGCCTACGAGGCGGGCGGCCCCGGCGCCTTCATG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   GCCGCCAACCACGGGCCGCCCCCCGGTGCCTACGAGGCGGGCGGCCCCGGCGCCTTCATG

Query  781   CACGGCGCGGGCGCCGCGTCCTCGCCAGTCTACGTGCCCACACCGCGGGTGCCCTCCTCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   CACGGCGCGGGCGCCGCGTCCTCGCCAGTCTACGTGCCCACACCGCGGGTGCCCTCCTCC

Query  841   GTGCTGGGCCTGTCCTACCTCCAGGGCGGAGGCGCGGGCTCTGCGTCCGGAGGCGCCTCG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   GTGCTGGGCCTGTCCTACCTCCAGGGCGGAGGCGCGGGCTCTGCGTCCGGAGGCGCCTCG

Query  901   GGCGGCAGCTCCGGTGGGGCCGCGTCTGGTGCGGGGCCCGGGACCCAGCAGGGCAGCCCG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   GGCGGCAGCTCCGGTGGGGCCGCGTCTGGTGCGGGGCCCGGGACCCAGCAGGGCAGCCCG

Query  961   GGATGGAGCCAGGCGGGAGCCGACGGAGCCGCTTACACCCCGCCGCCGGTGTCGCCGCGC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GGATGGAGCCAGGCGGGAGCCGACGGAGCCGCTTACACCCCGCCGCCGGTGTCGCCGCGC

Query  1021  TTCTCCTTCCCGGGGACCACCGGGTCCCTGGCGGCCGCCGCCGCCGCTGCCGCGGCCCGG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  TTCTCCTTCCCGGGGACCACCGGGTCCCTGGCGGCCGCCGCCGCCGCTGCCGCGGCCCGG

Query  1081  GAAGCTGCGGCCTACAGCAGTggcggcggagcggcgggtgcgggcctggcgggccgcgag
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  GAAGCTGCGGCCTACAGCAGTGGCGGCGGAGCGGCGGGTGCGGGCCTGGCGGGCCGCGAG

Query  1141  cagtacgggcgcgccggcttcgcgggCTCCTACTCCAGCCCCTACCCGGCTTACATGGCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  CAGTACGGGCGCGCCGGCTTCGCGGGCTCCTACTCCAGCCCCTACCCGGCTTACATGGCC

Query  1201  GACGTGGGCGCGTCCTGGGCCGCAGCCGCCGCCGCCTCCGCCGGCCCCTTCGACAGCCCG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  GACGTGGGCGCGTCCTGGGCCGCAGCCGCCGCCGCCTCCGCCGGCCCCTTCGACAGCCCG

Query  1261  GTCCTGCACAGCCTGCCCGGCCGGGCCAACCCGGCCGCCCGACACCCCAATCTCGATATG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GTCCTGCACAGCCTGCCCGGCCGGGCCAACCCGGCCGCCCGACACCCCAATCTCGATATG

Query  1321  TTTGACGACTTCTCAGAAGGCAGAGAGTGTGTCAACTGTGGGGCTATGTCCACCCCGCTC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  TTTGACGACTTCTCAGAAGGCAGAGAGTGTGTCAACTGTGGGGCTATGTCCACCCCGCTC

Query  1381  TGGAGGCGAGATGGGACGGGTCACTATCTGTGCAACGCCTGCGGCCTCTACCACAAGATG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  TGGAGGCGAGATGGGACGGGTCACTATCTGTGCAACGCCTGCGGCCTCTACCACAAGATG

Query  1441  AACGGCATCAACCGGCCGCTCATCAAGCCTCAGCGCCGGCTGTCCGCCTCCCGCCGAGTG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  AACGGCATCAACCGGCCGCTCATCAAGCCTCAGCGCCGGCTGTCCGCCTCCCGCCGAGTG

Query  1501  GGCCTCTCCTGTGCCAACTGCCAGACCACCACCACCACGCTGTGGCGCCGCAATGCGGAG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  GGCCTCTCCTGTGCCAACTGCCAGACCACCACCACCACGCTGTGGCGCCGCAATGCGGAG

Query  1561  GGCGAGCCTGTGTGCAATGCCTGCGGCCTCTACATGAAGCTCCACGGGGTCCCCAGGCCT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  GGCGAGCCTGTGTGCAATGCCTGCGGCCTCTACATGAAGCTCCACGGGGTCCCCAGGCCT

Query  1621  CTTGCAATGCGGAAAGAGGGGATCCAAACCAGAAAACGGAAGCCCAAGAACCTGAATAAA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  CTTGCAATGCGGAAAGAGGGGATCCAAACCAGAAAACGGAAGCCCAAGAACCTGAATAAA

Query  1681  TCTAAGACACCAGCAGCTCCTTCAGGCAGTGAGAGCCTTCCTCCCGCCAGCGGTGCTTCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1681  TCTAAGACACCAGCAGCTCCTTCAGGCAGTGAGAGCCTTCCTCCCGCCAGCGGTGCTTCC

Query  1741  AGCAACTCCAGCAACGCCACCACCAGCAGCAGCGAGGAGATGCGTCCCATCAAGACGGAG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1741  AGCAACTCCAGCAACGCCACCACCAGCAGCAGCGAGGAGATGCGTCCCATCAAGACGGAG

Query  1801  CCTGGCCTGTCATCTCACTACGGGCACAGCAGCTCCGTGTCCCAGACGTTCTCAGTCAGT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1801  CCTGGCCTGTCATCTCACTACGGGCACAGCAGCTCCGTGTCCCAGACGTTCTCAGTCAGT

Query  1861  GCGATGTCTGGCCATGGGCCCTCCATCCACCCTGTCCTCTCGGCCCTGAAGCTCTCCCCA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1861  GCGATGTCTGGCCATGGGCCCTCCATCCACCCTGTCCTCTCGGCCCTGAAGCTCTCCCCA

Query  1921  CAAGGCTATGCGTCTCCCGTCAGCCAGTCTCCACAGACCAGCTCCAAGCAGGACTCTTGG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1921  CAAGGCTATGCGTCTCCCGTCAGCCAGTCTCCACAGACCAGCTCCAAGCAGGACTCTTGG

Query  1981  AACAGCCTGGTCTTGGCCGGACAGTCACGGGGACATAATCACTGCGTAG  2029
              |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1981  AACAGCCTGGTCTTGGCCGGACAGTCACGGGGACATAATCACTGCGTAG  2029
```

Figure 6: BLASTn Sequence Alignment of the plasmid containing **TBX5**

Score	Expect	Identities	Gaps	Strand
4172 bits(2259)	0.0	2259/2259(100%)	0/2259(0%)	Plus/Plus

```
Query  1     ATGGAATCCCCTATACTAGGTTATTGGAAAATTAAGGGCCTTGTGCAACCCACTCGACTT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGGAATCCCCTATACTAGGTTATTGGAAAATTAAGGGCCTTGTGCAACCCACTCGACTT

Query  61    CTTTTGGAATATCTTGAAGAAAAATATGAAGAGCATTTGTATGAGCGCGATGAAGGTGAT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CTTTTGGAATATCTTGAAGAAAAATATGAAGAGCATTTGTATGAGCGCGATGAAGGTGAT

Query  121   AAATGGCGAAACAAAAAGTTTGAATTGGGTTTGGAGTTTCCCAATCTTCCTTATTATATT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   AAATGGCGAAACAAAAAGTTTGAATTGGGTTTGGAGTTTCCCAATCTTCCTTATTATATT

Query  181   GATGGTGATGTTAAATTAACACAGTCTATGGCCATCATACGTTATATAGCTGACAAGCAC
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   GATGGTGATGTTAAATTAACACAGTCTATGGCCATCATACGTTATATAGCTGACAAGCAC

Query  241   AACATGTTGGGTGGTTGTCCAAAAGAGCGTGCAGAGATTTCAATGCTTGAAGGAGCGGTT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   AACATGTTGGGTGGTTGTCCAAAAGAGCGTGCAGAGATTTCAATGCTTGAAGGAGCGGTT

Query  301   TTGGATATTAGATACGGTGTTTCGAGAATTGCATATAGTAAAGACTTTGAAACTCTCAAA
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   TTGGATATTAGATACGGTGTTTCGAGAATTGCATATAGTAAAGACTTTGAAACTCTCAAA

Query  361   GTTGATTTTCTTAGCAAGCTACCTGAAATGCTGAAAATGTTCGAAGATCGTTTATGTCAT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   GTTGATTTTCTTAGCAAGCTACCTGAAATGCTGAAAATGTTCGAAGATCGTTTATGTCAT

Query  421   AAAACATATTTAAATGGTGATCATGTAACCCATCCTGACTTCATGTTGTATGACGCTCTT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   AAAACATATTTAAATGGTGATCATGTAACCCATCCTGACTTCATGTTGTATGACGCTCTT

Query  481   GATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   GATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTT

Query  541   AAAAAACGTATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATA
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   AAAAAACGTATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATA

Query  601   GCATGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAAGAT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   GCATGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAAGAT

Query  661   TACGACATCCCAACGACCGAAAACCTGTATTTTCAGGGCATGGCCGACGCAGACGAGGGC
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   TACGACATCCCAACGACCGAAAACCTGTATTTTCAGGGCATGGCCGACGCAGACGAGGGC

Query  721   TTTGGCCTGGCGCACACGCCTCTGGAGCCTGACGCAAAAGACCTGCCCTGCGATTCGAAA
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   TTTGGCCTGGCGCACACGCCTCTGGAGCCTGACGCAAAAGACCTGCCCTGCGATTCGAAA

Query  781   CCCGAGAGCGCGCTCGGGGCCCCCAGCAAGTCCCCGTCGTCCCCGCAGGCCGCCTTCACC
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   CCCGAGAGCGCGCTCGGGGCCCCCAGCAAGTCCCCGTCGTCCCCGCAGGCCGCCTTCACC

Query  841   CAGCAGGGCATGGAGGGAATCAAAGTGTTTCTCCATGAAAGAGAACTGTGGCTAAAATTC
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   CAGCAGGGCATGGAGGGAATCAAAGTGTTTCTCCATGAAAGAGAACTGTGGCTAAAATTC

Query  901   CACGAAGTGGGCACGGAAATGATCATAACCAAGGCTGGAAGGCGGATGTTTCCCAGTTAC
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   CACGAAGTGGGCACGGAAATGATCATAACCAAGGCTGGAAGGCGGATGTTTCCCAGTTAC

Query  961   AAAGTGAAGGTGACGGGCCTTAATCCCAAAACGAAGTACATTCTTCTCATGGACATTGTA
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   AAAGTGAAGGTGACGGGCCTTAATCCCAAAACGAAGTACATTCTTCTCATGGACATTGTA

Query  1021  CCTGCCGACGATCACAGATACAAATTCGCAGATAATAAATGGTCTGTGACGGGCAAAGCT
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  CCTGCCGACGATCACAGATACAAATTCGCAGATAATAAATGGTCTGTGACGGGCAAAGCT
```

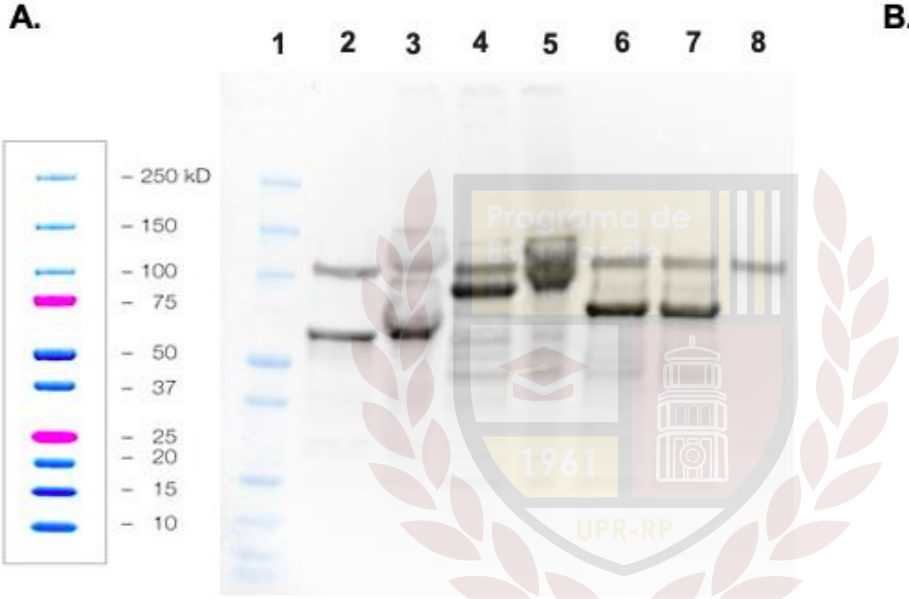
```
Query  1081  GAGCCCGCCATGCCTGGCCGCCTGTACGTGCACCCAGACTCCCCCGCCACCGGGGCGCAT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  GAGCCCGCCATGCCTGGCCGCCTGTACGTGCACCCAGACTCCCCCGCCACCGGGGCGCAT

Query  1141  TGGATGAGGCAGCTCGTCTCCTTCCAGAAACTCAAGCTCACCAACAACCACCTGGACCCA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  TGGATGAGGCAGCTCGTCTCCTTCCAGAAACTCAAGCTCACCAACAACCACCTGGACCCA

Query  1201  TTTGGGCATATTATTCTAAATTCCATGCACAAATACCAGCCTAGATTACACATCGTGAAA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  TTTGGGCATATTATTCTAAATTCCATGCACAAATACCAGCCTAGATTACACATCGTGAAA

Query  1261  GCGGATGAAAATAATGGATTTGGCTCAAAAAATACAGCGTTCTGCACTCACGTCTTTCCT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GCGGATGAAAATAATGGATTTGGCTCAAAAAATACAGCGTTCTGCACTCACGTCTTTCCT

Query  1321  GAGACTGCGTTTATAGCAGTGACTTCCTACCAGAACCACAAGATCACGCAATTAAAGATT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  GAGACTGCGTTTATAGCAGTGACTTCCTACCAGAACCACAAGATCACGCAATTAAAGATT

Query  1381  GAGAATAATCCCTTTGCCAAAGGATTTCGGGGCAGTGATGACATGGAGCTGCACAGAATG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  GAGAATAATCCCTTTGCCAAAGGATTTCGGGGCAGTGATGACATGGAGCTGCACAGAATG

Query  1441  TCAAGAATGCAAAGTAAAGAATATCCCGTGGTCCCCAGGAGCACCGTGAGGCAAAAAGTG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  TCAAGAATGCAAAGTAAAGAATATCCCGTGGTCCCCAGGAGCACCGTGAGGCAAAAAGTG

Query  1501  GCCTCCAACCACAGTCCTTTCAGCAGCGAGTCTCGAGCTCTCTCCACCTCATCCAATTTG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  GCCTCCAACCACAGTCCTTTCAGCAGCGAGTCTCGAGCTCTCTCCACCTCATCCAATTTG

Query  1561  GGGTCCCAATACCAGTGTGAGAATGGTGTTTCCGGCCCCTCCCAGGACCTCCTGCCTCCA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  GGGTCCCAATACCAGTGTGAGAATGGTGTTTCCGGCCCCTCCCAGGACCTCCTGCCTCCA

Query  1621  CCCAACCCATACCCACTGCCCCAGGAGCATAGCCAAATTTACCATTGTACCAAGAGGAAA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  CCCAACCCATACCCACTGCCCCAGGAGCATAGCCAAATTTACCATTGTACCAAGAGGAAA

Query  1681  GAGGAAGAATGTTCCACCACAGACCATCCCTATAAGAAGCCCTACATGGAGACATCACCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1681  GAGGAAGAATGTTCCACCACAGACCATCCCTATAAGAAGCCCTACATGGAGACATCACCC

Query  1741  AGTGAAGAAGATTCCTTCTACCGCTCTAGCTATCCACAGCAGCAGGGCCTGGGTGCCTCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1741  AGTGAAGAAGATTCCTTCTACCGCTCTAGCTATCCACAGCAGCAGGGCCTGGGTGCCTCC

Query  1801  TACAGGACAGAGTCGGCACAGCGGCAAGCTTGCATGTATGCCAGCTCTGCGCCCCCCAGC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1801  TACAGGACAGAGTCGGCACAGCGGCAAGCTTGCATGTATGCCAGCTCTGCGCCCCCCAGC

Query  1861  GAGCCTGTGCCCAGCCTAGAGGACATCAGCTGCAACACGTGGCCAAGCATGCCTTCCTAC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1861  GAGCCTGTGCCCAGCCTAGAGGACATCAGCTGCAACACGTGGCCAAGCATGCCTTCCTAC

Query  1921  AGCAGCTGCACCGTCACCACCGTGCAGCCCATGGACAGGCTACCCTACCAGCACTTCTCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1921  AGCAGCTGCACCGTCACCACCGTGCAGCCCATGGACAGGCTACCCTACCAGCACTTCTCC

Query  1981  GCTCACTTCACCTCGGGGCCCCTGGTCCCTCGGCTGGCTGGCATGGCCAACCATGGCTCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1981  GCTCACTTCACCTCGGGGCCCCTGGTCCCTCGGCTGGCTGGCATGGCCAACCATGGCTCC

Query  2041  CCACAGCTGGGAGAGGGAATGTTCCAGCACCAGACCTCCGTGGCCCACCAGCCTGTGGTC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2041  CCACAGCTGGGAGAGGGAATGTTCCAGCACCAGACCTCCGTGGCCCACCAGCCTGTGGTC

Query  2101  AGGCAGTGTGGGCCTCAGACTGGCCTGCAGTCCCCTGGCACCCTTCAGCCCCCTGAGTTC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2101  AGGCAGTGTGGGCCTCAGACTGGCCTGCAGTCCCCTGGCACCCTTCAGCCCCCTGAGTTC

Query  2161  CTCTACTCTCATGGCGTGCCAAGGACTCTATCCCCTCATCAGTACCACTCTGTGCACGGA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2161  CTCTACTCTCATGGCGTGCCAAGGACTCTATCCCCTCATCAGTACCACTCTGTGCACGGA

Query  2221  GTTGGCATGGTGCCAGAGTGGAGCGACAATAGCTTGTAG  2259
              |||||||||||||||||||||||||||||||||||||||
Sbjct  2221  GTTGGCATGGTGCCAGAGTGGAGCGACAATAGCTTGTAG  2259
```

4.3 Protein Expression

After using the wheat germ cell free protein synthesis system, the western blot showed that GATA4 and TBX5 were successfully expressed. Both proteins can be found at their expected molecular weight: GATA4 at 72kDa and TBX5 at 85kDa. **Figure 7** shows the results.

Figure 7: Western blot membrane after incubating it with anti-GST HRP-conjugated antibody overnight.

A.



B.

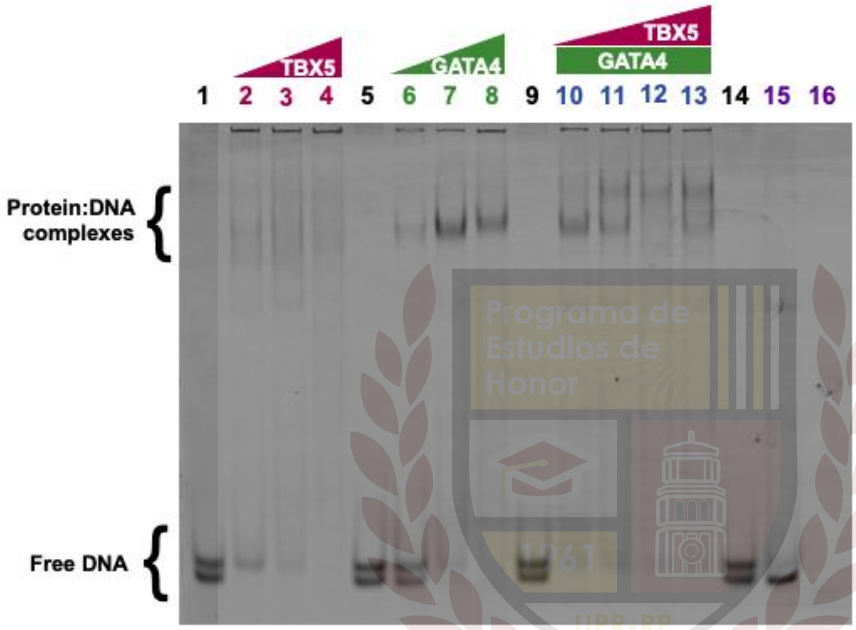
Lane	Protein	MW (kDa)
1	Ladder	-------
2, 3	NKX2.5	62
4,5	TBX5	85
6,7	GATA4	72
8	GFP Control	28

4.4 DNA-binding Validation With EMSA

To validate the DNA-binding activity of GATA4 and TBX5, we did an EMSA using the ANF gene as the DNA probe. We chose ANF because previous research has demonstrated that GATA4 and TBX5 regulate this gene. **Figure 8** suggests that GATA4 and TBX5 bind to this specific DNA sequence, both as monomers and in complex. As we increased the protein amount, the amount of free DNA decreased. The gel shift represents the migration of the TF:DNA complex. Lanes 2-4 show TBX5 bound to ANF while lanes 6-8 show GATA4 bound to ANF. Likewise, in lanes 10-13, the first band (from top to bottom) shows the GATA4:TBX5:DNA complex, and the second one represents GATA bound to DNA. The relative weight of the GATA4 shift suggested that this TF was bound to DNA as a homodimer. We confirmed this with the SELEX-seq data. This EMSA also helped us choose the protein dilutions and combinations that might have the best

resolution for the SELEX-seq. The first protein concentration we chose for the heteromeric complex was: 1/10 (1 in 10µl) final dilution of both GATA4 and TBX5. The second concentration was 1/5 (1 in 5µl) final dilution of both GATA and TBX5.

Figure 8: The EMSA native gel shows the free DNA at the bottom and the different protein:DNA complexes at the top. The (*) symbol in the table represents the dilutions that were chosen for the subsequent rounds of SELEX-seq.



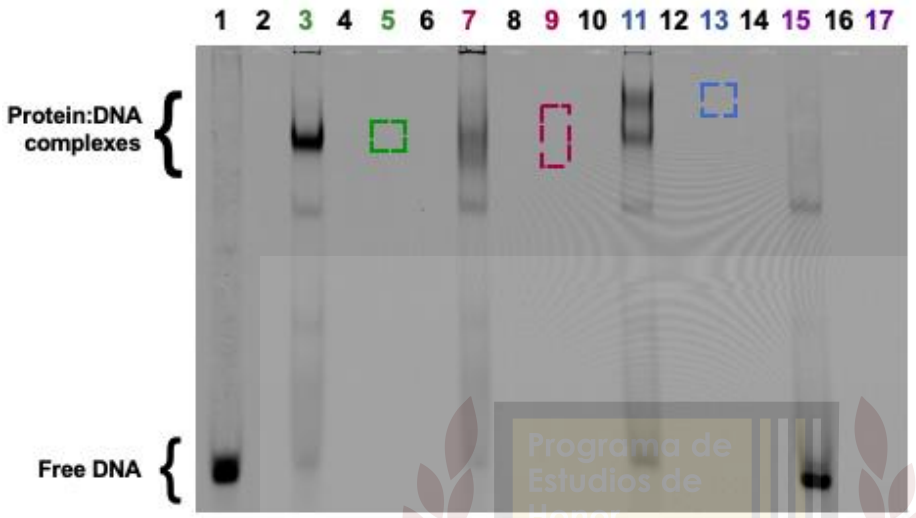
Lane	Dilution
1	Binding buffer
2	1/25 TBX5
3	1/5 TBX5
4	1/1 TBX5
5	Binding buffer
6	1/25 GATA4
7	1/5 GATA4
8	1/1 GATA4
9	Binding buffer
10	1/5 GATA4 1/25 TBX5
11	*1/5 GATA4 1/5 TBX5
12	1/5 GATA4 1/1 TBX5
13	*GATA4 TBX5 equimolar
14	Binding buffer
15	1/5 GFP
16	1/1 GFP

4.5 SELEX-seq Rounds

Since we chose two different protein dilutions for the GATA4:TBX5 complex, we made a duplicate for each round of SELEX-seq. The transcription factor heteromeric complex of **gel 1** contained the 1/10 final protein dilution. On the other hand, the TF complex of **gel 2** contained a 1/5 final protein dilution. In this experiment we ran EMSA gels with different binding reactions. For each protein (GATA4, TBX5 and GATA5:TBX5 Complex), we ran one binding reaction containing the ANF gene and a second reaction with the DNA library. The bands in the lanes with the ANF probe served as guides for us to cut the same position in the lanes with the DNA library, which were transparent. Lane 11 of each gel in every round exhibited 2 bands which appear to be the GATA4:TBX5 complex and the GATA4 homodimer, respectively. We confirmed this prediction using the results from the SELEX-seq.

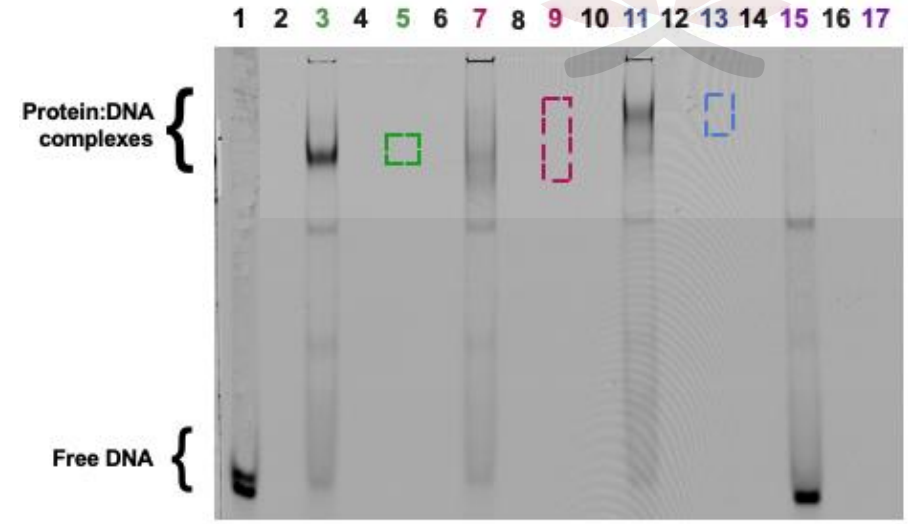
Round 1

Figure 9: Gel 1 of round 1.



Lane	Dilution
1	Binding buffer
2	Empty
3	1/5 GATA4 - ANF
4	Empty
5	1/5 GATA4 - Library
6	Empty
7	1/5 TBX5 - ANF
8	Empty
9	1/5 TBX5 - Library
10	Empty
11	1/10 GATA4 & TBX5 -ANF
12	Empty
13	1/10 GATA4 & TBX5 -Lib.
14	Empty
15	1/5 GFP - ANF
16	Empty
17	1/5 GFP - Library

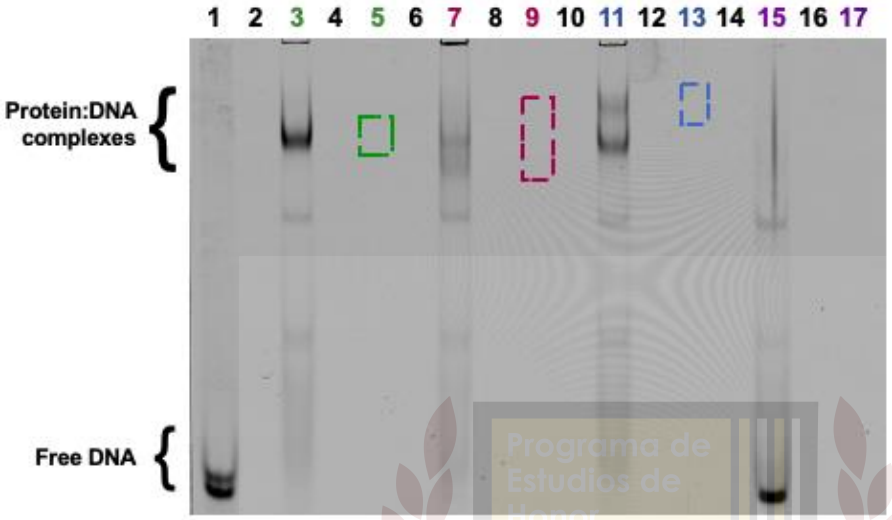
Figure 10: Gel 2 of round 1.



Lane	Dilution
1	Binding buffer
2	Empty
3	1/5 GATA4 - ANF
4	Empty
5	1/5 GATA4 - Library
6	Empty
7	1/5 TBX5 - ANF
8	Empty
9	1/5 TBX5 - Library
10	Empty
11	1/5 GATA4 & TBX5 -ANF
12	Empty
13	1/5 GATA4 & TBX5 -Lib.
14	Empty
15	1/5 GFP - ANF
16	Empty
17	1/5 GFP - Library

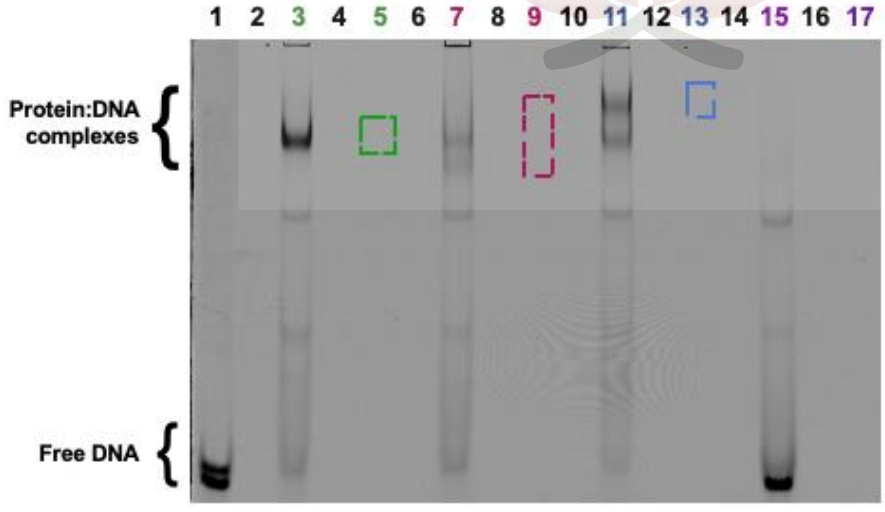
Round 2

Figure 11: Gel 1 of round 2.



Lane	Dilution
1	Binding buffer
2	Empty
3	1/5 GATA4 - ANF
4	Empty
5	1/5 GATA4 - Library
6	Empty
7	1/5 TBX5 - ANF
8	Empty
9	1/5 TBX5 - Library
10	Empty
11	1/10 GATA4 & TBX5 -ANF
12	Empty
13	1/10 GATA4 & TBX5 -Lib.
14	Empty
15	1/5 GFP - ANF
16	Empty
17	1/5 GFP - Library

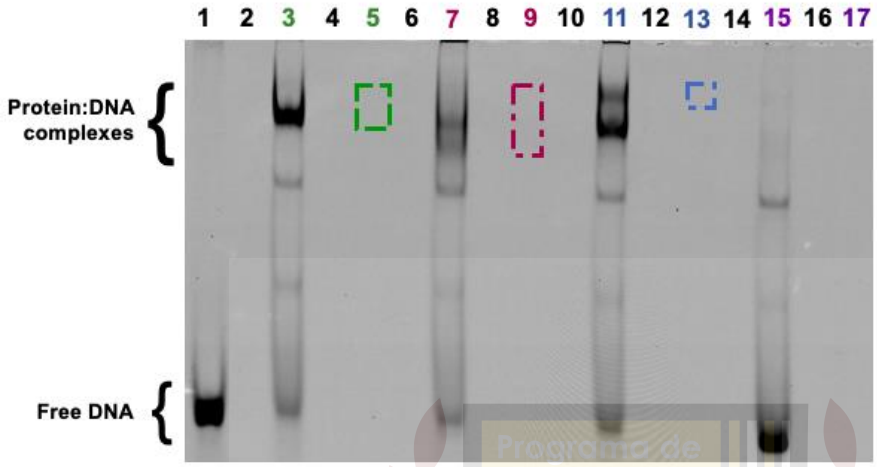
Figure 12: Gel 2 of round 2.



Lane	Dilution
1	Binding buffer
2	Empty
3	1/5 GATA4 - ANF
4	Empty
5	1/5 GATA4 - Library
6	Empty
7	1/5 TBX5 - ANF
8	Empty
9	1/5 TBX5 - Library
10	Empty
11	1/5 GATA4 & TBX5 -ANF
12	Empty
13	1/5 GATA4 & TBX5 -Lib.
14	Empty
15	1/5 GFP - ANF
16	Empty
17	1/5 GFP - Library

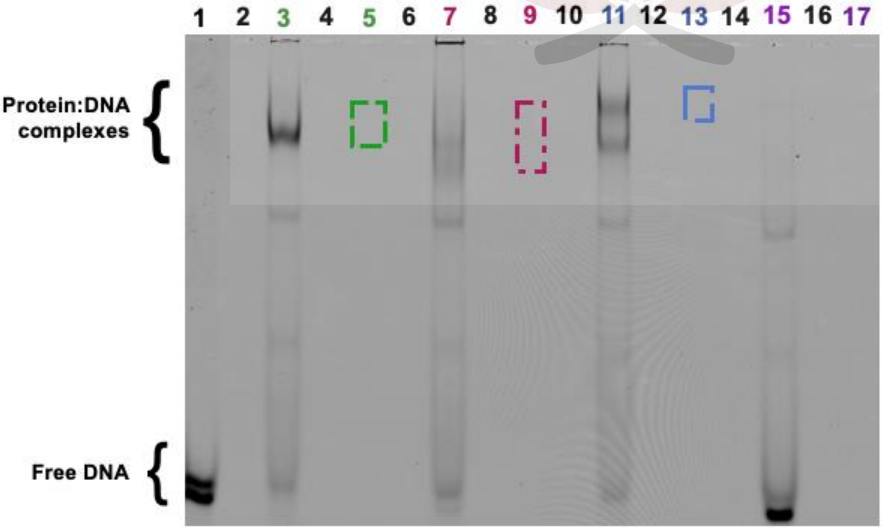
Round 3

Figure 13: Gel 1 of round 3.



Lane	Dilution
1	Binding buffer
2	Empty
3	1/5 GATA4 - ANF
4	Empty
5	1/5 GATA4 - Library
6	Empty
7	1/5 TBX5 - ANF
8	Empty
9	1/5 TBX5 - Library
10	Empty
11	1/10 GATA4 & TBX5 -ANF
12	Empty
13	1/10 GATA4 & TBX5 -Lib.
14	Empty
15	1/5 GFP - ANF
16	Empty
17	1/5 GFP - Library

Figure 14: Gel 2 of round 3.



Lane	Dilution
1	Binding buffer
2	Empty
3	1/5 GATA4 - ANF
4	Empty
5	1/5 GATA4 - Library
6	Empty
7	1/5 TBX5 - ANF
8	Empty
9	1/5 TBX5 - Library
10	Empty
11	1/5 GATA4 & TBX5 -ANF
12	Empty
13	1/5 GATA4 & TBX5 -Lib.
14	Empty
15	1/5 GFP - ANF
16	Empty
17	1/5 GFP - Library

4.6 DNA-binding Logos Determined by Autoseed

Monomeric GATA4

In this section we show a few examples of the GATA4's DNA-binding sequences that were enriched in round 3 of the SELEX-seq. GATA4 was found both as a monomer (**Figures 15.A** and **16.A**) and as a homodimer (**Figures 15.B-15.D** and **16.B-16.D**). These motifs demonstrate that the homodimer was co-bound to DNA with different spacings and orientations.

Figure 15: Examples of the DNA binding sites of GATA4 (Replicate 1)

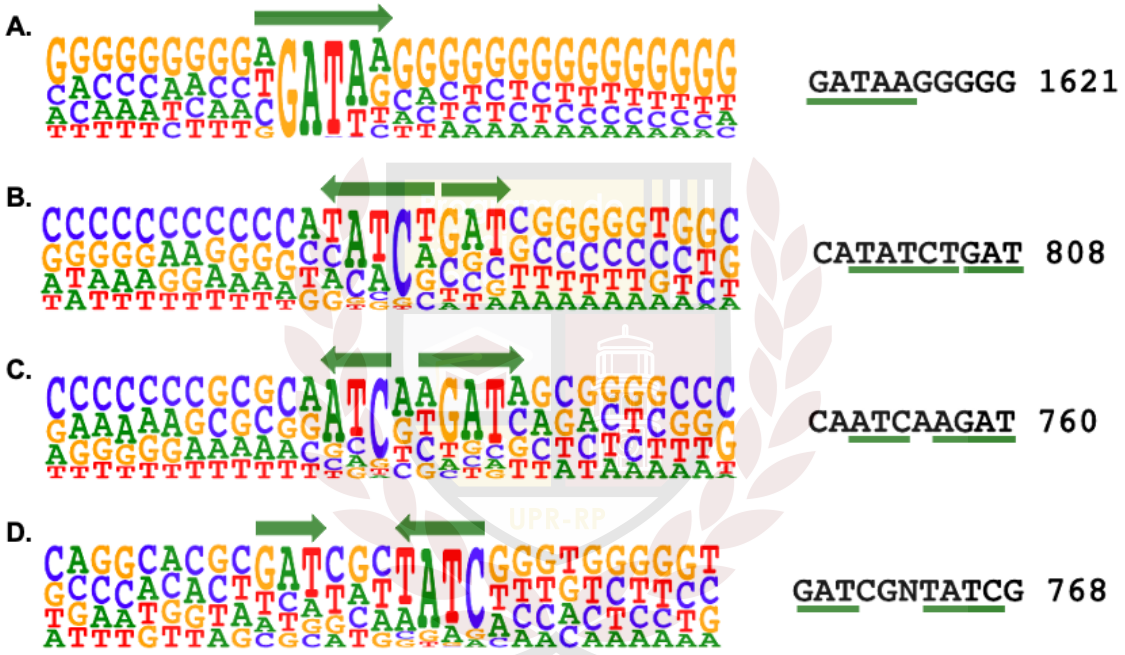
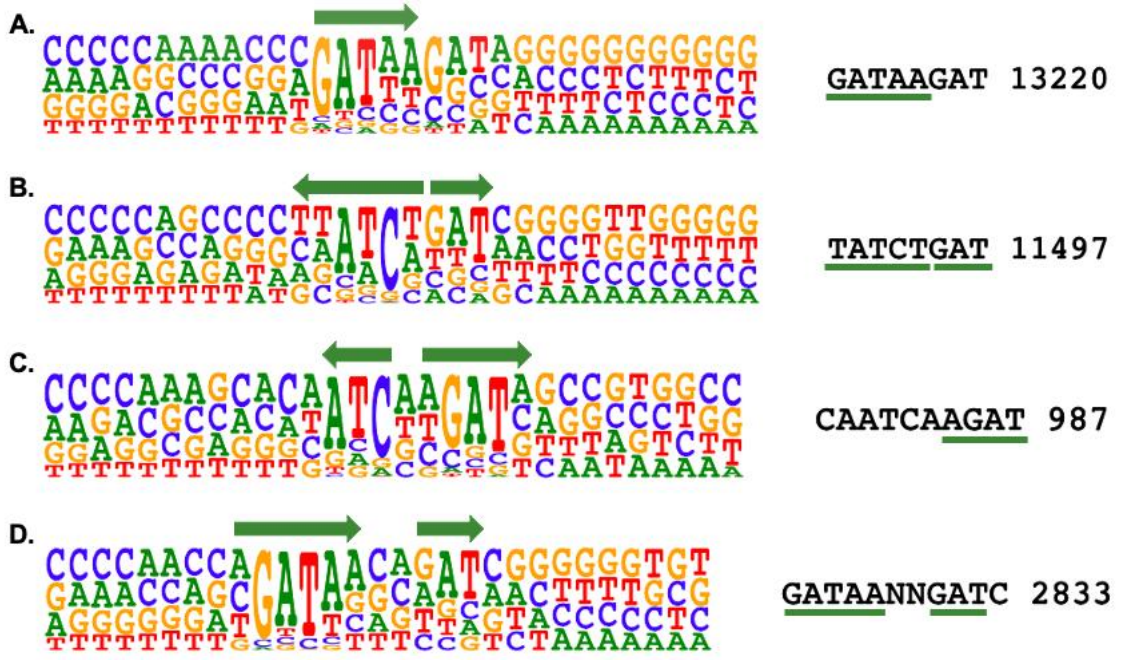


Figure 16: Examples of the DNA binding sites of GATA4 (Replicate 2)



Monomeric TBX5

The following figures exhibit a few examples of the TBX5's DNA-binding sequences that were enriched in round 3 of the SELEX-seq experiment. As expected, TBX5 was more enriched as a monomer (**Figure 17.A** and **18.A**). Surprisingly, it was also bound to DNA as a homodimer (**Figures 17.B** and **18.B**). **Figures 17.C** and **18.C** display a de novo DNA binding motif for TBX5.

Figure 17: Examples of the DNA binding sites of TBX5 (Replicate 1)

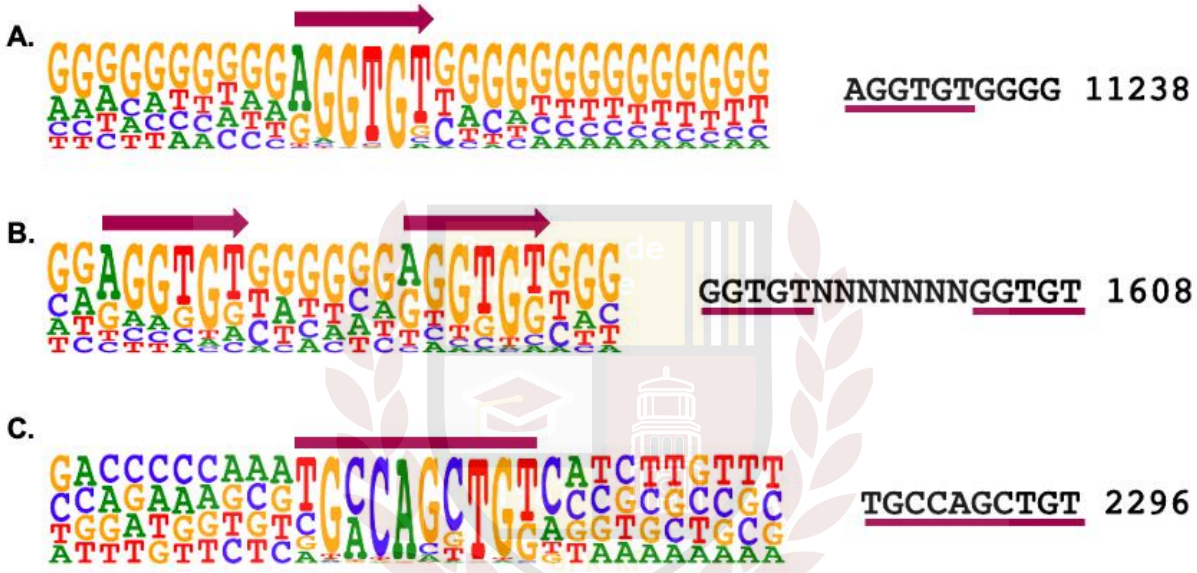
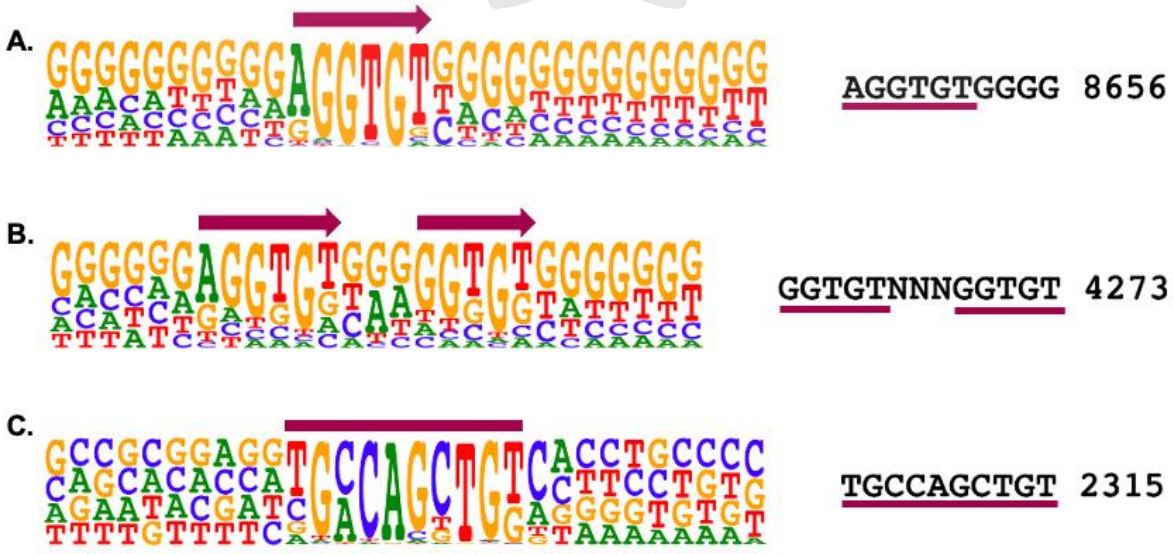


Figure 18: Examples of the DNA binding sites of TBX5 (Replicate 2)



GATA4:TBX5 Complex

Figures 19 and 20 exhibit some examples of the sequences co-bound by GATA4 and TBX5 in round 3 of the SELEX-seq experiment. As expected, the individual TF motifs of the complex can be found at different spacings and orientations.

Figure 19: Examples of the DNA binding sites of the GATA4:TBX5 complex (Replicate 1)

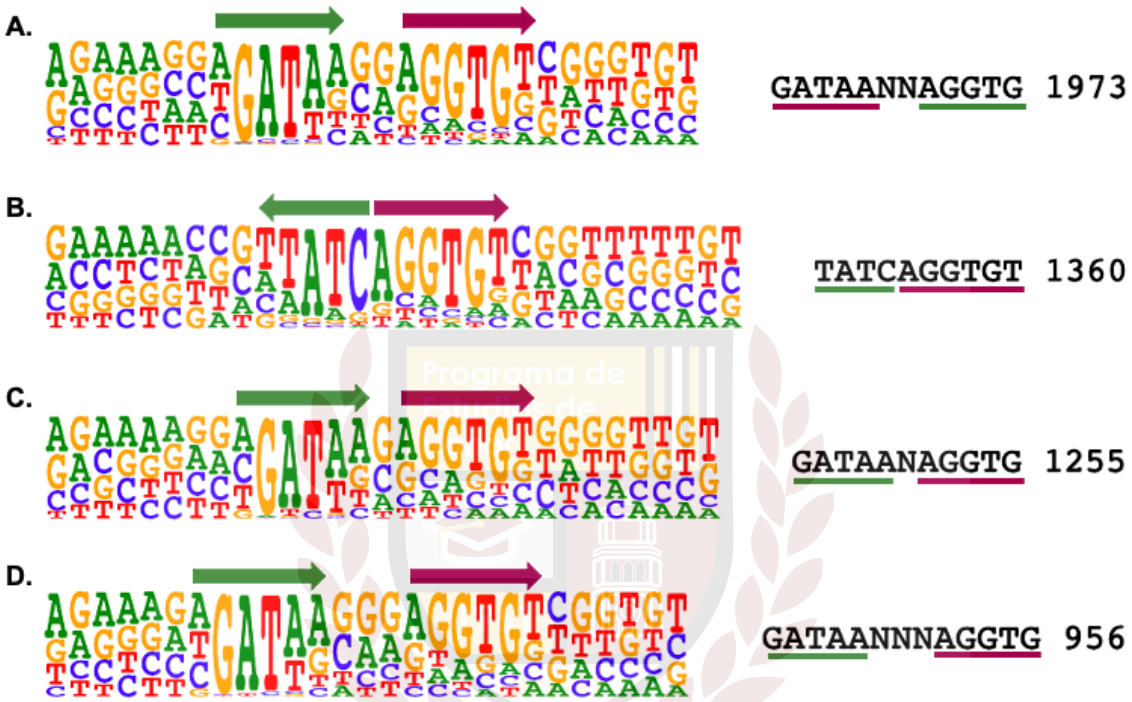
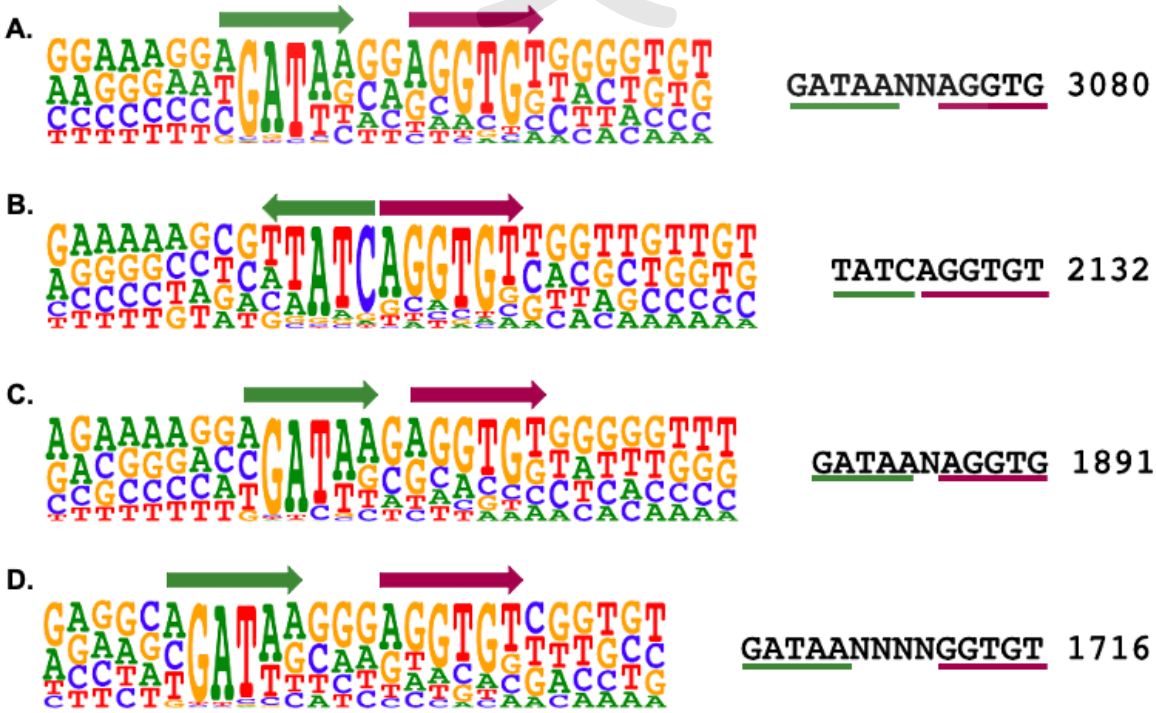


Figure 20: Examples of the DNA binding sites of the GATA4:TBX5 complex (Replicate 2)



Seeds Chosen for Future Computational Analysis

Table 7: This table summarizes the enriched seeds of gel 1, chosen for future computational analysis based on the fold number.

Transcription factor(s)	Chosen seeds	Fold	Total
GATA4	GATAAGGGG	26.7	353
	CCTTATCAC	18.5	124
	AGATAAAGGC	18.0	102
	GATANCGATC	17.7	192
	CCTTATCG	17.7	122
TBX5	AGGTGTGGG	104.8	952
	CCCACACCT	104.3	1,167
	GAGGTGTGG	116.3	997
	GAGGTGTGGG	77.4	1,929
	AGGTGTTG	76.0	671
GATA:TBX5 complex (1/10 dilution)	GATANNNAGGTG	43.9	624
	GATANNNNGGTGT	40.3	677
	GATAANNAGGTG	34.0	1,038
	ACACCNNNNNTTATC	30.1	1,181
	ACACNNNNTTATC	28.9	637

Table 8: This table summarizes the enriched seeds of gel 2, chosen for computational analysis, based on the fold number.

Transcription factor(s)	Chosen seeds	Fold	Total
GATA4	CCTTATCTC	30.6	179
	GAGATAAGG	30.6	153
	GATANNNNCGATC	27.6	337
	GATANNNTATC	25.1	416
	CGATNNGATA	25.0	404
TBX5	CGACACCTC	139.9	833
	AGGTGTCGG	124.6	801
	CAACACCTC	112.5	943
	GAGGTGTCGG	100.4	1433
	CCAACACCTC	93.9	1654
GATA4:TBX5 complex (1/5 dilution)	GATANNNAGGTG	50.5	1013
	GATAANNAGGTG	33.9	1535
	GATANNNNNNNNNGGTG	31.4	661
	GATANNNNNGTGT	31.2	651
	GATAANNNNGGTGT	29.9	1709
	ACACCTGATA	27.2	1005

5. DISCUSSION

As mentioned before, our **central hypothesis** stated that the DNA-binding sequences recognized by the GATA4:TBX5 complex will differ from the specific DNA sequences selected by the monomeric TFs. Additionally, we predicted that the GATA4:TBX5 complex will have strong spacing and orientation preferences. To test our hypotheses, we established two specific aims. First, we wanted to determine the specific DNA-binding sequences of GATA4 and TBX5 as monomers, and of the heteromeric GATA4:TBX5 complex. The second aim was to determine if the GATA4 and TBX5 DNA-binding motifs have strong orientation and spacing preferences when they co-bind to form a cooperative complex.

GATA4's binding motif has been well studied and it has been defined as: AGATAAGA in the forward direction and TCTTATCT in its reverse complement (CIS-BP data base). Our results are compatible with the published data. Figures 15.A and 16.A show that GATA4's monomeric binding motif is as follows: AGATAA. Its reverse complement is in the Supplemental File 1. The EMSA gel had suggested that GATA4 can be a homodimer and we confirmed this with our SELEX-seq results. We identified different spacers and orientations for the homodimer. We found a maximal spacer of 3 bps and a minimal of no bps. There were also different orientations: tail-to-tail (Figures 15.B-15.C and 16.B-16.C). head-to-head (Figure 15.A) and head-to-tail (Figure16.D). The fold number describes how many times a seed appeared in the sample in comparison to the DNA library (Round 0). Based on this parameter, the most enriched motif for GATA4 was the DNA sequence bound by its monomer (Tables 7 and 8). However, we obtained enriched seeds that corresponded to the GATA4 homodimer, even if the fold numbers were less than those of the monomer.

On the other hand, according to the CIS-BP database, TBX5's binding motif is AGGTGTGA in the forward direction and TCACACCT in the reverse direction. We corroborated this because the logos created with Autoseed showed that monomeric TBX5 recognizes the forward sequence: AGGTGT; its reverse complement is in the Supplemental File 1. Surprisingly, TBX5 was capable of binding as a homodimer. Similar to GATA4, we distinguished different spacers and orientations. We found a maximal spacer of 8 bps and a minimal of 1 bp (Supplemental File 1). Different orientations were also present: head-to-tail (Figures 17.B and 18.B) and tail-to-tail (Supplemental File 1). Figures 17.C and 18.C exhibit what could to be a *de novo* binding motif for TBX5. This

sequence [TGCCAGCTGT] appeared among the logos of both gels (duplicates) so we can rule out the possibility that there was an error in some step of the SELEX-seq experiment or its sequencing. However, we did not find this motif in the CIS-BP database or previous research studying this transcription factor. Once we repeat the experiment using His-tagged TFs instead of GST-tagged TFs, we will be able to know if we should further study it or discard it as a potential binding site. If we discover this motif among the results obtained with His-tagged proteins, we could then explore the possibility that it is a low-affinity binding site. Contrary to GATA4, among the enriched seeds chosen for TBX5, no homodimer sequence was found. Based on the fold number (Tables 7 and 8), the most enriched seed was the DNA sequence bound by monomeric TBX5.

The DNA motifs generated by Autoseed exhibited a difference between the monomers and the heteromeric complex. Among the enriched sequences, we detected multiple spacers and orientations. There were motifs with a maximal spacer of 8 bps and a minimal of no bps. The results also showed different orientations including head-to-tail (Figures 19.C, 19.D, 20.C and 20.D) and tail-to-tail (Figures 19.B and 20.B). Interestingly, the most enriched binding motif of the GATA:TBX5 complex has a 3 bp spacer and a head-to-tail orientation, as indicated by the fold umber in Tables 7 and 8. The multiple spacers and orientations distinguishing this cooperative complex demonstrate that TF interactions allow the recognition of new and unique DNA binding sites (Luna-Zurita et al., 2016).

The binding motifs discussed here are just preliminary results because the sequencing files we obtained did not have as many reads as expected. The sequencing step may have yielded less DNA reads due to an error in the process. Consequently, the data included in this thesis gave us preliminary insight into the grammar rules governing the GATA4:TBX5 complex. For this reason, we sent our samples for another round of sequencing to obtain more reads, which will provide us with more information and allow us to analyze the data in depth. Another limitation was that we manually curated the sequences to choose the seeds that will be computationally analyzed, which is an arbitrary approach. We are currently working with different programs such as the R Project to generate Positional Weight Matrices (PWMs) and other analyses that display more information about the affinity and grammar rules of the GATA4:TBX5 complex. Our next step will be to conduct SELEX-seq using His-tagged transcription factors. For future analysis, we will validate the DNA-binding sites of the heteromeric complex using EMSA. Additionally, we want to predict the complex's genome-wide binding sites by comparing our *in vitro* results with ChIP-seq databases from previous *in vivo* investigations. Furthermore, we want to contrast the wild-type binding motifs with the DNA sequences

bound by mutated GATA4 and TBX5. We would determine how the mutations disrupt the grammar rules governing the GATA4:TBX5 complex.

6. CONCLUDING REMARKS

We successfully accomplished our research aims. Our preliminary results demonstrated that the TF monomeric binding motifs differ from the DNA-binding sequences recognized by the heteromeric complex. Additionally, the preliminary logos created by Autoseed showed that the GATA4:TBX5 cooperative complex has spacing and orientation preferences. Based on the enrichment fold, this complex prefers a 3bp spacer. Since we are manually curating our data, we still cannot conclude which orientations are more enriched for the complex. Our next steps in the short term include using systematic computational analyses to better understand the grammar rules of this heteromeric complex and compare our dataset with the SELEX-experiments of His-tagged GATA4 and His-tagged TBX5. We expect that our data will allow us to make better predictions of gene regulatory networks driving heart development.

7. ACKNOWLEDGEMENTS

I want to specially thank my mentor José Rodríguez-Martínez and graduate student Jessica-Rodríguez for guiding and assisting me through this whole process. Additionally, I am grateful for the help provided by the thesis committee, all my all my lab partners and Dr. Peterson's laboratory. This work was supported by: National Institutes of Health (SC1GM127231), Puerto Rico Louis Stokes Alliance for Minority Participation (NSF HDR2008186), NSF Graduate Research Fellowships Program, the NSF Bridge to the Doctorate Program (1826558), and the Sequencing and Genomics Facility of the MSRC/UPR.

8. BIBLIOGRAPHY

Al-Qattan, M. M., & Abou Al-Shaar, H. (2015). Molecular basis of the clinical features of Holt-Oram syndrome resulting from missense and extended protein mutations of the TBX5 gene as well as TBX5 intragenic duplications. *Gene*, *560*(2), 129–136. https://doi.org/10.1016/j.gene.2015.02.017

Andrilenas, K. K., Penvose, A., & Siggers, T. (2015). Using protein-binding microarrays to study transcription factor specificity: Homologs, isoforms and complexes. *Briefings in Functional Genomics*, *14*(1), 17–29. https://doi.org/10.1093/bfgp/elu046

Ang, Y. S., Rivas, R. N., Ribeiro, A. J. S., Srivas, R., Rivera, J., Stone, N. R., … Srivastava, D. (2016). Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. *Cell*, *167*(7), 1734-1749.e22. https://doi.org/10.1016/j.cell.2016.11.033

Bass, J. I. F., Sahni, N., Shrestha, S., Garcia-gonzalez, A., Mori, A., Bhat, N., … Albertha, J. M. (2015). Human Gene-Centered Transcription Factor Networks for Enhancers and Disease Variants. *Cell*, *161*(3), 661–673. https://doi.org/10.1016/j.cell.2015.03.003.Human

Borok, M. J., Papaioannou, V. E., & Sussel, L. (2015). Unique functions of Gata4 in mouse liver induction and heart development. *Developmental Biology Journal*, *410*(2), 213–222. https://doi.org/10.1016/j.ydbio.2015.12.007.Unique

Bruneau, B. G., Nemer, G., Schmitt, J. P., Charron, F., Robitaille, L., Caron, S., … Seidman, J. G. (2001). A murine model of Holt-Oram syndrome defines roles of the T-Box transcription factor Tbx5 in cardiogenesis and disease. *Cell*, *106*(6), 709–721. https://doi.org/10.1016/S0092-8674(01)00493-7

Dixon, J. E., Dick, E., Rajamohan, D., Shakesheff, K. M., & Denning, C. (2011). Directed differentiation of human embryonic stem cells to interrogate the cardiac gene regulatory network. *Molecular Therapy*, *19*(9), 1695–1703. https://doi.org/10.1038/mt.2011.125

Garg, V., Kathiriya, I. S., Barnes, R., Schluterman, M. K., King, I. N., Butler, C. A., … Srivastava, D. (2003). GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature*, *424*(6947), 443–447. https://doi.org/10.1038/nature01827

Jia, J., Ye, T., Cui, P., Hua, Q., Zeng, H., & Zhao, D. (2016). AP-1 transcription factor mediates VEGF-induced endothelial cell migration and proliferation. *Microvascular Research*, *105*, 103–108. https://doi.org/10.1016/j.mvr.2016.02.004

Jimenez-Sanchez, G., Childs, B., & Valle, D. (2001). Human Disease Genes Database. *Nature*, *409*, 853–855.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., … Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, *20*(6), 861–873. https://doi.org/10.1101/gr.100552.109

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., … Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, *152*(1–2), 327–339. https://doi.org/10.1016/j.cell.2012.12.009

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., … Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, *527*(7578), 384–388. https://doi.org/10.1038/nature15518

Judith F. Kribelbauer, Chaitanya Rastogi, Harmen J. Bussemaker, R. S. M. (2019). Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol.*, *35*(3), 357–379. https://doi.org/10.1146/annurev-cellbio-100617-062719.Low-Affinity

Kribelbauer, J. F., Loker, R. E., Feng, S., Rastogi, C., Abe, N., Rube, H. T., … Mann, R. S. (2020). Context-Dependent Gene Regulation by Homeodomain Transcription Factor Complexes Revealed by Shape-Readout Deficient Proteins. *Molecular Cell*, *78*(1), 152-167.e11. https://doi.org/10.1016/j.molcel.2020.01.027

Yadav, R. K., Chauhan, A. S., Zhuang, L., & Gan, B. (2018). FoxO transcription factors in cancer metabolism. *Seminars in Cancer Biology*, *50*, 65–76. https://doi.org/10.1016/j.semcancer.2018.01.004

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., … Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Lodish, H., Berk, A., Zipursky, L., Matsudaira, P., & Baltimore, D. (2000). *Molecular Cell Biology* (4th ed.). New York.

Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., … Bruneau, B. G. (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell*, *164*(5), 999–1014. https://doi.org/10.1002/anie.201602763.Digital

Luscombe, N. M., Austin, S. E., Berman, H. M., & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biology*, *1*(1), 1–37. https://doi.org/10.1186/gb-2000-1-1-reviews001

Maitra, M., Schluterman, M. K., Nichols, H. A., Richardson, J. A., Lo, C. W., Srivastava, D., & Garg, V. (2010). Interaction of Gata4 and Gata6 with Tbx5 is critical for normal heart cardiac development. *Developmental Biology Journal*, *326*(2), 368–377. https://doi.org/10.1016/j.ydbio.2008.11.004.Interaction

Marson, A., Kretschmer, K., Frampton, G. M., Jacobsen, E. S., Polansky, J. K., MacIsaac, K. D., … Young, R. A. (2007). Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature*, *445*(7130), 931–935. https://doi.org/10.1038/nature05478

Misra, C., Chang, S. W., Basu, M., Huang, N., & Garg, V. (2014). Disruption of myocardial Gata4 and Tbx5 results in defects in cardiomyocyte proliferation and atrioventricular septation. *Human Molecular Genetics*, *23*(19), 5025–5035. https://doi.org/10.1093/hmg/ddu215

Morgunova, E., & Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology*, *47*, 1–8. https://doi.org/10.1016/j.sbi.2017.03.006

Nemer, G., & Nemer, M. (2010). *GATA4 in Heart Development and Disease*. *Heart Development and Regeneration* (Vol. II). Elsevier Inc. https://doi.org/10.1016/B978-0-12-381332-9.00027-X

Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, *37*(SUPPL. 1), 77–82. https://doi.org/10.1093/nar/gkn660

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., … Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *ELife*, *2015*(4), 1–20. https://doi.org/10.7554/eLife.04837

Orenstein, Y., & Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, *42*(8). https://doi.org/10.1093/nar/gku117

Orenstein, Y., & Shamir, R. (2017). Modeling protein-DNA binding via high-throughput in vitro technologies. *Briefings in Functional Genomics*, *16*(3), 171–180. https://doi.org/10.1093/bfgp/elw030

Pu, W. T., Ishiwata, T., Juraszek, A. L., Ma, Q., & Izumo, S. (2004). GATA4 is a dosage-sensitive regulator of cardiac morphogenesis. *Developmental Biology*, *275*(1), 235–244. https://doi.org/10.1016/j.ydbio.2004.08.008

Puerto Rico Health Department. (2014). *Informe Anual 2014 Sistema de Vigilancia y Prevención de Defectos Congénitos de Puerto Rico*.

Puerto Rico Health Department. (2017). Surveillance and Prevention System 2017.

Sánchez, M., Jennings, P. A., & Murre, C. (1997). Conformational changes induced in Hoxb-8/Pbx-1 heterodimers in solution and upon interaction with specific DNA. *Molecular and Cellular Biology*, *17*(9), 5369–5376. https://doi.org/10.1128/mcb.17.9.5369

Siggers, T., & Gordân, R. (2014). Protein-DNA binding: Complexities and multi-protein codes. *Nucleic Acids Research*, *42*(4), 2099–2111. https://doi.org/10.1093/nar/gkt1112

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., … Mann, R. S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, *147*(6), 1270–1282. https://doi.org/10.1016/j.cell.2011.10.053.Cofactor

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., & Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences*, *39*(9), 381–399. https://doi.org/10.1016/j.tibs.2014.07.002

Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., … Odom, D. T. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, *154*(3), 530–540. https://doi.org/10.1016/j.cell.2013.07.007

Stormo, G. D., & Zhao, Y. (2010). Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics*, *11*(11), 751–760. https://doi.org/10.1038/nrg2845

Stormo, G. D. (2013). *Introduction to Protein-DNA Interactions: Structure, Thermodynamics and Bioinformatics* (1st ed.). New York: Cold Spring Harbor Laboratory Press.

Suzuki, Y. J. (2011). Cell signalling pathways for the regulation of GATA4 transcription factor: Implications for cell growth and apoptosis. *Cell Signal*, *23*(7), 1–16. https://doi.org/10.1016/j.cellsig.2011.02.007.Cell

Triedman, J. K., & Newburger, J. W. (2016). Trends in congenital heart disease. *Circulation*, *133*(25), 2716–2733. https://doi.org/10.1161/CIRCULATIONAHA.116.023544

Vashee, S., Melcher, K., Ding, W. V., Albert, S., & Kodadek, T. (1998). Evidence for two modes of cooperative DNA binding, 452–458.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, *10*(4), 252–263. https://doi.org/10.1038/nrg2538

Wilkinson, A. C., Nakauchi, H., & Göttgens, B. (2017). Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. *Cell Systems*, *5*(4), 319–331. https://doi.org/10.1016/j.cels.2017.07.004

Zeisberg, E., Ma, Q., Juraszek, A., Moses, K., Schwartz, R., Izumo, S., & Pu, W. (2005). Morphogenesis of the right ventricle requires myocardial expression of Gata4. *Journal of Clinical Investigation*, *115*(6), 1522–1531. https://doi.org/10.1172/JCI23769DS1

Zhang, J., Dong, J., Qin, W., Cao, C., Wen, Y., Tang, Y., & Yuan, S. (2019). Ovol2, a zinc finger transcription factor, is dispensable for spermatogenesis in mice. *Reproductive Biology and Endocrinology*, *17*(1), 2–5. https://doi.org/10.1186/s12958-019-0542-3

Zhu, T., Qiao, L., Wang, Q., Mi, R., Chen, J., Lu, Y., … Zheng, Q. P. (2017). T-box family of transcription factor-TBX5, insights in development and disease. *American Journal of Translational Research*, *9*(2), 442–453.