Constructing Temporal Transcriptional Regulatory Cascades in the Context of Development and Cell Differentiation

Dissertation

for the award of the degree

Doctor of Philosophy Ph.D.

Division of Mathematics and Natural Sciences

of the Georg-August-University, Göttingen

within the doctoral Program for Environmental Informatics (PEI)
of the Georg-August University School of Science (GAUSS)

submitted by

Rayan Daou

From
Lebanon

Göttingen
2020

Thesis advisory committee:

Prof. Dr. Edgar Wingender	Dept. of Medical Bioinformatics, University Medical Center Göttingen
Prof. Dr. Stephan Waack	Institute of Computer Science, Georg-August University of Göttingen
Prof. Dr. Tim Beißbarth	Dept. of Medical Bioinformatics, University Medical Center Göttingen

Members of the examination board:

Referee: Prof. Dr. Edgar Wingender	Dept. of Medical Bioinformatics, University Medical Center Göttingen
Co-referee : Prof. Dr. Stephan Waack	Institute of Computer Science, Georg-August University of Göttingen

Other members of the examination board:

Prof. Dr. Burkhard Morgenstern	Institute for Microbiology and Genetics Dept. of Bioinformatics Georg-August University of Göttingen
Prof. Dr. Winfried Kurth	Department Ecoinformatics, Biometrics & Forest Growth Georg-August University of Göttingen
Prof. Dr. Tim Beißbarth	Dept. of Medical Bioinformatics, University Medical Center Göttingen
Prof. Dr. Ulrich Sax	Dept. of Medical Informatics, University Medical Center Göttingen

Date of the Oral examination: 30.03.2020

Abstract

Cell differentiation is a complex process orchestrated by sets of regulators appearing at precise temporal points, resulting in regulatory cascades that affect the expression of broader sets of genes, ending up in the formation of different tissues and organ parts. The identification of stage-specific master regulators and the mechanism by which they activate each other is a key to understanding and controlling differentiation and still a challenging quest, particularly in the fields of tissue regeneration and organoid engineering.

To tackle this quest I developed a novel workflow and a model I call the Temporal Regulatory Cascade (TRC). The TRC workflow combines a comprehensive general regulatory network based on binding site predictions with user-provided temporal gene expression data, to generate a series of connected stage-specific regulatory networks. The TRC identifies those regulators that are unique for each time point and the regulatory interactions between them, taking into consideration the temporal order of their appearance. The TRC model is represented in the form of a regulatory cascade that shows the emergence of these regulators and regulatory interactions across time. The TRC workflow was implemented in the form of a user-friendly tool with a visual web interface that requires no expert knowledge in programming or statistics, making it directly usable for scientists with no strong computational background. In addition to generating TRCs, the tool links multiple interactive visual workflows, in which a user can track and investigate further different regulators, target genes, and interactions, directing the tool along the way into biologically sensible results based on the given dataset.

The workflow was used to analyze a high-quality dataset that documents the gene expression levels across multiple time points during the differentiation of stem cells into mature cardiomyocytes. In addition to the main dataset, we applied the TRC model to several different time-series expression datasets coming from different contexts such as neural development. The model was successful in identifying previously-known and new potential key regulators, in addition to the particular time points to which these regulators are associated. These results were highly supported by GO enrichment, experimental knowledge and literature. Compared to other methods, our approach showed an advantage in terms of computational time, and the density of the important regulators identified in such small cascades. The workflow is now available publicly at TF-Investigator.sybig.de/TRC.

Zusammenfassung

Die Differenzierung von Zellen ist ein komplexer Prozess, welcher durch eine Reihe von Regulatoren geleitet wird. Das temporäre Auftreten ist genau abgestimmt ist und wird durch zusammen regulatorische Signalkaskaden gesteuert, die die Expression von Genen beeinflussen und damit letzten Endes zur Bildung von unterschiedlichen Geweben und Organen führt. Die Identifikation von Zustands-spezifischen Masterregulatoren und deren gegenseitige Aktivierung ist einer der Schlüsselaspekte um, Zelldifferenzierung verstehen und kontrollieren zu können und ist gegenwärtig eine herausfordernde Aufgabe in den Gebieten der Gewebsregeneration und Organzüchtung.

Um diese Herausforderung anzugehen, habe ich einen neuen Workflow und das Modell Temporal Regulatory Cascade (TRC) entwickelt. Der TRC-Workflow kombiniert ein allgemeingültige globales regulatorisches Netzwerk, dass auf der Vorhersage von DNA-Bindungsstellen basiert, um auf Basis von Benutzern zur Verfügung gestellten Expressionsdaten von Zeitreihenexperimenten Signalkaskaden von zusammenhängenden zeitpunktspezifischen, regulatorischen Netzwerke zu erstellen. Der TRC-Workflow identifiziert exklusive Regulatoren für einen bestimmten Zeitpunkt, sowie deren regulatorisches Zusammenspiel, wobei das zeitliche Auftreten der Regulatoren berücksichtigt wird. Das TRC-Modell wird in Form einer regulatorischen Kaskade repräsentiert, die das Auftreten und die regulatorischen Interaktionen der Regulatoren über die Zeit hinweg darstellt. Der TRC-Workflow wurde als benutzerfreundliches Programm inklusive einer Weboberfläche implementiert, welche ohne großes Expertenwissen in Informatik oder Statistik verwendet werden kann. Zusätzlich zur Erstellung von TRCs verbindet das Programm mehrere interaktive Workflows in denen Nutzer unterschiedliche Regulatoren, Zielgene und Interaktionen identifizieren können.

Der TRC-Workflow wurde angewendet, um einen hochqualitativen Datensatz zu analysieren, der die Genexpressionsstärken zwischen einer Vielzahl an unterschiedlichen Zeitpunkten der Differenzierung von Stammzellen in reife Herzmuskelzellen abbildet. Zusätzlich zu diesem Datensatz, habe ich das TRC auf mehrere Zeitreihen-Expressionsdaten von unterschiedlichen Hintergründen, wie zum Beispiel die neuronale Entwicklung angewendet. Das Modell hat erfolgreich bereits bekannte und neue potentielle Masterregulatoren in den Zeitpunkten zu denen diese Regulatoren ursprünglich zugeordnet wurden identifiziert. Die Ergebnisse wurden mit Hilfe von *GO-Enrichment*, Expertenwissen und Literaturstudien belegt. Im Vergleich zu anderen Methoden zeigt mein Ansatz einen Vorteil hinsichtlich Rechenzeit und der Dichte der identifizierten Regulatoren in kleinen Kaskaden. Der Workflow ist über TF-Investigator.sybig.de/TRC öffentlich verfügbar.

Acknowledgments

As I reflect back on my PhD period, I can clearly see that reaching this goal has only been possible with the help, motivation, and contributions of the people I had the luck to be surrounded with in this journey.

First of all, I would like to express my gratitude for Prof. Dr. Edgar Wingender for having me at the department of bioinformatics under his supervision, through which I gained the experience and knowledge needed to write this thesis. He helped me polish my ideas through his vast knowledge, constructive criticism and sharp observation.

A big thanks as well for Martin Haubrock, whose dedication to his work is truly inspiring, for his patience, valuable expert suggestions, his guidance through every step of my journey, and for giving the time to answer my questions no matter how busy he was.

I want to thank Prof. Dr. Tim Beißbarth for giving me the time to finish my PhD and for bringing a great addition of colleagues to the department.

Further, I would like to acknowledge and thank all the members of my thesis committee: Prof. Dr. Stephan Waack, Prof. Dr. Burkhard Morgenstern, Prof. Dr. Winfried Kurth, and Prof. Prof. Dr. Ulrich Sax. They were kind with their valuable time, I cannot thank them enough for that!

A great deal of thanks goes for my mentor Kifah who, despite the distance, managed to help all the time.

In my daily work at the department, I was blessed with a friendly and cheerful group of fellow students and scientists whose impact was extremely valuable on the academic and personal levels. From Sebastian who helped me get started, to Mehmet who showed me how to efficiently get things done, to Torsten for his technical support and Doris for her impressive organizational skills. Colleagues whom I saw on a daily basis, Gregory, Maren, Halima, Darius, and many others, had a big influence on my work as well. And of course, a big special thanks to Conni for being such a great office mate and a dear friend, and for the daily scientific and non-scientific conversations that brought joy to the office.

I can't thank Natalie enough for her patience and precious help in different aspects of my life. Having such a positive person around was crucial in stressful times.

Eventually, I want to thank my family and friends in Lebanon for their motivation and support through all my studies, the support that got me where I am today.

Table of Contents

List of Figures

List of Tables

Acronyms

DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic acid
GO	Gene Ontology
GRN	Gene Regulatory Network
hPSC	Human Pluripotent Stem Cell
iPSC	Induced Pluripotent Stem Cell
miRNA	Micro RNA
mRNA	Messanger Ribonucleid Acid
PS	Peak Strength
PSC	Peak Strength of a Cascade
PSS	Peak Strength of a Stage
PWM	Positional Weight Matrix
RNA	Ribonucleic acid
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TPP	Template Peak Pattern
TRC	Temporal Regulatory Cascade

1 Introduction

Cell differentiation, the driving force in development, is responsible for the diversity of cell types and organs that is behind the complexity of eukaryotic organisms. Orchestrating such precise differentiation events is the work of a set of complex regulatory programs that exert the needed control on the timing, cell type, and spatial coordinates of the differentiating cells. Typically, a handful of master regulators start the regulatory mechanism that results in the activation or repression of other regulators and non-regulatory genes, which by themselves express proteins that lead to the activation or repression of other genes. These regulatory waves emerge in exact temporal order and pace, to give rise, through consecutive unique stages, to different cell types and tissue layers that end up forming complex functioning organs. Our understanding of such regulatory programs and their dynamics is still in its infancy. However, a massive wave of scientific interest and research has been ongoing recently to decode and reverse engineer such programs.

The wave of cell differentiation research started mainly with the discovery of induced pluripotent stem cells (iPSCs), that coincided with the decreasing prices of genetic high-throughput methods such as RNA-Seq. Through this wave, medical applications based on manipulating cell differentiation emerged and experiments geared towards developing stem cell therapies and organoid engineering became increasingly popular. Scientists were able to run various differentiation experiments and take transcriptional snapshots of tens of thousands of genes at different time points of the experiment. These experiments led to the generation of a significant number of temporal gene expression datasets that needed to be analyzed to provide a basis to reconstruct the underlying regulatory programs that drove this expression.

Various computational approaches were applied to analyze such sets, and most of them aim at either identifying a candidate list of genes or deriving relevant gene regulatory networks (GRNs). However, both quests turned out to be challenging, as these general methods do not take into consideration the unique properties of cell differentiation. Candidate gene lists generated by methods such as those that identify differentially expressed genes (DEGs) were usually long and contained only few context-relevant regulators and genes, which is a challenge for the experimentalists that usually look for concise sets of candidates for experimental verifications. Methods for constructing gene regulatory networks from these temporal datasets suffered from challenges such as excessive computational time, and the considerable difference between the number of the genes under study, usually in the order of tens of thousands, compared to the number of time points in the experiment, which were

typically less than ten. In addition to that, the results of the constructed GRNs were large networks of thousands of nodes and hundreds of thousands of interactions, which are hard to distill into useful starting points for experimentalists.

As these gaps in these computational approaches persisted, I decided to take the challenge of developing a method to reconstruct transcriptional regulatory programs in the context of cell differentiation. To tackle this challenge, I needed to create a model and a workflow that can do the following:

- Integrates protein-DNA binding information
- Integrates temporal gene expression data effectively
- Utilizes the temporal order and integrates into a cascade-like architecture
- Identifies stage-specific master regulators
- Proves to be biologically relevant
- Generates concise, information-dense results
- Is computationally efficient
- Can be used by experimentalists with no computational background

Based on these points, I developed a model and a workflow that constructs a series of concise interconnected stage-specific regulatory networks that form a temporal cascade. This model, which I call the Temporal Regulatory Cascade (TRC) model, uses time-series gene expression data combined with a comprehensive regulatory network based on transcription factor binding site predictions to generate the regulatory cascade. In this cascade architecture, stage-specific regulators are identified based on their expression pattern and placed accordingly in their temporal order, and then relevant regulatory interactions are queried from the background network. The model gives a glimpse on the emergence and disappearance of the regulatory waves across time, as well as the potential role of particular regulators within these waves. This workflow was implemented in the form of a web service where a user can upload his own time series dataset and automatically get a visual representation of the custom regulatory cascade based on the relevant experiment. The web service included other workflows that allow the user to explore aspects of co-expression and co-regulation in his dataset interactively, to obtain biologically sensible results. The web service is fast, user-friendly, visual and easily usable by scientists with no statistical or programming background and is publicly available at http://tf-investigator.sybig.de/TRC.

In order to investigate the ability of this method to deliver biologically sensible results, I applied the TRC workflow to construct a cardiac differentiation regulatory cascade based on a high-quality dataset that is generated from an experiment that monitored the differentiation of stem cells to mature cardiomyocytes. The workflow was successful in capturing a set of previously known cardiac regulators and identifying their precise temporal

role during differentiation, as well as identifying new potential regulators that might enhance the differentiation process. The workflow was also applied to various datasets and cell differentiation contexts, and consistently had similar positive results which were analyzed in details and compared with the existing literature and experiments, to merely lay in place some pieces of the big puzzle of developmental biology.

1.1 Thesis Structure

The remainder of this thesis is structured as follows. In Chapter 2, I present an overview of the biological facts and mechanisms that provided the basis for the work done as well as the current and potential medical applications of manipulating cell differentiation. In Chapter 3, I go through some of the main state of the art computational approaches that share common aims with the model I developed, such as regulatory network inference methods and approaches for analyzing temporal gene expression, highlighting the advantages and disadvantages of these methods and tools. In Chapter 4, I introduce the main material and methods that were used throughout the thesis, such as experimental methods, databases, software libraries, tools and data sources. Afterwards, I present the main results of my PhD work, from enhancing some existing methods to developing the novel TRC model and the accompanying web tool to applying the developed workflows to several data sets and evaluating the results from a biological point of view. This is followed by a discussion in Chapter 6, which covers aspects such as optimizing the model evaluating its significance and comparing it to other existing comparable tools. The thesis ends with a conclusion part that summarizes the work done and provides insights for future work that can extend the work done.

1.2 Impact

Publications:

I published the TRC model with the accompanying tool in addition to applying it to analyze differentiation experiments in the following manuscript:

1. Daou, R, Beißbarth, T, Wingender, E, Gültas, M, Haubrock, M (2019). *Constructing temporal regulatory cascades in the context of cell differentiation.* PLoS ONE. (Under revision).

Conferences and Workshops:

The work presented in this thesis was presented as posters in the following workshops and conferences:

- Workshop on Bioinformatics of Gene Regulation (Göttingen 2018)
- RECOMB/ISCB Conference on Regulatory and Systems Genomics (New York City 2018)

Web-Service:

I provided a web-server that incorporates different workflows into an interactive visual webtool for investigating regulatory forces in temporal gene expression data. The tool is publicly available at the following URL: http://tf-investigator.sybig.de/TRC/ .

Student Projects and Supervision:

The author supervised the following students:

- Alessandro Consorte**:** *Predictions based on co-expression data and probable co-regulation identifies interesting roles and functions of early expressed genes in human heart development.* (Project)
- Sofia Marina Guerin Darvas: *Co-expression and co-regulation analysis of transcription factor-gene pairs associated to cardiomyocyte differentiation.* (Project)
- Liza Vinhoven: *Attempt to find candidate Master Regulators in Human Heart Development and Comparison of Tools* (Project)
- *Tobias Haar: Identifying differentially expressed mircroRNAs in RTQPCR datasets (Project).*
- Christian Steinmeyer: *Webtooling for Co-expression Analysis* (Project).
- Lukas Faiss: *The new tool NetFader - Prediction of gene clusters and their related transcription factors in human heart development* (Project).

2 Biological Background

Throughout millions of years of evolution, from the simple single-celled organisms to the more complex eukaryotic organisms, biological systems became more diverse and complex. Understanding and reverse engineering elements of such systems have always been the pursuit of biologists throughout history. With the discovery of microscopes, a whole world opened for scientists, the world of a cell. From that point on, a succession of discoveries led to a deeper understanding of cellular components and cell division. The next breakthrough was with the discovery of the DNA's (Deoxyribonucleic Acid) structure, by James Watson and Francis Crick. Genetics took another leap; thousands of researchers went on quests that unraveled different aspects and complexities and gave rise to a set of new biological questions. In the following subsections, we introduce some basic yet essential biochemical components and concepts that lay the bedrock for the research done in this manuscript.

2.1 The Genomic Organization

Deoxyribonucleic acid (DNA), which is considered to be the blueprint for living things, is a molecule that contains information used in everyday metabolism and enables cells to develop and work together to form a fully functional body. A given DNA strand contains a sequence of bases: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). These bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a backbone of sugar and phosphate molecules. A nucleotide refers to the combination of a base along with a sugar and a phosphate moiety. Those nucleotides are pieced together in two long strands forming a spiral called a double helix (Figure 1).

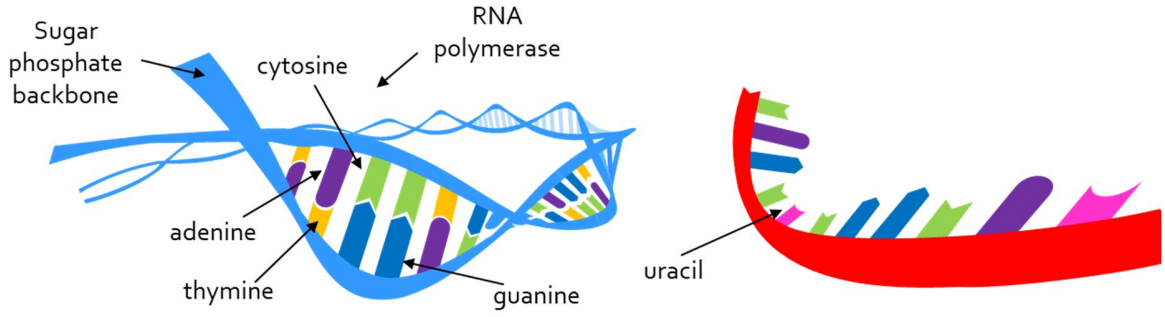

Figure 1. (Left) The structure of the DNA with the different base pairs forming the double helix (Right) An RNA single strand with Uracil instead of Thymine.

In a eukaryotic cell, the DNA is packaged tightly to prevent it from being damaged and the strands from being entangled. For this packaging, the cell uses histones, positively charged proteins, around which the negatively charged DNA wraps forming complexes called nucleosomes. Nucleosomes fold up to form the chromatin fiber, which is compressed, folded, and coiled, forming the chromatid of a chromosome.

A gene is a section of the DNA that can range from hundreds to millions of base pairs that carry the instructions for the synthesis of a product that could be RNA or protein. A gene has a Transcription Start Site (TSS) that indicates the beginning of the gene and a transcription stop site that marks its end. A gene has also a start codon and a stop codon, its importance covered in the next section. Upstream from each gene is a promoter region, where certain regulators can bind to the DNA and control the activity of the associated gene. (Figure 2)

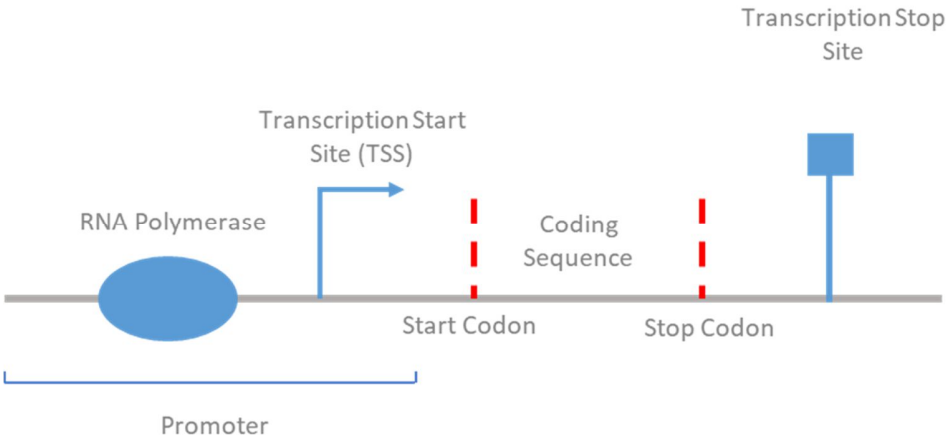

Figure 2. The genomic composition around a gene, with the promoter region upstream from the TSS and the transcription stop site in the end.

2.2 Mechanisms of Gene Expression

Since the DNA is merely a blueprint, genes still have to undergo a process in which their code is read and used to produce corresponding proteins upon which the cell and the whole organism would function. For this to happen, two main steps have to occur, transcription and translation.

The main molecule involved in these processes is the RNA (Ribonucleic Acid), which is a polymer similar to the DNA in some aspects (Figure 1). It is made out of a single strand of nucleotides that can fold on itself. The sequence of bases similar to that of the DNA for the exception of Uracil (U) instead of Thymine (T). RNA molecules are typically much shorter

than DNA polymers, not exceeding a few thousand base pairs in length. Different types of RNA, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), play different roles in the regulation and protein synthesis.

Transcription is the process by which an RNA molecule is synthesized using a DNA segment as a template. This process starts with an enzyme called RNA polymerase binding to the promoter of a gene, separating the DNA strands and adding matching RNA nucleotides to one of them. The transcription ends with the newly synthesized RNA strand separating, to be later translated into proteins. (Figure 3)

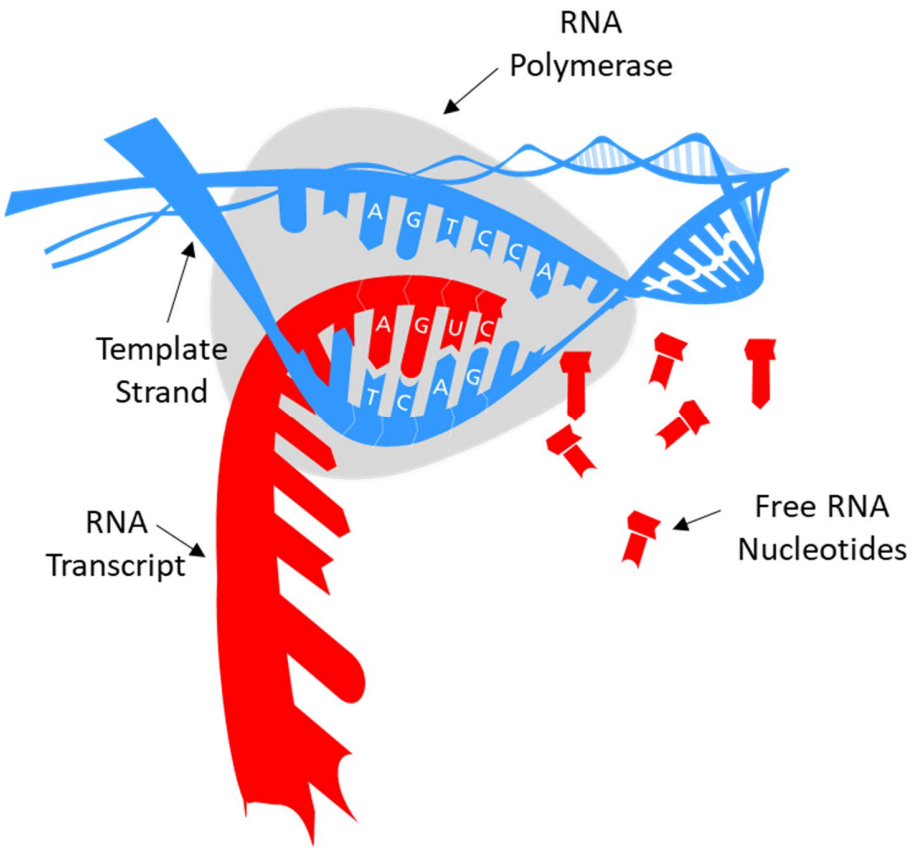

Figure 3. The process of transcription in action. The RNA polymerase separating the DNA strands and adding the RNA nucleotides to form an RNA transcript.

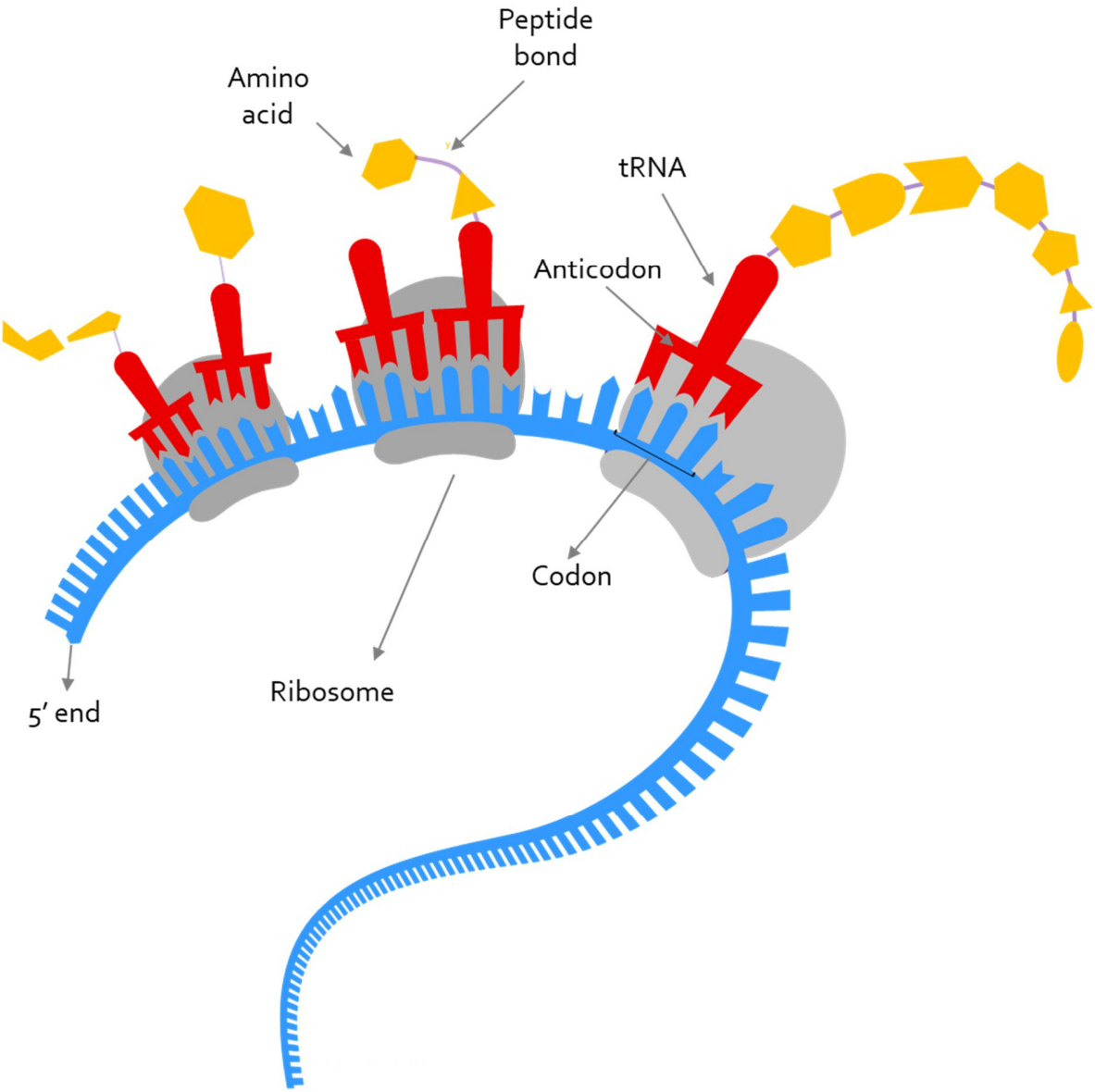

Figure 4. The process of translating an mRNA segment. The ribosome assembling itself around the mRNA and tRNAs binding to matching codons extending the amino acid chain.

Translation is the process of synthesizing proteins based on mRNA templates. It goes through the following stages in eukaryotic organisms:

Initiation: The ribosome initiates the translation assembling itself around the mRNA, and a tRNA carrying the amino acid methionine attaches itself to the matching codon AUG, known as the start codon.

Elongation: It is the stage where the amino acid polypeptide chain is extended, one amino acid at a time. A tRNA binds to a new codon, and the carried amino acid is linked to the existing chain. The next codon in the mRNA is then exposed for reading, and the process repeats.

Termination: When the ribosome encounters a stop codon, it starts the process of separating the chain from the tRNA and ejects it out.

Afterwards, the polypeptide chain goes on to fold into a 3D shape or combines with other polypeptides forming a functional protein (Figure 4).

2.3 Gene Regulation

Despite having the same DNA, cells within the same organism differ in terms of their type, shape, functions and the proteins synthesized. And despite being present in the code, not all genes are expressed at the same time and conditions; they are rather used selectively by regulatory mechanisms. Gene regulation is the mechanism by which gene expression is controlled, either positively by activating the gene or negatively by repressing its expression. Regulation in most genes can occur at one or more of the following levels:

Chromatin level: The accessibility of the chromatin is a determining factor of whether a gene gets expressed or not. Open chromatin around the region of the gene makes it possible for regulators and the transcription machinery to access and start transcribing the gene, while a tightly packed one can be a barrier.

Transcriptional level: It is the primary regulatory level and the main focus of this manuscript. The leading players in the transcriptional regulation are transcription factors (TFs).

A TF is a protein that has the affinity to bind to the DNA, particularly in the promoter regions of genes controlling their activity (Figure 5). Where the TF binds to DNA is determined by the

DNA-binding domain (DBD) of the TF matching a particular associated nucleotide pattern in the DNA called a binding site. Many genes are regulated by several transcription factors, with a specific combination needed to turn the gene on. TFs, in a way, allow the cell to use molecular logic and process information to turn on and off genes depending on the type of the tissue, environmental stress, and many other variables.

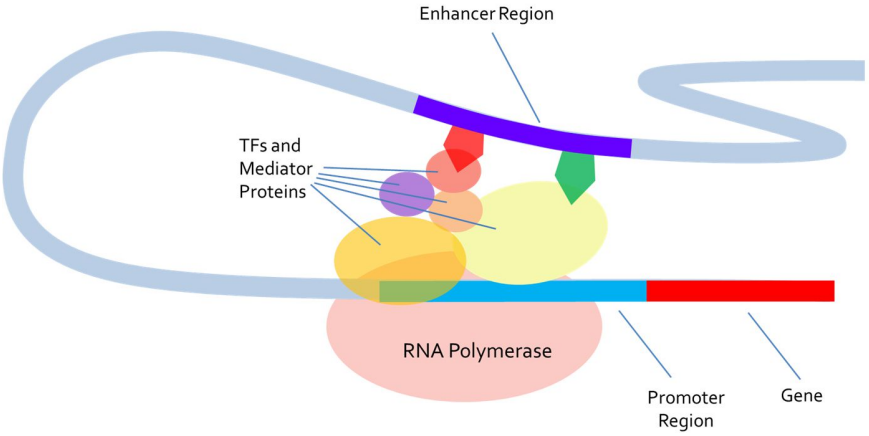

Figure 5. Elements of the proximal and distal regulatory mechanisms in action. Some regulatory proteins binding to the promoter and others binding further away but bending the DNA accordingly to contribute to the regulatory complex.

Post-transcriptional level: mRNA segments resulting from transcription undergo different modifications before they reach the translation stage. Manipulations such as capping, slicing, alternative splicing, editing, and the addition of poly(A) tail to the RNA segment, regulate the final sequence, availability, and half-life of the mRNA that is ready for translation. MiRNAs are small RNAs that have the capability to bind to mRNA segments and chop them, effectively suppressing the expression of the corresponding gene. Depending on how well it matches in its binding to the mRNA, sometimes miRNAs can block the process of translation of an mRNA segment rather than causing its degradation.

Protein level: Proteins undergo editing, cleaving, and folding with the help of various other molecules, which affects their activity and behavior. Phosphorylation is another common post-translational regulatory mechanism, where a phosphate group attaches to a protein activating, deactivating, or modifying its behavior.

2.4 Cell Differentiation

Cell differentiation is the process in which cells change their type, functionally or morphologically, as they divide and multiply. Differentiation is the essence of eukaryotic development, where cells multiply and morph into drastically different types in the right time and place, giving the rise for different types of tissues and organs. Most cells are originally derived from stem cells.

Stem cells are cells that have the ability to differentiate into different specialized cell types. These special cells are usually of embryonic origins, prominently found in the cell mass and blastocysts during the early stages of development. Stem cells can also be obtained after development, through the blood from the umbilical cord after birth, or even the bone marrow, adipose, or the blood of an adult, and referred to as somatic stem cells in such cases. Embryonic stem cells, which are typically hard to obtain, are pluripotent, meaning they have the ability to differentiate into any cell type. While somatic stem cells, though much easier to obtain and isolate, are multipotent, meaning they can differentiate only into particular closely related cell types.

In the year 2002, Shinya Yamanaka made a breakthrough by discovering a method to produce pluripotent stem cells from fibroblasts by adding a small set of TFs (Myc, Oct3/4, Sox2, and Klf4), reverting these adult cells into a pluripotency stage [1]. This discovery led to a development in the induced pluripotent stem cell research and later through the years, multiple groups of researchers successfully generated better and better qualities of iPSC. Scientists later utilized these iPSCs and differentiated them into different types of cells such as neural cells and even reprogramming them to create a whole organ such as a liver. Stem cells proved to be a handy tool to study development, differentiation, and gene regulation. However, in order to understand and effectively manipulate stem cells, a deep understanding of the regulatory mechanism that governs cell differentiation is necessary.

2.5 Medical application

One of the biggest motivations that have driven scientists throughout history to understand the human body was to overcome common diseases. Nowadays, a handful of diseases are responsible for most of the deaths in the world and constitute the most prominent challenges to our health. Surpassing cancer, chronic respiratory diseases, and diabetes, cardiac related diseases contribute annually to more deaths than any other disease (Figure 6). It is estimated that around 17.9 million people die each year because of heart-related issues. A myocardial infarction takes place every 25 seconds and up to half of these heart attacks are ultimately

fatal, and around 320 billion dollars are spent annually on these issues by the health care system in the US alone. Those staggering numbers have motivated more medical and biological research in the direction of understanding the cardiovascular diseases and the heart on all levels. Harnessing that knowledge into a medical application that can save thousands of lives every month from premature death.

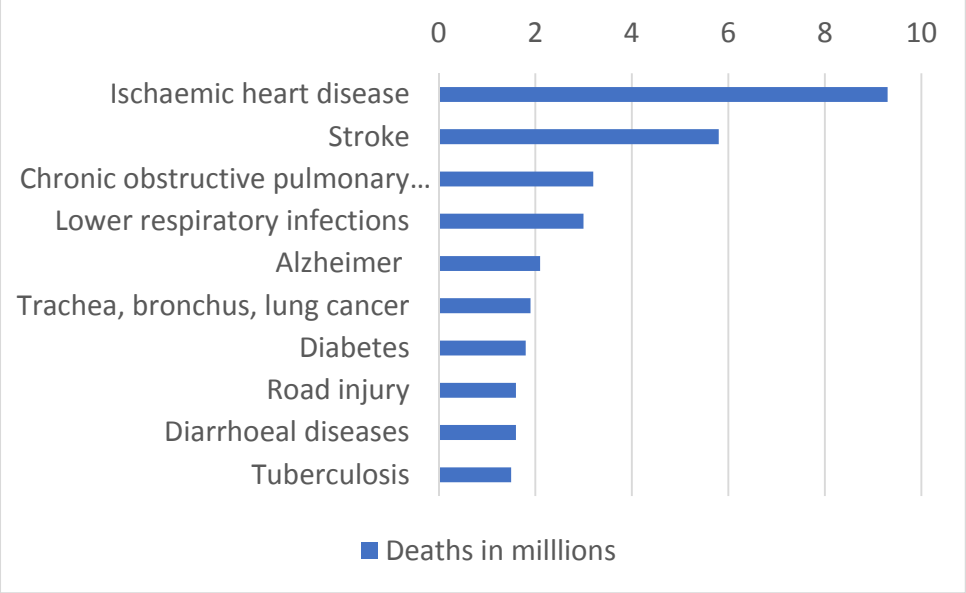

Figure 6. Top ten global causes of deaths and their associated number of deaths in 2016. The dominance of cardiac related diseases is outstanding.

One major approach for the treatment of cardiac diseases has been through the use of drugs such as beta-blockers or various calcium channel blockers. However, this approach is inadequate to restore cardiac function. A heart transplant is another way of dealing with heart failure, but its impracticality is evident when it comes to providing enough hearts for the millions that need it, aside from the other problems such as the high risk of rejection. Devices such as the Implantable cardioverter defibrillator (ICD) or Left ventricular assist device (LVAD) can be used to temporarily enhance cardiac function in the case of heart failure, but they are expensive, cumbersome and entail many complications and problems.

The main reason behind the high rates of cardiovascular diseases is the fragile nature of the human heart. Despite its efficiency in pumping blood to meet the demands of the different tissues and organs of the body, the function of the human heart can easily and fatally be disrupted. For example, in the case of myocardial infarction, the blood flow to a portion of the heart is blocked through a cholesterol blockage, which leads to the death of billions of heart muscle cells within hours. These muscle cells are replaced by a scar, and this causes dilation of the left ventricular chamber, and its ability to contract and squeeze blood is compromised, causing eventually heart failure. Impacts like these lead to a series of changes

in the structure and function of the heart that include fibrosis, which causes stiffening of the heart loss of pump function and cardiac arrhythmia, where the heart develops an abnormal electrical beat, and ultimately these changes can culminate in complete heart failure or sudden death.

This would not have been as big of a problem if the heart had regenerative properties like other organs such as the human liver or skin. For example, if you cut the human skin, it can repair itself completely seamlessly, so does the liver if one surgically removes fifty percent of it, the rest will grow back a completely new liver exactly the same size and structure as the original liver. That mechanism has been lost in the hearts of mammals, including humans, leaving the adult human heart unable to generate new heart muscle cells and repair itself. However, some organisms like fish and salamanders have remarkable regenerative powers. If, for example, the fins, the legs or the tail of a salamander or a fish are amputated, they will grow back the limb to exactly the same size, structure, and function as the original limb, and interestingly the same goes for their heart. If half of the heart of a salamander or a fish is surgically amputated, they will grow the heart right back to the exact same size and structure. While this feature is not present in adult mammals, scientists discovered that neo-fetal mice displayed the ability to regenerate their hearts. Newly born mice were taken, and twenty percent of the apex of their heart was surgically amputated, remarkably it was found that these hearts could completely regenerate back to normal (Figure 7). On the first day they observed a clot formation plugging the leak in the ventricle, then on day two they saw inflammation of this region, by day 7 they found muscle cell proliferation, and by day 21 there was a complete disappearance of the wound and the heart was completely restored to normal structure and function. However, if the amputations are delayed and the heart was injured one week after birth, the regenerative process starts to diminish, and more delay would cause a bigger scar and less regeneration. This experiment is quite important because it indicates that the mechanisms, genes, proteins, and signals that are required to regenerate the heart really do exist in a mammal such as a mouse and presumably in a human, but somehow these are silenced later in life [2]. This leads to the hypothesis that there must be biological pathways, genes, and mechanisms that can do this regeneration and raised the question of why they are switched off in the adult part of humans. Finding the biochemical key to unlock such process could have enormous high implications.

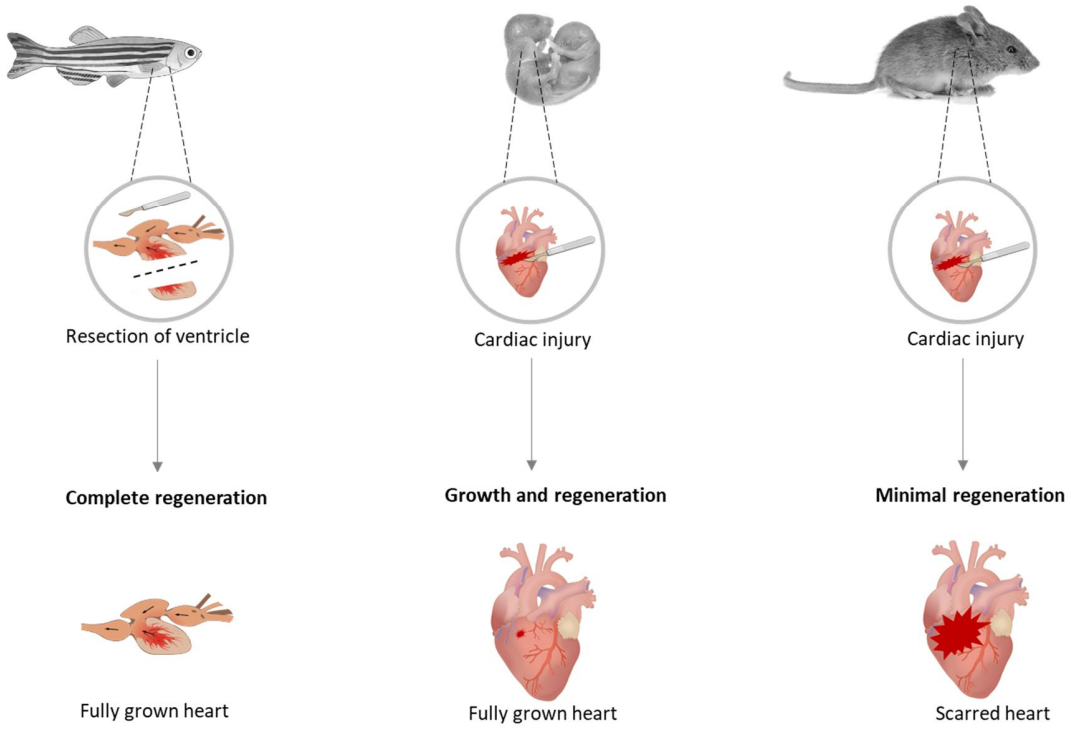

Figure 7. (Left) Adult zebrafish regenerating cardiac muscle lost from resection of the ventricular apex. (Middle) Neonatal mice showing a regenerative response to cardiac injury (Right) Adult mice show minimal regeneration in response to injury (based on a figure from [2]*).*

New approaches emerged to harness stem cell technology, and developments on the medical application level have been made in the past years. Cardiomyocytes derived from iPSCs are used as patches that are transplanted into affected areas of the heart (Figure 8). However, this approach still faces many challenges and still requires optimization. The heart is highly electrically integrated, and the disruption of this electrical integration by injecting foreign cells into the heart can cause arrhythmias, where the new cells will be pulsing at a different rate and intensity than the rest of the heart. Adjusting such variables and tuning in the cells requires additional extensive research and further experimentation to decode the exact mechanisms that govern them.

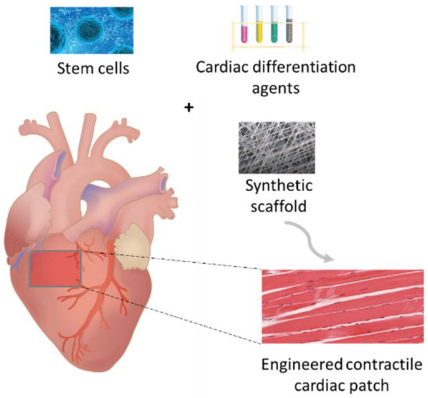

Figure 8. Using stem cell-derived cardiomyocytes for cardiac regeneration. Stem cells are differentiated into mature cardiomyocyte, assisted with synthetic or natural scaffolds, which are transplanted cardiomyocytes into the affected area in the heart to stimulate re-growth.

Ultimately, the aim of medical research in this field is to find a way to directly reprogram resident fibroblasts that exist in the human heart, via manipulating genetic and regulatory programs using factors and drugs, turning them into cardiomyocyte-like cells, without the need of a surgery or a transplant. To reach this point, medical research is aimed towards understanding cardiac differentiation on a deep molecular level, and currently, large funds are allocated towards solving the puzzle of the regenerative human heart.

3 Bioinformatics Background

3.1 Regulatory Networks

Gene regulatory networks are usually represented as graphs where nodes represent different genes and edges, which are typically directed, represent the potential effects of one gene on the other. The edges can hold more specific information about such interactions and their types in more complex types of networks.

The construction of such regulatory networks has always been a challenge. Depending on the type and quality of the data used for such construction, the liability, size, and type of the reconstructed networks vary. Some methods are based on expression data as an input and try to predict the effects of genes on each other based solely on the variation of expression levels across different conditions or time points. Other methods use ChIP-seq data and other experimental inputs that are based on detecting regions of DNA bound by certain TF proteins. More complex methods evolved, combining several approaches and data inputs to generate more robust networks that could aid later in experimental design, decisions, and conclusions. What follows in this section is an overview of some of the main approaches for the construction of GRNs.

Boolean networks provide a basis for one of the simplest methods for deriving GRNs. Using a threshold-based discretization, gene expression levels are presented in terms of two states, 1 for expressed and 0 for non-expressed [3]. It then attempts to find Boolean functions for every gene in the network. However, this classical method suffers from information loss due to the harsh discretization and the threshold choice. Certain methods such as Reverse Engineering algorithm (REVEAL) extend this classic approach by adding the in-degree value of genes and utilizes mutual information but suffer from extensive computational time, thus suitable for analyzing a smaller set of genes [4]. Other methods that are based on the same principle have been developed, such as probabilistic Boolean networks, although improve on the original model still suffers from some of the inherited disadvantages [5].

Bayesian networks (BNs), which effectively represent probabilistic relations between variables, are popular models for deriving GRNs [6]. Despite its efficiency in dealing with noisy data, the classical BN model cannot deal particularly with time-series data and feedback regulations, nor do they take into consideration the time lagging that usually occur in real GRNs. Dynamic Bayesian networks (DBNs) were developed with advanced features

that allowed them to handle time-series data, as well as hidden variables and missing data points effectively at the expense of computational time [7]. Versions of DBNs were developed depending excessively on prior knowledge, particularly information about transcriptional regulation to increase prediction accuracy [8].

Differential equations were also an intuitive base for several GRN models. Ordinary differential equations (ODEs) can effectively model feedback loops and show good accuracy, especially in small scale networks. However, besides its expensive computational time, it is hard to describe the non-additive logic of gene regulation in ODEs. The difference equations model, unlike ODEs, uses discrete variables, which leads to information loss, but gives it the edge when it comes to dealing with time-series data [9].

Association networks are undirected graphs that are used to describe GRNs. It draws an edge between two genes that are, for example, co-expressed, without indicating which is regulating the other. ARACNE uses mutual information in combination with information about TFs and their binding sites to infer GRNS [10]. Graphical Gaussian models (GGMs) attempt to infer large GRNs using partial correlation [11].

Dynamic Regulatory Events Miner (DREM) attempts to reconstruct dynamic regulatory networks from time series expression data and protein-DNA interaction data. DREM uses a Hidden Markov Model (HMM) and identifies the genes associated with each bifurcation point [12]. A more elaborate version, iDREM, was developed to visually represent the bifurcation points and integrates several other sources [13]. However, this method still suffers from the significant difference between the size of the gene set and the number of time points and has an excessive computational time.

Collateral-Fuzzy Gene Regulatory Network Reconstruction (CF-GeNe) uses a fuzzy c-means clustering algorithm to construct GRNs, which allows it to deal with noisy and missing data [14]. Other models such as Finite State linear model [15], State-space model [16], and many other approaches and methods emerged and been used for particular cases successfully and unsuccessfully in the past years, and many other will continue to be developed and optimized especially with the growing scale of data available [5] [17] .

3.2 Binding Site Analysis

A Position weight matrix (PWM) is a model representation of a pattern or profile. While a PWM can be used to represent different types of profiles, in this manuscript, we refer to the PWMs that represent the binding profiles of TFs. A PWM summarizes the frequency by which a specific nucleotide appears at a particular position in the profile. It is extracted from the alignment of TF binding sites sequences, identified by techniques like DNase-seq and ChIP-seqs, and the occurrence of each nucleotide at each position is counted and summarized in a matrix. This matrix can be visualized, for example using a logo plot. PWMs are later used as an indicator to evaluate the likelihood of a particular transcription factor to bind to a specific segment of DNA, thus used via different algorithms to predict potential binding sites along the whole genome (Figure 9).

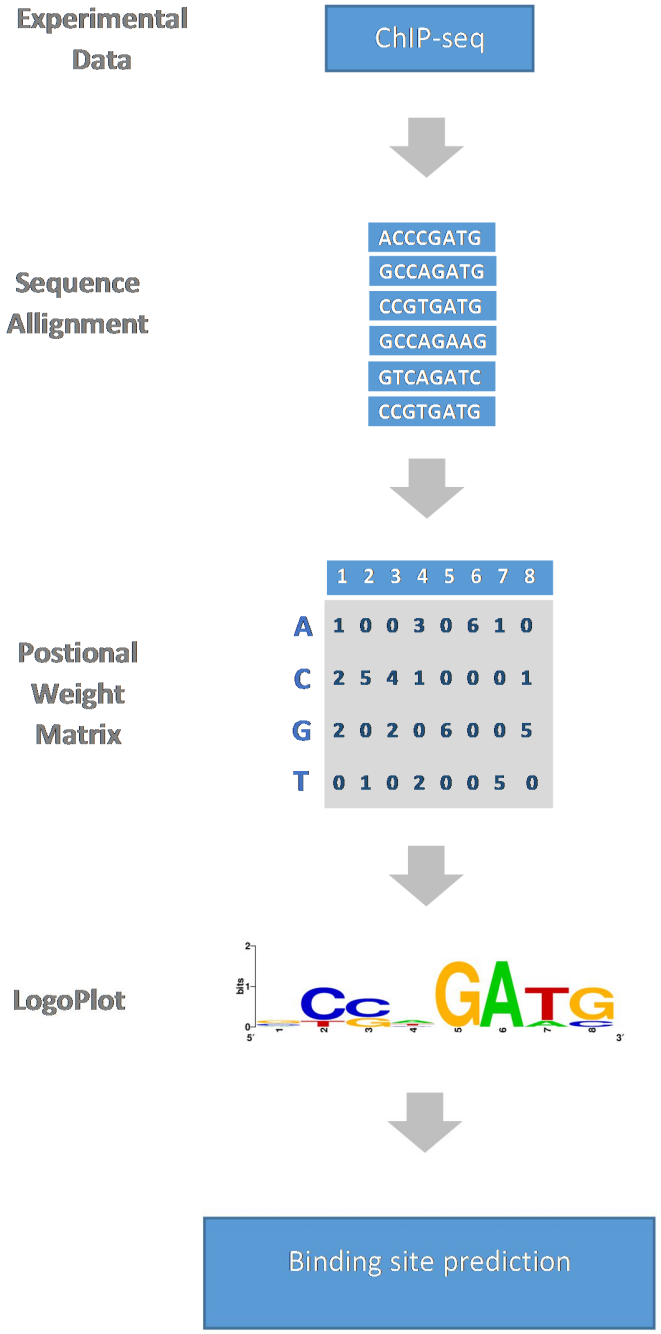

Figure 9. A workflow that illustrates the typical steps for binding site predictions from deriving motifs to utilizing them for predicting binding sites.

3.3 Gene Expression Analysis

Owing to their decreasing prices, methods like RNA-seq and microarrays have generated and keep generating thousands of gene expression datasets. These high throughput technologies allowed the parallel analysis of tens of thousands of genes and their transcripts with a single experiment. Among these sets, I developed a particular interest in time-series datasets. Gene expression time series experiments provide insight into the molecular biology processes inside an organism over time. Time series experiments attempt to study the variation in transcription after stress such as starvation or a drug application or on the gene activation through an evolving process such as differentiation or organ development as in most of the cases covered in this thesis. The result of such experiments is typically a series of snapshots of the gene expression at different consecutive time points are obtained, compiled, formatted, and normalized accordingly. Such datasets not only provide a glimpse of the gene expression in a cell or group of cells but also shows the dynamics of such expression and its change across time, providing more information to capture than the static sets.

With gene expression data sets, scientists face the challenge of having to analyze in parallel thousands of genes with usually only a few conditions or time points and sometimes no replicates. The experiment if not well designed, can add the problem of under-sampling where a lot of key information is missed, and the accuracy of the results is affected. Another challenge is in the biological variability between individuals and even cells of the same individual, which might be in a different cell-cycle stage. In the process of collecting the cells and preparing them for a process such as RNA-seq, these cells are actually destroyed and the data for the next time point or condition is taken from different cells where the variability mentioned before might arise.

Despite these challenges, different methods were developed to analyze temporal expression datasets. These methods vary vastly in their approach and objective, answering different questions and generating different types of results. The following paragraphs attempt to summarize some of the common methods and tools that are used for identifying Differentially Expressed Genes (DEGs), detecting gene clusters and other various approaches usually applied to analyze gene expression data.

To identify DEGs between two gene expression time courses, a method that uses the maximal difference of the area between the linear or spline interpolated gene expression measurements across time was proposed.

A number of methods generate a gene ranking based on differential expression across time. A multivariate empirical Bayes approach can be used to sort genes according to their

differential expression within one or between two or more gene expression temporal datasets [18]. An approach by Kalaitzis et al. ranks differentially expressed genes using a likelihood ratio quotient or a Bayes factor after modeling gene trajectories by Gaussian process regression [19]. Mean Absolute Rank Difference (MARD) constructs gene relationship networks for the control and treatment time courses, measures the differences in the neighbourhood of each gene between the two networks, eventually identifying DEGs based on the significant changes in their estimated neighbourhood [20].

Other methods directly model the gene expression under various conditions and experimental designs directly on the discrete sampled time series. An example of that would be the regression-based statistical modeling used in combination with permutation tests to find significantly differentially expressed genes [21]. Limma attempts to fit linear models to the gene expression values and uses moderated tests in the analysis of variance (ANOVA) framework to assign significance to its findings [22][23]. ANOVA models were also applied in combination with *F*– or permutation tests to identify significant time-group-interactions or the effects of experimental groups [24]. In order to remove the variance caused by individual differences, a modified repeated measure of ANOVA was proposed [25] . The idea of utilizing a principal component analysis (PCA) for a dimension reduction of the estimated parameters from an ANOVA model in multiple series time course experiments was also suggested and applied [26].

Alternatively some tools use Hidden Markov models (HMM) for identifying DEGs in gene expression time-course experiments. Non-homogeneous HMMs are to classify genes between the two states equally expressed and differentially expressed at each time point [27]. Hidden spatial-temporal Markov random fields are used to identify genes, which are differentially expressed at each time point in the context of known biological pathways [28].

Other approaches model the measured gene expression trajectory as a continuous function in time. Gene-wise hypotheses testing can be used on the integral of the quadratic difference between the B-spline curves of two aligned gene expression time-series experiments [29]. Extraction of Differential Gene Expression (EDGE) Identifies differentially expressed genes via a procedure that fits a natural cubic spline representation of the gene expression trajectory under the alternative hypothesis and a constant mean curve under the null hypothesis. Permutation testing based on the residual sums of squares of both models assigns significance to the detected differentially expressed genes [30]. A functional hierarchical model that uses basis expansion to model gene expression trajectories was utilized by Hong and Li to identify temporally differentially expressed (TDE) genes [31]. Bayesian Analysis of Time Series (BATS) is a popular tool that analyzes one-sample time series [30] [31]. The functional Bayesian approach expands the gene temporal profiles over an orthonormal basis and assigns significance for differential gene expression in the form of

Bayes factors. A functional ANOVA mixed-effects model can be used to identify either non-parallel differentially expressed genes or parallel differentially expressed genes [34]. A functional principal component analysis can also be used to test for changes in the temporal gene expression under different conditions [35].

Different Clustering algorithms are typically applied to identify modules of co-expressed genes that have similar expression patterns over time. The common hypothesis behind these clustering approaches is that the genes that are expressed in a similar manner across time are likely to be co-regulated by a set of common regulators and/or are involved in the same biological process or functions. The clustering methods can be divided into three main fields, the similarity-based approaches, the model-based procedures, and template-based methods, which attempt to recognize genes with a gene expression time profile similar to predefined patterns.

Weighted correlation network analysis (WGCNA) is a clustering method that uses a modified Pearson correlation coefficient to detect gene modules. WGCNA was implemented a popular R software package that includes a collection of other functions for constructing networks, topological analysis, and visualization [36].

Some clustering algorithms need a predetermined total number of clusters as in the case of the k-means procedure [37] or in the self-organizing map (SOM) framework [38]. In order to group genes with unknown function to clusters with a priori known function Brown and Grundy supervised a learning algorithm based on support vector machines (SVMs) [39]. CLICK is an algorithm that identifies homogeneous gene expression clusters based on graph-theoretical and statistical techniques [40]. First and second-order differences between adjacent time points can also be used to evaluate the similarity and cluster genes accordingly [41]. Gene Shaving is an algorithm that applies sequential PCA techniques to identify those genes, which are largely varying across time and coherent to each other at the same time [42]. Clustering can also be based on a rank order-preserving matrix framework or by identifying minimum mean squared residue clusters [43]. TimeClust is a tool that implements different clustering techniques like Bayesian clustering [44].

On the other hand, other approaches cluster genes by model fitting their expression trajectory in time, and/or applying a specific clustering model. A corrupted clique graph model can be used efficiently for the non-hierarchical clustering of genes [45]. An algorithm that attempts to fit a mixture of multivariate Gaussian distributions to the gene expression values can be found in the popular package MCLUST [46]. Genes can also be clustered based on their involvement in a specific biological process based on a biological kinetic model [47]. Expectation Maximization (EM) is an algorithm that is used to cluster genes on the basis of their cubic spline representation in a predefined number of sets [48]. Cluster analysis of gene expression dynamics (CAGED) is a pseudo-Bayesian agglomerative clustering approach applied on auto-regressive gene expression models [49]. A similar approach based on polynomial models for describing the gene expression trajectory in the framework of a

Bayesian hierarchical mixture model was also published [50]. The EM algorithm can be used to fit a mixed-effects model on the B-spline representations of the gene expression profiles [51] or for modeling a mixture of simplified differential equations in order to cluster genes according to their temporal expression [21]. Clusters can also be identified using the EM algorithm to fit mixtures of linear models or linear mixed models [52] . A rejection-controlled EM algorithm is used in a mixture of mixed-effects models, in order to estimate the class assignment and the corresponding mean expression curves is used [32]. Another model was developed based on clustering linear HMMs, and the Graphical Query Language (GQL) [53]. An approach for the analysis of gene expression time series infer gene clusters from finite mixtures of HMMs while using prior information in a semi-supervised learning framework was also proposed [54]. Microarray Significant Profiles (MaSigPro) identifies gene clusters of differentially expressed genes by a two-step regression approach, where the algorithm is based on the similarity of the gene-wise regression model coefficients [55]. A Bayesian hierarchical clustering of nonlinear regression spline representation of the temporal trajectories was also proposed [56].

Some clustering approaches attempt to identify statistically significant patterns of expression in the data, and the genes associated with them, based on permutation or resampling procedures. EPIG is a method that uses a multi-step filtering procedure to generate representative candidate patterns from the gene expression data [57]. An order-restricted inference methodology defining candidate temporal profiles in terms of inequalities among the mean expression levels at the time points was proposed [58].The ORICC algorithm groups the gene trajectories according to an order-restricted information criterion to pre-specified candidate inequality profiles [59]. StepMiner aims to detect genes with one or more binary transitions across the gene expression time series by modeling segment-wise constant adaptive regression [60]. GOALIE uses linear time logics to identify spans in the time series and separate gene clusters with similar gene expression patterns in these spans [61]. Short Time-series Expression Miner (STEM) matches the gene expression profiles on data-independent, chosen model profiles and applies a time point permutation test to assign significance to the corresponding gene clusters [62]; afterwards, a Fisher test is used to identify GO gene sets enriched with genes from significant clusters. Springer et al. proposed a data-driven selection of model profiles, which gains a better fit to the data structure, but with the drawback of losing the significance assessment for the identified clusters [63].

Some methods have a different approach for analysing gene expression data compared to the previously described DEG and clustering methods. In order to generate hypotheses about the function of genes not yet annotated to any predefined GO gene set, Hvidsten et al. used a systematic supervised learning approach based on learning a classification rule model within the rough set framework and then evaluating it by cross validation [64]. A unified mixed effects model is constructed for the mean trajectory of every gene set to capture those sets where 20 – 50 % of the genes follow the same trend [65]. On the other hand the more elaborate GlobalANCOVA fits a linear model to the gene expression value for every gene set and identifies those groups, in which a design factor such as treatment-time interaction, is significant in contrast to a reduced model [66]. A nonparametric Wald-type test statistic is also used in combination with a permutation-based test to detect treatment effects or treatment-time interactions in predefined sets of genes [67]. Principal components Analysis through Conditional Expectation (PACE) proved effective in estimating the mean trajectory function for sparse longitudinal data [68]. MaSigFun fits regression models to the gene set expression values in the time series assuming that all group genes follow the same underlying trajectory. PCA-maSigFun is a version of the latter method where more than one model profile per group is allowed [69].

A general drawback of most of the computational methods described in this section is that they very often have to deal with large data sets, listing hundreds of genes, making it hard for the biologists to go through each manually and renders the results too general and broad to be conclusive. Another drawback is the typical black box, where the biologists find it hard to understand exactly how the results where computationally produced, thus less confidence in using the results for the next experimental validation. As most computational approaches develop black-box algorithms, there is a demand for developing ready to use interactive visual tools. Experimentalists could use these tools to explore dynamically and track different genes and other aspects of gene regulation which can involve their biological intuition and deep understanding of the experimental context which is usually and understandably unavailable at the computational side.

3.4 Network Visualization

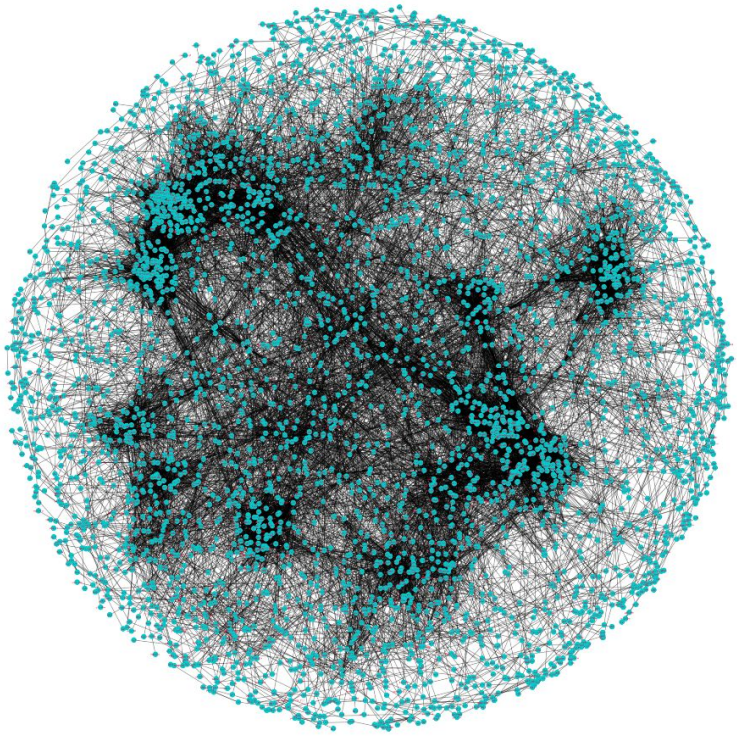

Figure 10. A classic hairball view representing the problem of visualizing large networks.

Humans are visual creatures whose brains have evolved to recognize and classify visual patterns. Looking at a network representation as an edge list or an adjacency matrix might be easy for a computer program but can prove useless for the human brain. On the other hand, representing a network as a visual graph is intuitive when it comes to the human eye. Thus, since the early days of bioinformatics, biological networks have been displayed and drawn as simple basic graphs. But this worked until the complexity and size of such networks increased, where the simple method leads to what is described in visualization as the hairball problem. Networks with thousands of nodes and millions of edges, when presented as a graph, can give little information besides a general view on the topology of such a network (Figure 10). Several solutions have been developed to be able to make sense visually of such networks. One common approach is clustering nodes into modules that are highly connected internally and loosely connected to each other, providing a summary of the different components of the network. This clustering can be based on different features such as biological function and connectivity.

Cytoscape is one the most popular network visualization tools in the bioinformatics community. Cytoscape is a software that provides a framework where different visualization apps can be integrated [70]. Cytoscape includes a large number of apps and is continuously expanding due to the contributors in the Cytoscape community [71]. Since this thesis is highly based on time series data, those apps that can integrate temporal expression data and visualize dynamic networks are of main focus. Some apps stand out in Cytoscape as being able to deal with dynamic networks. DyNet is a plugin that compares two or more networks to identify the most 'rewired' nodes [72]. CyDataSeries allows the user to enrich networks with data series such as gene expression measurements across multiple time points or conditions and supports some basic manipulation with the data series [73]. DynNetwork is a plugin for importing, visualizing, and analyzing dynamic networks in Cytoscape. The dynamics of nodes, edges, attributes can be visualized at the desired time points. To improve visualization, the nodes can be animated with a dynamic force layout algorithm. Some other approaches require that the user enters the precomputed dynamic network in the form of successive networks and then visualizes each of them separately in consecutive snapshots. This leads to a problem where the nodes change places making it hard to track a particular node through the different time points.

Most of the bioinformatics visual apps and software have to be locally installed and mostly require importing a resulting network from another pipeline. This aspect discourages and intimidates non-bioinformaticians from exploring the possibilities and potential of such tools. Hence, one of the goals of the developed workflow is to have it implemented also in a webtool making it simple to used requiring merely few clicks and no programming prerequisites thus reaching for a wider audience. Another problem I wanted to avoid is presenting networks that are too big to be analyzed visually. Thus, I focused on finding a way to distill large networks into concise ones that contain the high-confidence results.

4 Materials and Methods

4.1 RNA-seq

RNA-seq is one of the popular methods for quantifying gene expression. An RNA-seq experiment starts with isolating the RNA material from the sample under study, then generating a cDNA complementary fragment for each RNA fragment. Afterwards, the RNA is washed out, leaving the cDNA to be sequenced and mapped to the genome. The result is a quantification of these transcripts, which is transformed to gene expression levels. RNA-Seq typically costs more than a microarray experiment but the added benefits outway the additional costs most of the time. As the prices of RNA-Seq decrease rapidly, its utilization in experiments increases and an expanding resource of RNA-Seq based gene expression data is available.

4.2 ChIP-seq

ChIP-seq is a method used to determine where specific proteins, like transcription factors, structural proteins, and protein modifications, bind to the DNA. Chip-Seq uses chromatin immunoprecipitation (ChIP) followed by sequencing through the following steps (Figure 11):

Step 1: Cross-link the protein to be studied to DNA.

Step 2: Destroy the cell and shred the DNA strands with the attached proteins into segments of hundreds of base pairs each using sonication.

Step 3: Add bead-attached antibodies into the DNA segments. These antibodies will bind to the target protein and precipitate them along with the DNA segments that are connected to it.

Step 4: Unlink the protein from the DNA strands.

Step 5: Sequence the DNA segments.

Step 6: Map the sequenced segments onto the genome.

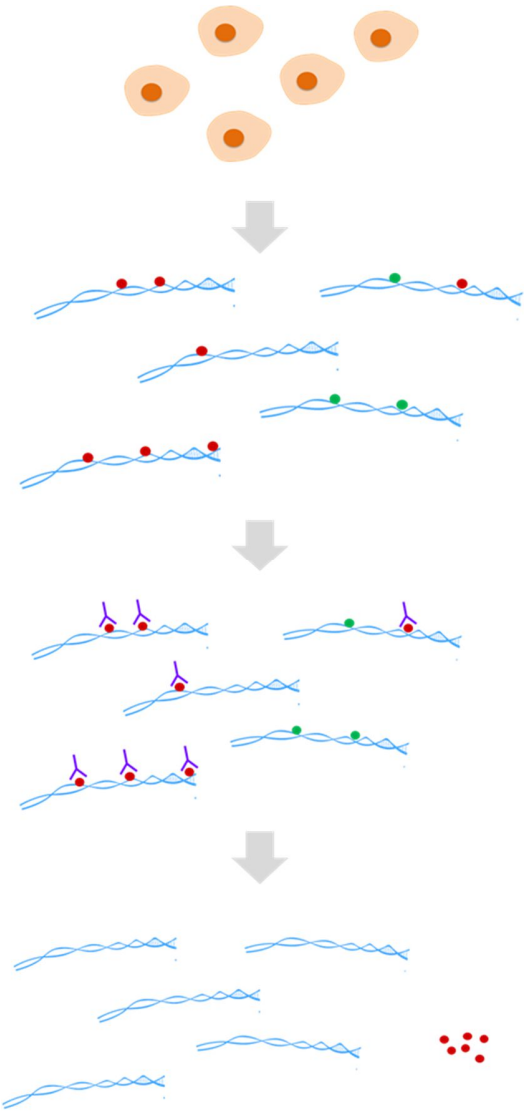

Figure 11. The ChIP-seq experimental workflow.

.

Eventually, Chip-Seq generates a list of target DNA binding sites for the TF under study. These sites are usually hundreds of base pairs long which decreases the accuracy of finding the exact actual active site where that protein binds within that region. However, different techniques have been used to estimate that based on the location of the peak of the Chip-seq signal.

4.3 TRANSFAC®

TRANSFAC® is the most comprehensive database that stores information which revolves around transcriptional regulation in eukaryotes [74]. TRANSFAC® is manually curated, maintained, and available from GeneXplain at http://genexplain.com. It stores millions of experimentally verified DNA binding sites along with with details such as the exact position in the genome, the experimental method used and the DNA sequence corresponding to the site. The binding sites are aligned in order to construct a comprehensive library of thousands of PWMs. This PWM library provides a good quality resource for the different workflows available at GeneXplain (Figure 12).

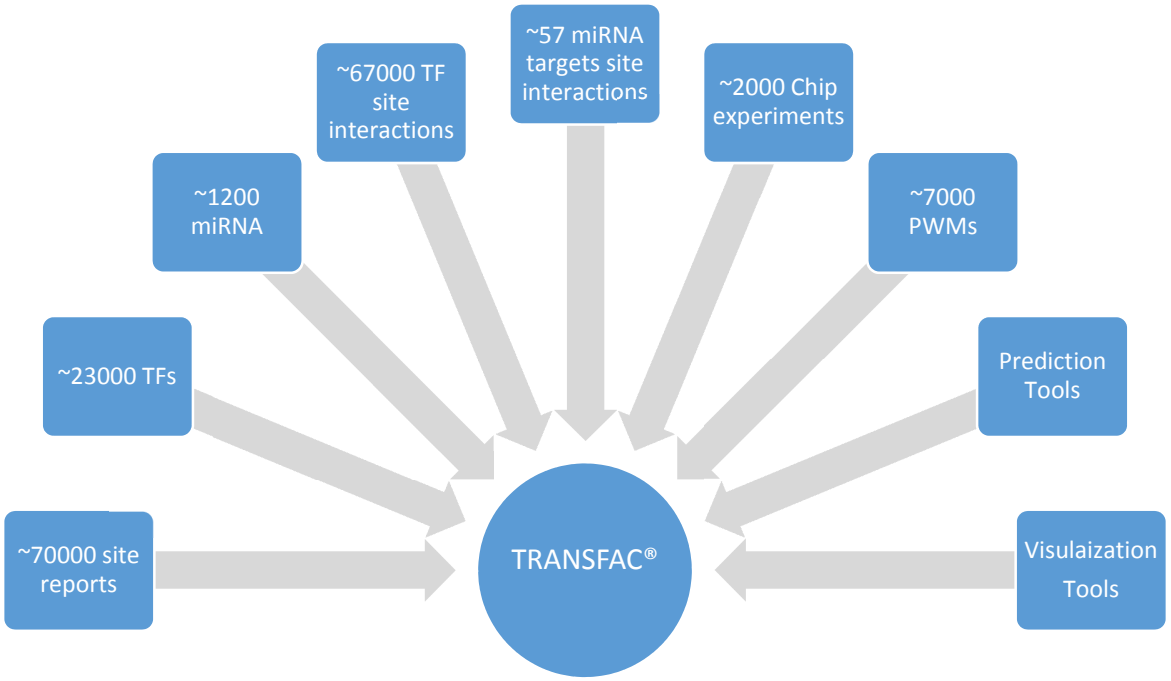

Figure 12. Key Features of TRANSFAC™ (Source http://genexplain.com/transfac/#section2 17.11.2019).

4.4 MATCH™

MATCH™ is a tool for predicting potential binding sites of TFs in a given DNA sequence based on a PWM library [75]. Although other tools have been developed for identifying potential TFBS, the power of this tool stems from its usage of the TRANSFAC® library of matrices, built-in cutoff optimizations, and its utilization of information vectors in its algorithms. Match uses

two scores to evaluate how similar a sequence of nucleotides is to the matrix under search. These scores are:

1. **MSS** or matrix similarity score, calculated using the following formula:

   $$\boldsymbol{MSS} = \frac{Current - \ Min}{Max - Min}$$

   Where:

   - $Current = \sum_{i=1}^{L} I(i) f_{\ i,\, bi}$
   - $f_{i,B}$ *:* frequency of nucleotide B to occur in position i of the matrix.
   - $Min = \sum_{i=1}^{L} I(i) f^{min}{}_{i,\, bi}$
   - $f_{i,B}$ *:* lowest *frequency in position i of the matrix.*
   - $Min = \sum_{i=1}^{L} I(i) f^{max}{}_{i,\, bi}$
   - $f_{i,B}$ *:* highest frequency in position i of the matrix.
   - $I(i) = \sum_{B} f_{i,B} \ln(4 f_{i,B})$, i=1,2,..L

2. **CSS** or core similarity score: is calculated similarly to MSS but using the core positions only ie. the first five most conserved consecutive positions of a matrix.

The lowest score **0.0** corresponds to the state where Current=Max meaning every base in the sequence matches with the base with the lowest score in that position in the matrix.

The highest score **1.0** is obtained in the state where every matching nucleotide matches with the base with the highest frequency in the matrix.

A cutoff for each of following metrics is either automatically set by the program or can be set by the user, and only the matches that have both scores higher than the thresholds are included in the result. Depending on his objective a user can choose one the precalculated cutoffs for each matrix which are:

a) **minFN :** Cut-offs minimizing false negatives.
b) **minFP :** Cut-offs minimizing false positives.
c) **minSum :** Cut-offs minimizing the sum of both false negative and false positive rates.

The tool is available publicly at:

- http://www.gene-regulation.com/pub/programs.html#match
- http://compel.bionet.nsc.ru/Match/Match.html
- http://www.biobase.de : A more advanced version Match™ Professional, allows the user to construct his/her own matrices.

4.5 TFClass

TFClass is a resource that provides a classification of eukaryotic transcription factors based on their DNA-binding domains (DBDs). It comprises six levels:

1. **Superclass**: Based on the similarity in the general DBD topology.
2. **Class:** Based on the similarity in the structure of the DBD.
3. **Family:** Based on the similarity in the sequence and function.
4. **Subfamily:** A subgrouping based on the similarity in the sequence (Optional).
5. **Genus:** The TF gene.
6. **Factor Species:** The TF gene product, in the case of isoforms (optional).

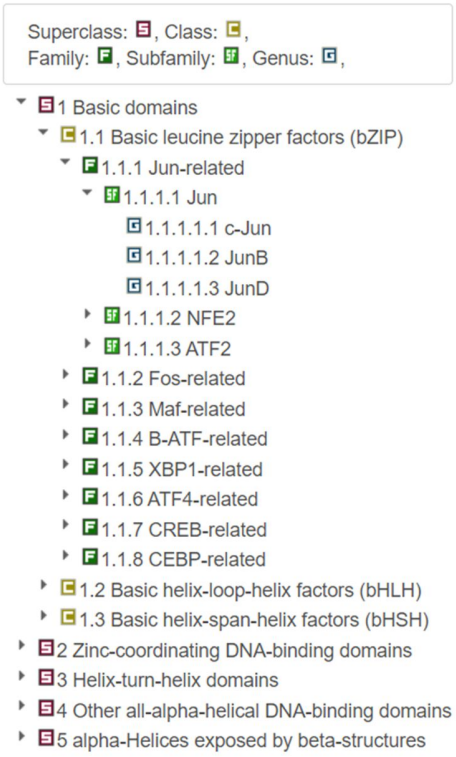

Figure 13. A snapshot from the TFClass web interface showing a part of the hierarchical classification tree (Source : http://tfclass.bioinf.med.uni-goettingen.de , 05.11.2019)

TFClass can be accessed via a web interface where the user can search for a particular TF or label or explore the ontology (Figure 13). Along with with the classification a more detailed view that contains particular information about the item clicked such as the consensus binding sequence, logoPlots of the DNA binding sequence and other specific

information for the different levels. The detailed view contains also links to online resources such as Ensemble and UniProt which can be practical for further exploration and investigation.

TFClass is publicly available at http://tfclass.bioinf.med.uni-goettingen.de/.

4.6 PC-TraFF

Potentially Collaborating Transcription Factor Finder (PC-TraFF), is an algorithm for identifying potentially collaborating TF pairs based on their binding sites distribution and relative distance in the genomic regions of interest [76] . PC-TraFF is implemented in the form of a user-friendly web service that takes a set of genes or alternatively a set of sequences, and a number of parameters such the minimal and maximal distance between to TFBSs, number base pairs upstream from a gene to look for TFBS pairs, and a Z-Score threshold. As a result, the algorithm generates a table of significant TFBS pairs which can be later associated with certain TFs (Table 1). The web service also offers the option of viewing the results as an interactive network that links the collaborating pairs to each other through edges and clusters the network via a Markov clustering algorithm. PC-TraFF is freely accessible at http://pctraff.bioinf.med.uni-goettingen.de/.

Table 1. A sample tabular output from PC-TraFF . Potential collaborating TFBS pairs along with their Z-Score and references that support the prediction

Pairs	**Z-Score**	**Reference**
V$EGR_Q6 - V$SP1_Q2_01	6.19841	BioGRID
V$CETS1P54_01 - V$SP1_Q2_01	5.77081	TRANSCompel
V$MYCMAX_B - V$SP1_Q2_01	5.26869	BioGRID
V$AP1_Q2_01 - V$AP1_Q4_01	4.85265	TRANSCompel, BioGRID
V$CEBP_Q2 - V$STAT6_01	4.71775	TRANSCompel, BioGRID
V$CETS1P54_01 - V$EGR_Q6	4.69386	
V$SP1_Q6 - V$SP1_Q2_01	4.64687	TRANSCompel, BioGRID

V$CETS1P54_01 - V$PEBP_Q6	4.57839	BioGRID
V$SP1_Q4_01 - V$SP1_Q2_01	4.56386	TRANSCompel, BioGRID

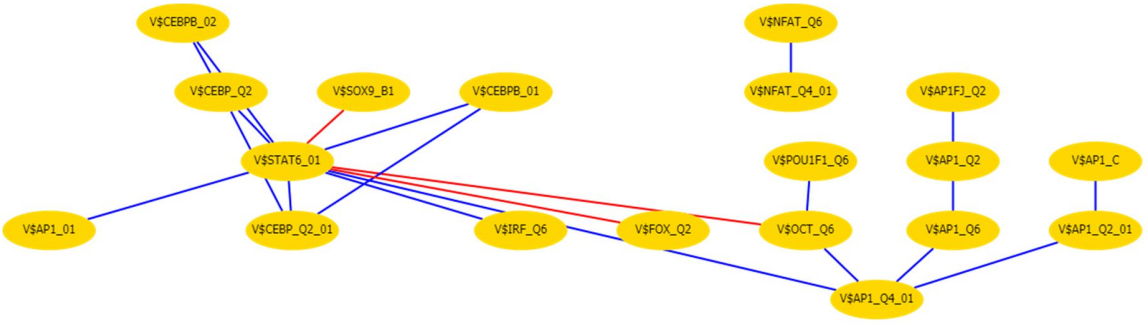

Figure 14. A visual representation of the collaborating TFBS pairs in the form of an interactive network generated by the web-interface of PC-TraFF. The blue edges denote interactions between TFs whose importance is experimentally verified whereas red edges indicate potential interactions between transcription factors that have not been experimentally validated yet. (Source:http://pctraff.bioinf.med.uni-goettingen.de/network/web/, 05.11.2019).

4.7 Network Construction

A library of position weight matrices (PWMs) from TRANSFAC™ [74]version 2013.1 was used in combination with the MATCH™ [75] program to predict binding sites of transcription factors in the conserved promoter regions of the human genome as described in a previous publication [77] (Fig 1D). The promoter region was defined as 1000 base pairs upstream from the RefSeq-defined TSS of the gene. The conservation is based on the alignment between the human, mouse, dog, and cow genomes. As a cutoff, the top 5% of the binding site predictions for each matrix, ranked according to the MATCH score, were considered.

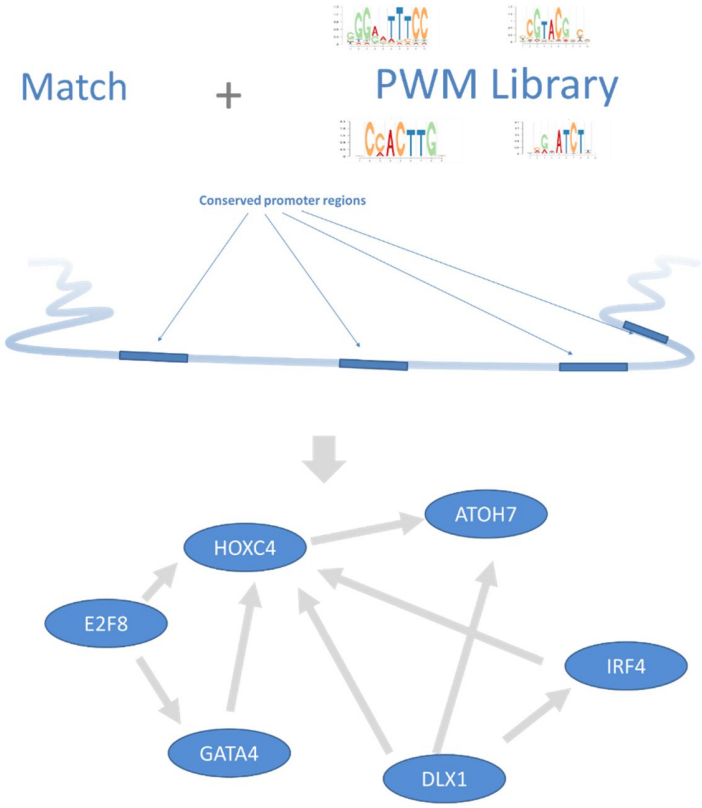

Figure 15. Constructing the regulatory background network. Applying MATCH™, which utilizes a TRANSFAC® PWM library, to conserved promoter regions ended up with a network that summarizes these predictions.

The results were stored in the form of the regulatory network were the TFs and their target genes are represented as nodes, and a directed edge is drawn from the TF to the target gene if the TF has a potential binding site in the promoter of the target. The core network included 829 TFs and their 16354 targets summing up to 749949 interactions. Another expanded for of the network, which contains additional microRNA binding predictions based on mirTrans [78], was constructed and contained 2239 regulators and 20160 targets. This network was computed once and is independent in the process of its derivation from the expression data,

making it usable with every human expression dataset (Fig 1E). The conservation property of these sites makes the prediction ideal for the differentiation context since several pieces of research have shown that conserved regions in the DNA are critical binding sites for development and differentiation [79]–[82].

4.8 Neo4j

Neo4j is an open-source graph database engine that is considered one of the leading engines in its domain[83].

In this project, Neo4j was chosen as an environment to implement the graph database for the following reasons:

- It is an open-source software
- It uses native graph storage optimized for fast traversals.
- It has a simple core java API that can manipulate the finest details in the database
- It has APIs built for several languages including Python PHP, R, Perl, Ruby, NET and many other plugins that are being added with time.
- It provides very clear documentation and active discussion forums.
- It is very scalable (up to billions of nodes/relations), which is very convenient for growing data.
- It has built-in common graph algorithms already implemented and very flexible for direct usage
- It has a built-in visualizer that is customizable, user-friendly and directly linked to the database.
- It has a built-in web interface in which the database can be accessed and queried remotely with security controls.

Through this project, the Neo4j 2.1.2 community version was installed, and the corresponding Neo4j java library was used in the java implementations of the database.

Neo4j is publicly available for download at https://neo4j.com.

4.9 Gene Ontology

Gene Ontology (GO) is a knowledge base that stores information about the biological role of genes and their products on different levels. It is a popular, reliable, widely accepted resource that has been cited thousands of times in the scientific literature[82] [83] .

Its backbone is an ontological structure that stores terms of biological functions terms and defines their relationships with each other. The main types of relationships are:

- "is a": when term X is a term Y, it indicates that X is a subtype of Y.
- "part of": when X is a part of Y, it indicates that the presence of the X implies the presence of Y, but the opposite is not necessarily true.
- "has part": when X has Y as a part, it means that X necessarily has part Y. If X exists, Y will always exist; however, if Y exists, we cannot say for certain that X exists.
- "regulates": Indicates the ability of a specific process or term to regulate another.

The second component is the GO annotations. Through these annotations, genes and their products are associated with terms in the ontology. This summarizes the different known biological roles of a gene in different contexts. The annotations are based on evidence from literature and different curated databases.

To evaluate the relevance of the gene sets at each stage, a GO enrichment analysis using the biological processes and a Fisher's Exact test on each column in these cascades was applied using one set at a time as an input. Terms that had a p-value less than 0.05 after the Bonferroni correction were sorted by their fold enrichment and the top terms were examined (Table 2). These terms were evaluated based on their consistency with the stage under observation at that time point.

Table 2. An example of a Gene Ontology analysis result table. The terms are sorted here by fold enrichment and all have a p-value less than 0.05.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrichment	+/ -	P - value
ventricular cardiac muscle cell differentiation	18	3	.01	> 100	+	9.20E-04
pharyngeal system development	26	3	.01	> 100	+	2.52E-03
cell fate determination	42	3	.02	> 100	+	9.76E-03
cardiac ventricle morphogenesis	72	3	.03	87.48	+	4.61E-02
cardiocyte differentiation	116	4	.06	72.40	+	1.85E-03
cardiac muscle tissue development	159	5	.08	66.03	+	5.96E-05
cellular response to steroid hormone stimulus	187	4	.09	44.91	+	1.19E-02
mesenchyme development	216	4	.10	38.88	+	2.09E-02
striated muscle tissue development	285	5	.14	36.84	+	1.03E-03
muscle tissue development	298	5	.14	35.23	+	1.28E-03
heart morphogenesis	249	4	.12	33.73	+	3.65E-02
regulation of animal organ morphogenesis	256	4	.12	32.81	+	4.06E-02
gland development	417	5	.20	25.18	+	6.63E-03
heart development	528	6	.25	23.86	+	4.52E-04
chordate embryonic development	640	7	.30	22.96	+	2.52E-05
tissue morphogenesis	560	6	.27	22.50	+	6.39E-04

4.10 Cytoscape.js

Cytoscape.js is an open-source JavaScript library that specializes in graph analysis and network visualization. It is the evolved version of an older library called Cytoscape Web,

which was created at the Donnelly Centre at the University of Toronto within the wider framework of Cytoscape [84][85].

Cytoscape.js allows the display and manipulation of graphs in desktop and mobile browsers. This library allows the user to interact with the graph and allows the client to hook into user events and includes basic to advanced features such as pinch-to-zoom, box selection, panning, and others. It also contains a collection of different layouts which is continuously growing and adapted for different uses and graph types.

Aside from visualization, this library has various built-in graph analysis algorithms and functions such as the Breadthfirst algorithm, which can be easily used and built on for the development of more complex algorithms.

We used Cytoscape.js as a bedrock upon which we extended and built the JavaScript visualizations in our web tool. New layouts, animations and features were built to suit our purposes and can be reused and extended in the future to branch into more complex web-based network visualizations.

4.11 Data

4.11.1 The Heart Development Dataset

The main project I worked on during my PhD period involved a collaboration with the Institute of Pharmacology and Toxicology in the University Medical Center Göttingen. The project revolved around examining the process of differentiation of cardiomyocytes from human pluripotent stem cells (hPSCs) on the genetic level.

The project was based on an experimental model that simulates human heart muscle development the generation of bioengineered heart muscle (BHM) directly from hPSCs. This experiment is the first detailed characterization of a novel cardiac organoid model, generated by a single step tissue engineering approach directly from hPSCs, with organotypic contractile functionality.

The experiment revealed that BHMs indeed traverse through defined in-utero like developmental stages with characteristic transcriptome profiles, are composed of mainly mesodermal cells (cardiomyocytes and fibroblast-like cells), display continuous functional maturation over time with enhanced contractile performance on the cellular level and develop in late cultures (day 60) a functional neural crest component with resemblance to the cardiac sympathetic nervous system. A detailed characterization on the molecular, cellular and functional level of this model is described in the dissertation of F. Raad PhD, in the following we present a brief summary of the underlying experimental procedure.

HESC cultures of HES2 cell line were maintained on IR-HFF layers in HES medium. The HESCs were then cultured in basal medium and different factors were added to the medium as indicated in Figure 16.

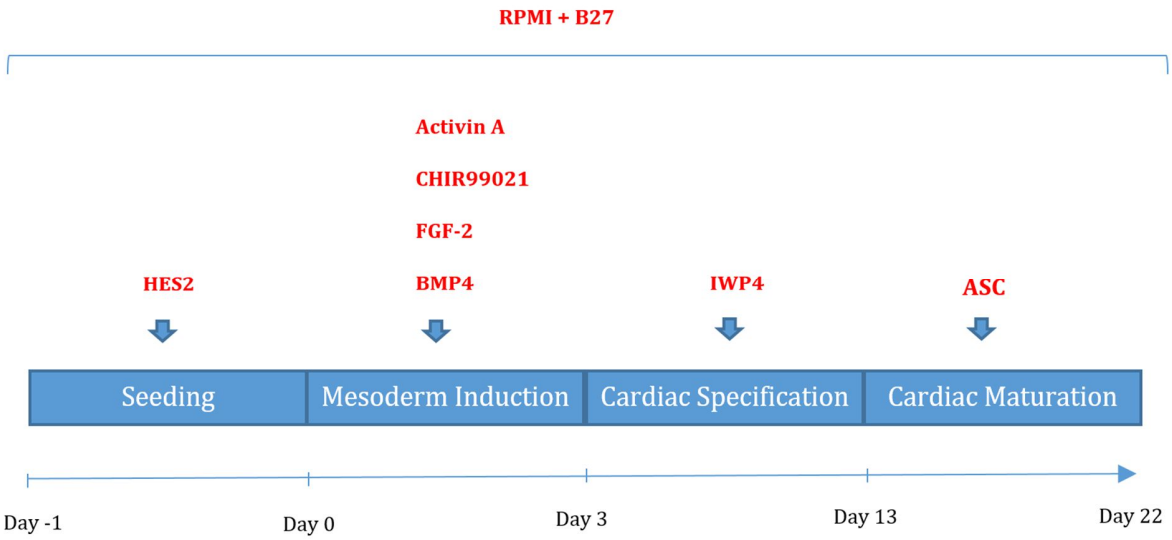

Figure 16. Different agents added at different time point through the experiment.

At day -1 cells were passaged and plated. 24 hours later, cells were washed with basal medium before the start of the mesoderm induction phase (3 days); following growth factors were added to the basal medium for mesoderm induction: CHIR99021, FGF-2, BMP4, Activin-A. Culture medium was exchanged daily until day 3 .

Subsequently, the cardiac specification was induced for 10 days with the addition of an inhibitor of Wnt production IWP4.

Thereafter, cardiomyocytes were maintained in basal medium with medium exchanges every 2 days . The beating was observed starting day 13, and the cells were harvested on day 22. The cells were afterwards rinsed and collected.

The experiment was followed with a quantification using RNA-Seq which produced transcript counts of around 20,000 genes across eight different time point with four replicates each. The time points used were denoted as the following days: D_1 (control) , D0, D3, D8, D13, D22, D29, D60. The data also contained RNA-Seq counts for two time points in 2D culture, however they were mostly irrelevant to the analysis which focused on the more complete 3D culture time points. The data was normalized using the FPKM method and the

resulting normalized set was used for the analysis. This dataset is referred to as the heart development dataset throughout this manuscript.

4.11.2 Other Sources

The first secondary dataset, Dataset 2, was derived from the normalized expression datasets from the previously published study by Qing Liu et al [88], publicly available in the GEO repository under the accession number GSE85332. The dataset covered the RNA-Seq profiling of the differentiation of four different stem cell lines into cardiomyocytes, 2 hiPSC lines (C15 and C20) and 2 hESC lines (H1 and H9). The gene expression for the genes in each cell line was taken at four stages: pluripotent stem cells (day 0), mesoderm (day 2), cardiac mesoderm (day 4), and differentiated cardiomyocytes (day 30). The assembled and formatted data can be found in S2 Dataset.

Dataset3 was assembled using public RNA-Seq data that is captured during the differentiation of H1 derived human neuronal precursor cells (NPCs) across the days 0,1,2,4,5,11, and 18 after induction of neuronal differentiation. Publicly available DEGs and GO enrichment analysis on the same dataset was used for comparison. The dataset and the analysis results could be found in the expression Atlas under the accession E-GEOD-56785. The assembled and formatted data can be found in S1 Dataset.

Dataset4 was based on the temporal profiling of HIV SupT1 CD4+ T cells detailed in a manuscript published by Golumbeanu et al. [89]. The cells were either Mock-infected or infected with an HIV-GFP-based vector. The expression of the genes was profiled at five time points of the viral replication process. The data is publicly available from NCBI's Gene Expression Omnibus under the accession number GSE100587.

5 Results

This section represents the various results of my PhD work, from the methods developed, to their implementation, their application to real life datasets, and the biological conclusions and implications of the outputs.

In the first subsection, I describe the work I have done on enhancing the existing regulatory network previously published by M. Haubrock [90], by adding additional relevant sources, which resulted in different networks that can be used according to the context. Furthermore, in this subsection, I describe the strategy I used to store the resulting networks in a fitting database structure to facilitate its usage in the next parts of the work.

The central and novel part of the results comes in the following part of the results. In the second subsection, I introduce the concept of the TRC model, the computational and mathematical model behind, and the detailed steps of the TRC workflow.

Afterwards, in the third subsection, I introduce the web service, which I developed throughout my doctoral studies, which is an implementation of the TRC workflow as well as other workflows that cover aspects such as co-expression and co-regulation.

Subsequently, the application of the theoretical concepts and the tools developed on different experimental datasets will be described. The detailed, in-depth analysis of the heart development dataset is covered in the fourth subsection using the TRC model and the accompanying workflows.

The final subsections cover the analysis of other datasets, however, in varying details depending on the dataset. The work is entirely reproducible using the webserver, the mentioned resources and tools, and the datasets, which will be attached as supplementary files.

5.1 Background Regulatory Network

Different contexts require different background regulatory networks depending on the questions the scientist is trying to answer. For that purpose, I attempted to enhance and expand the regulatory network developed by M.Haubrock [90], adapting it to be suitable as a backbone for the different workflows and tools I constructed.

As the binding site predictions are the backbone of the regulatory network, the match scores of these predictions can be a good measure of the quality of the network. For each PWM, these predictions are ranked based on the match score, and the top *n* predictions are considered to derive what Haubrock refers to as an *n*-prf network. The 1-prf network is the smallest and is based only on predictions that belong to the highest top 1 percent, and the 100-prf network is the largest, based on all the predictions that satisfy certain thresholds and cutoffs that are detailed in Haubrock's work. In the work that follows, whenever not specified, we refer by default to the 5-prf regulatory network as a background regulatory network.

5.1.1 Network Enhancement

We wanted to integrate additional resources to the network to enhance the information content, check for experimentally verified regulatory predictions, and evaluate the overall predictive quality of the background regulatory network.

As a first step, we attempted to integrate all the available ChIP-seq data resulting from experiments done on human cells in different tissues and contexts, which are stored in the ENCODE project [2][3]. We chose to integrate the data into the 100-prf network, thus covering all the possible predictions, allowing us additionally to study whether a higher match score corresponds to a higher probability of having experimental validation. For this integration to happen, we had to align all the predicted binding sites with all the ChIP-seq fragments from all the experiments. For each binding site prediction corresponding to a PWM, we needed to search for all the ChIP-seq fragments where a TF that is associated with this PWM binds in the region where the site lies Figure 17. This task proved to be very expensive in terms of computational time, so we had to develop a system that facilitates this kind of overlap and produces the desired results in an acceptable time.

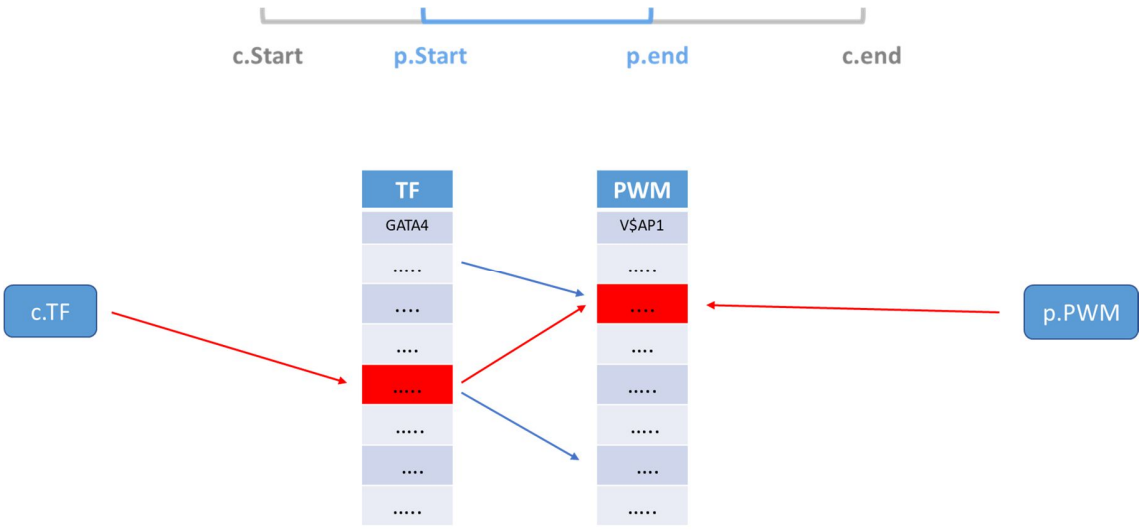

Figure 17. The prediction-ChIP overlap criteria. If a binding site prediction p overlaps with a ChIP-seq segment c, where p.start>c.start and p.end<c.end, and the TF in that ChIP-seq experiment is associated with the PWM upon which this prediction is based, then this prediction is considered to be experimentally validated.

For that purpose, we developed an overlap system we call the Parallel Tables System (PTS). The PTS takes advantage of the fact that the predictions and ChIP-seq fragments are both based on the same version of the human genome, hg19. That means if we divide the genome into smaller regions, we only need to compare the predictions in this region with the ChIP-seq fragments in the same corresponding region, rather than comparing them with every ChIP-seq fragment Figure 18. This system was to speed up overlapping between predictions and experimental data by spreading the data on several smaller tables. This reduced the complexity significantly and allowed the overlap to happen in a feasible time frame. The criterion of spreading these tables was to be chosen carefully to minimize as possible the number of comparisons and mismatches. These tables should fit into the main memory easily and at the same time, be independent of the rest of the tables in the group so that during the joining process, each table is loaded and used once. The PTS system can be proven mathematically to cover all the overlaps needed without any information loss. The complexity of such a system is far lower due to the reduced number of comparisons to be made compared to joining the table that contains all the predictions with one that has all the ChIP-seq.

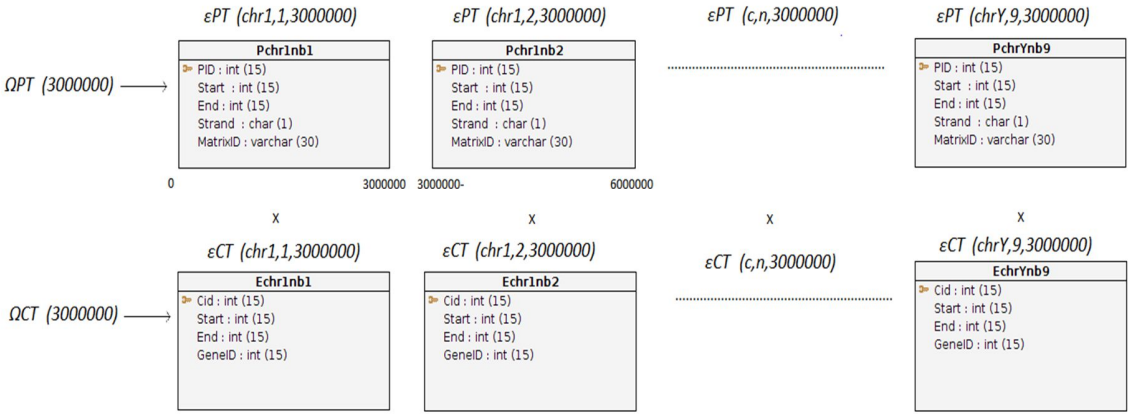

Figure 18. Spreading the ChIP-seq and the predicted binding sites tables using the PTS. The genome is divided into slightly overlapping regions of 3000000 bp, and a table is dedicated to each region for both the ChIP-seq and the predictions that fall into that region. Each prediction table is joined with the corresponding table from the ChIP-seq tables. The join results are then collected and summarized in one result. In this particular configuration, the PTS was 1050 times faster than a regular table join.

Not only that, but we found that the same system can be reused for solving any similar genomic overlap problem, such as overlapping ATAC-Seq, DNase-Seq, and other data sources that are based on genomic locations of DNA spans.

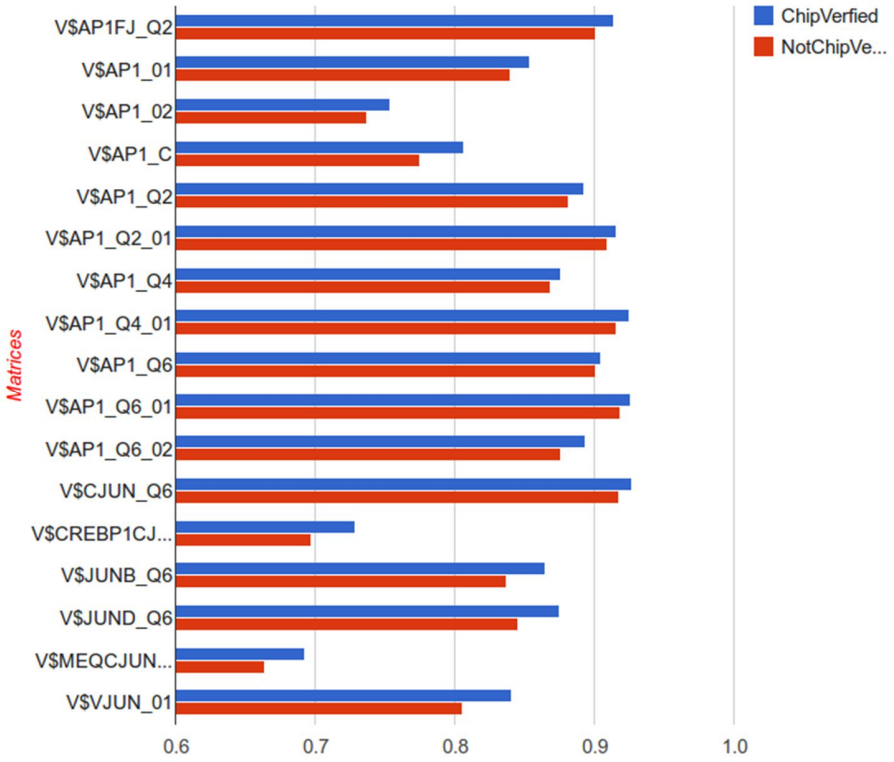

Figure 19. A comparison between the average scores of the ChIP-verified predictions and the non-ChIP-verified prediction for some of the matrices in the study. The ChIP verified predictions turned out to have a higher match score on average for most of the matrices, confirming the relevance of the match score in the likelihood of the prediction to be true.

After overlapping the ChIP-seq experiments with the predictions, we performed a study that compared the average match score of the binding sites with experimental validation vs. those that had no ChIP-seq experiment supporting them. This was done for each of the PWMs separately, studying the set of the binding sites associated with each PWM. The associated binding sites set for the PWM is separated into experimentally verified, those which overlapped with a ChIP-seq segment of an experiment that used a TF associated with this PWM, and those which did not overlap with any. For 98% of the PWMs, the experimentally verified binding sites set had a higher average match score than the non-experimentally verified Figure 19. This showed the importance of the match score in determining the likelihood that the binding site prediction is used in a biological context, and gave us a reason to use the regulatory networks with lower prfs, such as the 1-prf and the 5-prf networks.

5.1.2 Network Storage

At first, binding site predictions or regulatory interactions were stored in an SQL database. However, the traditional relational database model is not well suited to store highly connected entities such as regulatory interactions, which are, in their essence, more of a network of interactions than a table. Queries on such interactions were complex and were

based on graph theory, and algorithms as simple as finding the shortest regulatory path from one gene to the other were computationally expensive in such a tabular database structure. For that reason, we needed to upgrade the storage system from an SQL database to a NOSQL graph database. Neo4j was the engine and database system of choice for the many reasons mentioned in the materials and methods. Thus I designed two simple graph database structures, one designed to contain detailed information, mapping TF gene names to PWMs to binding sites in target genes, and the other summarizes all the information from the first in the form of regulatory interactions and integrates a summary the ChIP-seq experiments associated with each regulatory interaction. The two databases were used for different questions, depending on the level of detail and the type of query. The second database, which was structured as a regulatory network with direct regulatory interactions, provided the backbone for most of the work that follows.

The first database was designed using three different kinds of nodes (Figure 20):

1. **Gene:** a node where a gene is represented, and information such as its official symbol and name are stored as properties of the node.
2. **Matrix:** a node where a matrix is represented and information such as its name and the associated PWM.
3. **Binding Site:** a node where a binding site is stored and information such as the nucleotide sequence of the site, the strand, and chromosome on which this site lays, its coordinates, and a boolean value that tells whether this site is verified through the ChIP-seq overlap previously described.

These nodes were connected by three types of relationships:

1. **UsesMotif:** connects a gene node with a matrix node associating a regulatory gene with the binding motifs that could be used to determine its binding site.
2. **Matches:** connects a matrix node with a binding site node associating a motif with the binding site that was found based on its PWM.
3. **InPromoter:** connects a binding site node with a gene node associating the binding site that was found in a promoter region with the gene associated with that promoter.

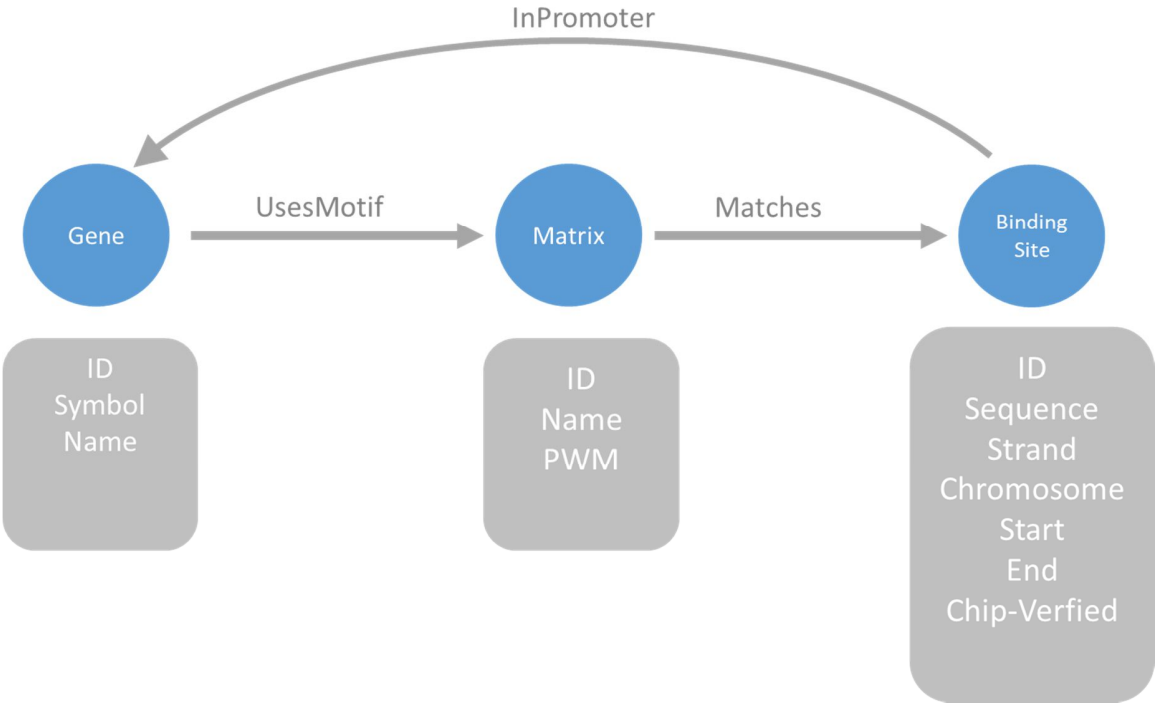

Figure 20. The graph database schema used for the first database. Three types of nodes and three types of relationships were used to store the information previously stored in an SQL database, improving the performance and allowing more complex graph queries to be ran efficiently.

The second database was designed using only one kind of nodes (Figure 21):

1. **Gene:** Similar to the previous database, it contains basic information such as the official symbol and the name of the gene.

The second database contained one relationship type. However, this relationship contained properties that summarized some of the information contained in the previous database:

1. **Regulates:** Connects two gene nodes associating a gene, a regulatory one, with a target gene. If a gene has at least one directed connection in the first database with three hops: UsesMotif then Matches, then InPromoter to another gene to this relationship is represented in the second database via this single relationship. In other words, if a regulatory gene has a potential binding site in the promoter of another gene, then it is considered potentially regulating it, and they are connected by an edge in the second database. Four properties were stored in the relationship:
    - ChipNum: The number of ChIP-verified binding sites supporting this regulatory relationship.
    - ChipScore: The average score of the ChIP-verified binding sites.

- NonChipNum: The number of binding sites related to this regulatory relationship that did not overlap with any ChIP-seq fragment.
- NonChipScore: The average score of the non-ChIP-verified binding sites.

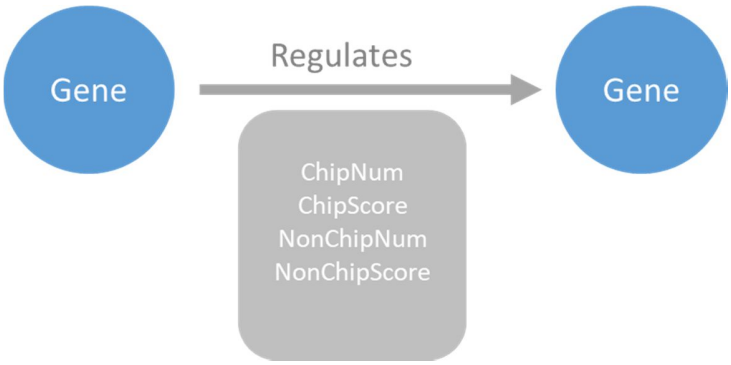

Figure 21. The graph database schema used for the second database. It uses only one type of node and one type of relationship, summarizing the essence of the information in the previous database. It allows the user to query regulatory interactions between genes and filter them based on experimental evidence.

These databases can be accessed and used via the built-in Neo4j interface using the Cypher query language. For the integration of these databases in the workflows, a Neo4j Java API that was provided by the Neo4j community was used (Figure 22).

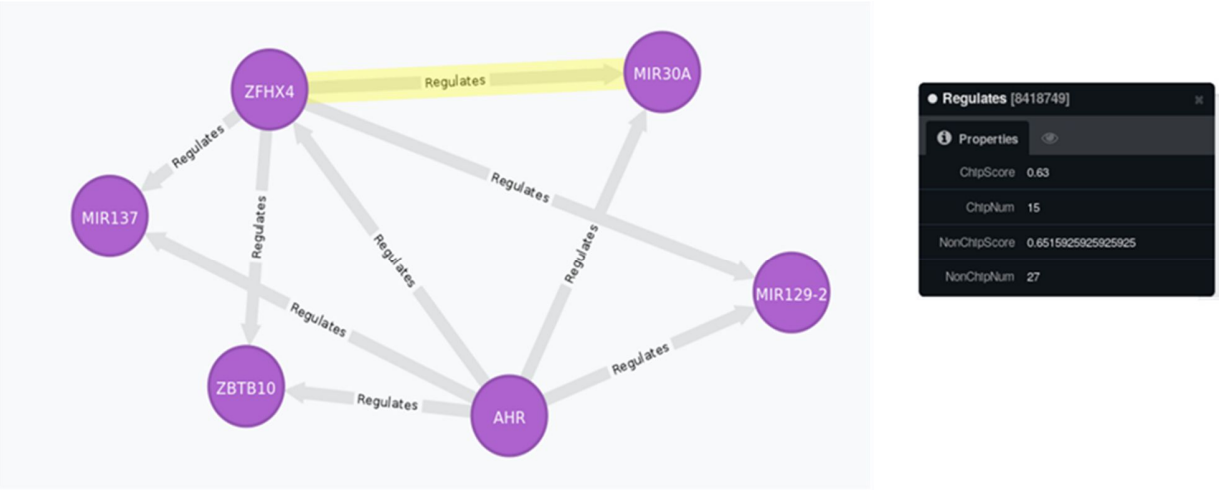

Figure 22. A snapshot of the Neo4j local web interface displaying a visual result of a Cypher query. Regulatory interactions between different genes are shown as well as the information contained in one of the "Regulates" relationships between two genes. The properties of the relationship indicate that there are 15 ChIP verified binding site predictions that support this interaction and 27 non-verified ones.

5.2 Temporal Regulatory Cascades

The Temporal Regulatory Cascade (TRC) model is a series of connected stage-specific regulatory networks that form a cascade-like architecture incorporating temporal order and gene expression as well as conserved TF binding sites information. The TRC is constructed via a workflow that has two main elements (Figure 23):

1. The background regulatory network, which was described previously, which provides an independent source of regulatory interactions and gives rise to the edges in the TRC.
2. The temporal gene expression data, which varies depending on the experiment, and provides the basis for identifying stage-specific regulatory and non-regulatory genes by detecting single peaks in their expression pattern. These stage-specific genes, typically regulatory genes, are represented as nodes in the TRC.

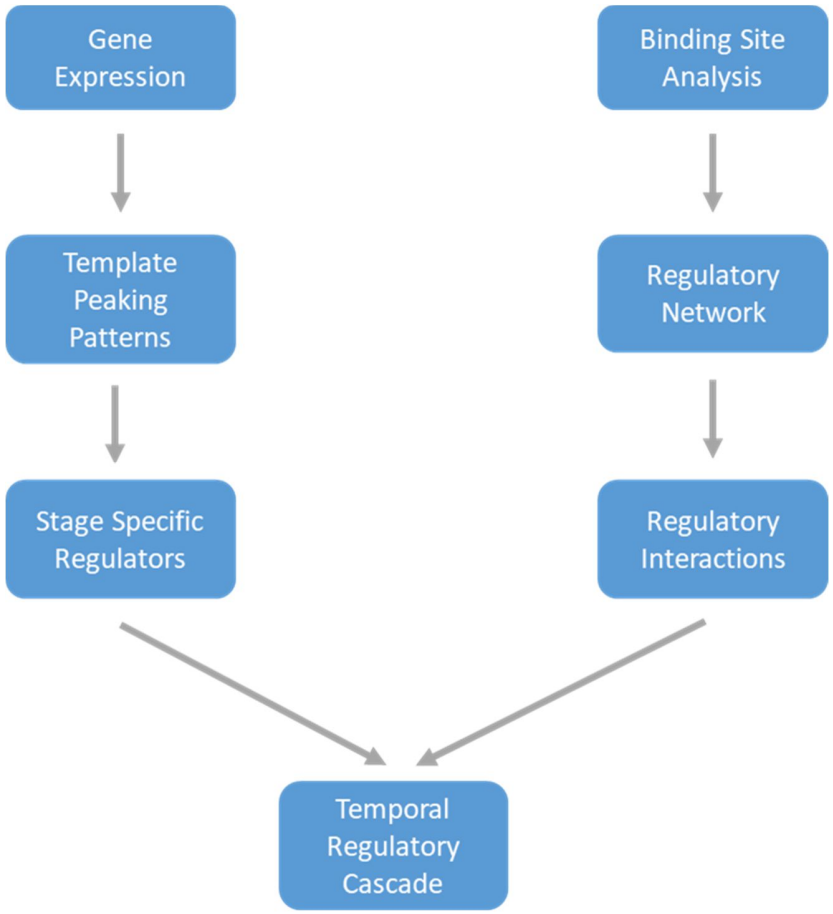

Figure 23. The TRC workflow in a nutshell. Combining gene expression and binding site information to construct a temporal regulatory cascade.

For simplicity purposes in explaining the methodologies that follow, we assume that the experiment is neither over or under-sampled thus the terms “stage” and “time point” might be used interchangeably, however it is worth defining the two terms as follows:

- **Stage:** A stage where distinct biological events happen in the experimental context under study. A stage could be represented by one or multiple time points in the data set, or not represented at all in case no gene expression snapshot was taken through it.

- **Time point:** A point represented typically by a column of expression values taken at a particular time in the course of the experiment. A time point might represent a biological stage or not, depending on the experimental design.

5.2.1 Template Peak Patterns

The TRC method utilizes the concept of constructing artificial template patterns of interest and attracting genes that behave similarly to these patterns using correlation. The template patterns used were stage-specific patterns, peaking at one time point each, and denoted by template peak patterns (TPPs). While different kinds of template patterns can be used, we chose the single peak TPPs as a default for its ability to attract stage-specific regulators that are unique to each time point.

$Let\ T = \{\ t_0, \dots,\ t_m\ \}$ denote the set of time points in an experiment and $G = \{\ g_0, \dots,\ g_n\ \}$ denote the set of genes under study.

$Let\ D = \begin{bmatrix} d_{00} & \cdots & d_{m0} \\ \vdots & \ddots & \vdots \\ d_{n0} & \cdots & d_{nm} \end{bmatrix}$ denote the expression data matrix where d_{ij} is the expression of g_i at time point *j.*

$Let\ R(N, E)$ denote the regulatory network graph. With N denoting the set of nodes and E denoting the set of edges in the network.

Definition 1: $L = \{ A_0, \ \ldots, A_m \}$ a library of artificial template peaking gene expression patterns where:

$$A_j = \{ a_0 \ldots a_m : \ a_i \ = 0 \quad for \quad i \neq j \quad and \quad a_i = 100 \quad for \, i = j \}$$

A Template Peak Pattern or ***TPP*** (*Denoted by* $\mathbf{A_j}$ in the previous defenition) associated to time point t (Figure 24) is an expression pattern constructed such that the expression is:

- 100 at time point t.
- 0 at every other time point.

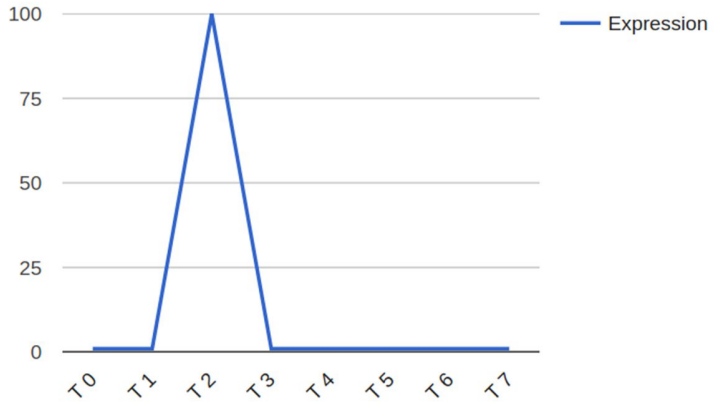

Figure 24. The TPP associated with time point T2. The pattern shows a maximal expression(100%) at T2 and 0 at the other time points

For simplicity purposes in explaining the model and the workflow, we are assuming the data has one replicate per time point. However, when replicates exist for each time point in real datasets, we apply for each replicate what applies for the time point. For example, In the case where the time point has several replicates associated with it, as it often is, the peak spans through all the replicates. Becoming as follows:

An Artificial Peak Pattern or ***TPP*** associated to time point *t* (Figure 25) is an expression pattern constructed such that the expression is:

- 100 at every replicate of time point *t*.
- 0 at every replicate of every other time point

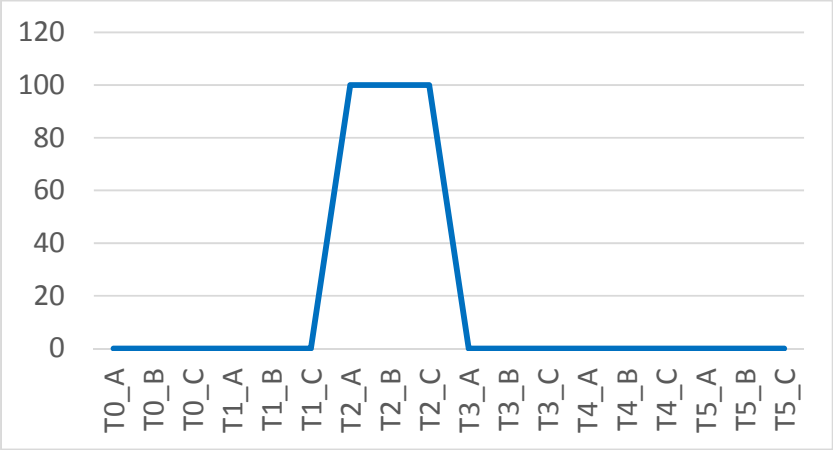

Figure 25. The TPP of T2 with multiple replicates per time point. The same principle as the normal TPP, but the maximal expression spans through each replicate of the time point.

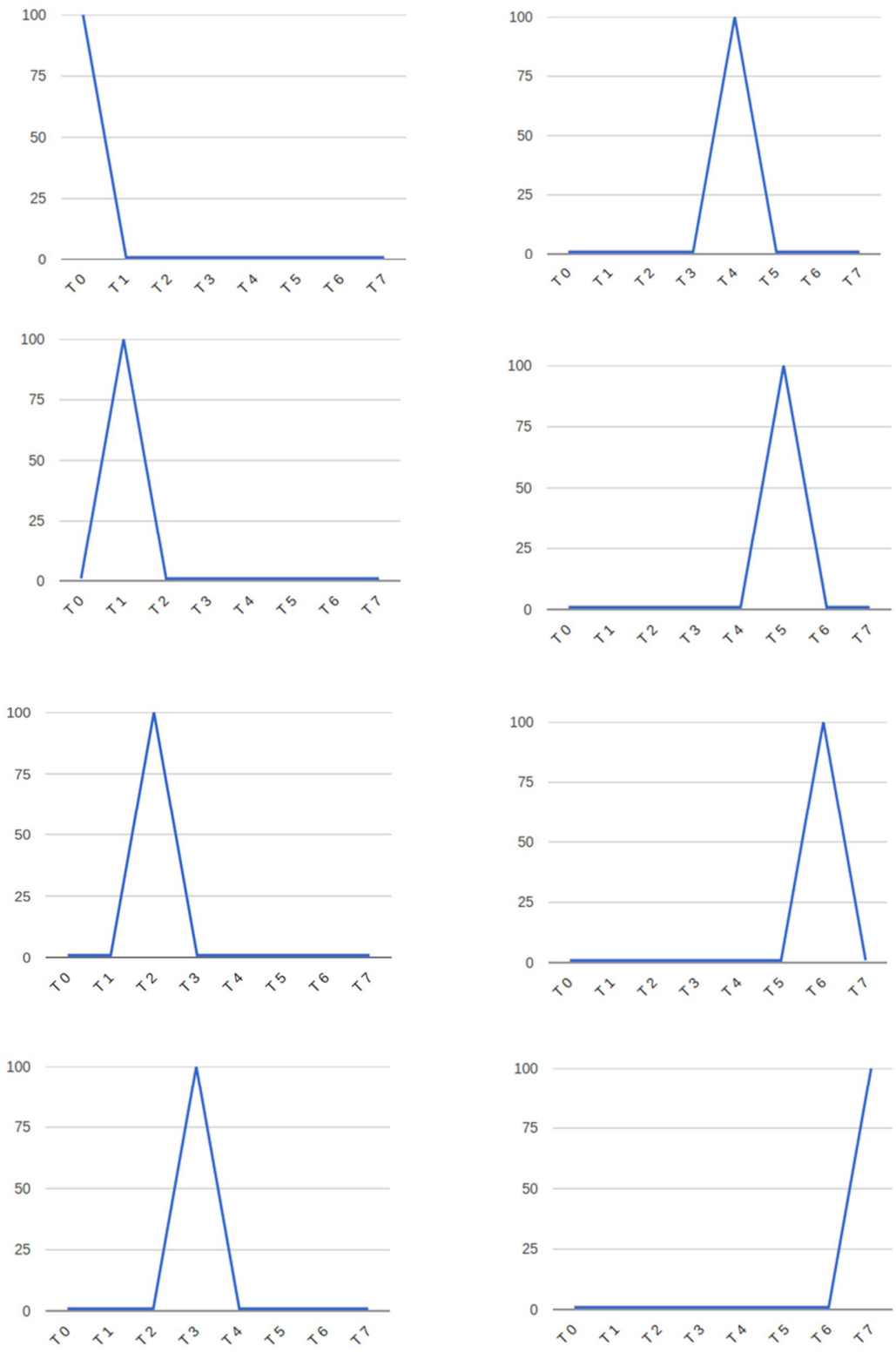

Figure 26. A library of TPPs (Denoted by ***L*** *in the previous definition) based on an 8-Time point dataset. One TPP for each time point in the data set.*

5.2.2 Identifying Stage-Specific Regulators

After a TPP library is constructed, a set of stage-specific regulators is constructed for each time point based on the TPP associated with that time point.

Definition 2: $P = \{ P_0, \ \ldots, P_n \}$ a sequence of gene sets where:

$$P_j = \{ g : \ \exists i \ \mid \ g[i] \in G \ \ and \ \ cor(D[i], L[j]) > threshold \}$$

The correlation between each TPP and the expression pattern of every regulatory gene the gene set is calculated. Genes that have a high correlation above a certain threshold, with the TPP of a time point, form the initial stage-specific regulatory set associated with that time point (Figure 27). Although it is not obligatory to use regulatory genes, it is a default to restrict the TRC analysis to regulatory genes as they provide a good starting point for different analysis, contribute to a larger set of regulatory interactions, and keep the TRC concise as the number of the non-regulatory genes is at least 20 fold larger. Unless otherwise specified, the TRC corresponds to a regulatory-genes-only TRC. Each stage-specific regulatory set is later trimmed based on size parameter maxS, and only the highest correlated genes to the TPP are taken in the TRC.

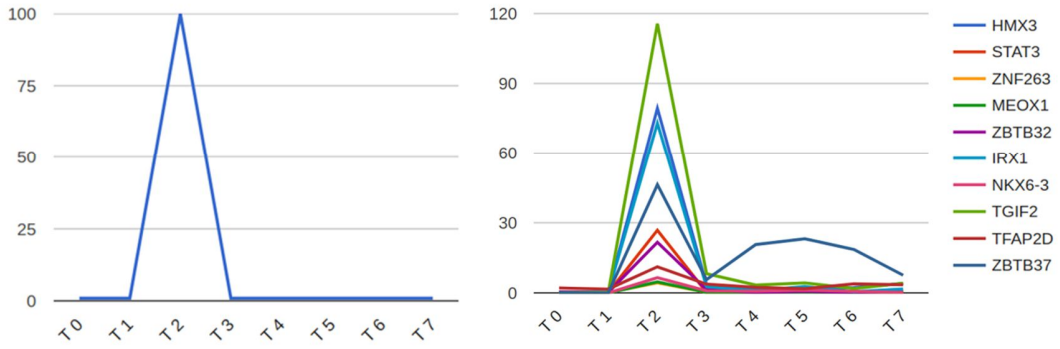

Figure 27. (Left) The TPP of T2. (Right) The top 10 correlated regulators to that TPP and their expression patterns. One can clearly observe the peaks in their expression at T2.

5.2.3 Mapping Regulatory Interactions

After the stage-specific regulators are identified, the regulatory interactions between those regulators are queried. However, in order to take the factor of time and temporal order, only regulatory interactions between regulators of the same stage, or the interactions between regulators of a stage and the next time point are queried. As it does not fit into the biological logic, for example, for a TF that is active only in the beginning to directly activate another gene that is active in the end, or for a TF that is active only in the end to have activated another that was present only in the beginning.

A regulatory mapping between the two sets is the set of edges that are present in the regulatory network and have a source gene belonging to the first set and a target gene belonging to the second set. Note that this mapping is a directed one where the direction of edges goes from the first set to the second.

Definition 3: $M(X,Y) = \{\ e(x,y) \in R\ \ : x\ \in X\ \ and\ y\ \in Y\}$

For each time point, a regulatory mapping is performed for the corresponding stage regulators' set against itself. This generates the regulatory interactions between the regulators of the stage; we refer to them as intra-regulatory edges.

Definition 4: $Eintra = \{E_0,\ \ldots,\ E_{m-1}\}$ where $Eintra_j = M(P_j\ ,\ P_j\ )$

For each time point, except the last time point, a regulatory mapping is performed for the corresponding stage regulators' set against the regulators' set of the next stage. This generates the regulatory interactions where the regulators of a stage have the potential to activate the upcoming regulators of the following one; we refer to them as inter-regulatory edges.

Definition 5: $Einter = \{E_0,\ \ldots,\ E_{m-2}\}$ where $Einter_j = M(P_j\ ,\ P_{j+1}\ )$

The cascade is the combination of the peaking regulators' sets, intra-regulatory edges, and inter-regulatory edges.

Definition 6: $C(P,E) = \{C_0,\ \ldots,\ C_m\}$ *whe*re: $C_j\ = \{P_j\ ,\ Eintra_j\ U\ Einter_j\ \}$

5.2.4 Parameters

To adjust the temporal regulatory cascade, we use three primary parameters:

- **minE:** The minimum required expression level of a gene. This represents a threshold that filters out genes that are lowly expressed despite their peaking patterns. A gene that does not have an expression level higher than this threshold in any of the replicates or time points is eliminated and omitted from the calculation that leads to the TRC.

- **minC:** The minimum required correlation between a gene and a TPP to qualify as a peaking gene. This cutoff threshold eliminates every gene that is not correlated to the TPP related to that stage with a high enough correlation. Genes that make it above this threshold and the previous one, provide the initial set of regulators associated with each stage.

- **maxS:** The maximum size of stage-specific regulator's set, which is the maximum number of genes that can be associated with a specific time point. The initial regulators associated with a time point based on minC are sorted by their correlation to the TPP of that stage, and the top n (maxS) regulators are picked to be in the column associated with the stage. If the initial regulators set of a stage has fewer genes, then the whole set is taken. The max number of nodes in the cascade is maxS multiplied by the number of time points.

5.2.5 Algorithm

Input:

- *Array of genes G[n]*
- *Array of time points T[m]*
- *Gene expression matrix D[n][m]*
- *RegulatoryNetwork*
- *minC*
- *minE*
- *maxS*
- *OnlyRegulators (Boolean)*

```
If ( onlyRegulators )
  FilterOutNonRegulatoryGenes (D, G, RegulatoryNetwork)

FilterOutLowExpressedGenes (D, G, minE)

For t in (0 … m):
      TPP [t] = array [m] (0)
     TPP [t][t] = 100

For t in (0 … m):
      intialSet=set()
     For i in (0 … n):
         If ( correlation( TPP[t] , D[i] ) > minC )
            intialSet.add(object(G[i], correlation( TPP[t] , D[i] ))
    RankByCorrelation(intialSet)
   P[t]= top(intialSet, maxS)

For t in (0 … m):
      Find all regulatory interactions between the genes in P[t]
      Find all regulatory interaction between the Genes in P[t] and the Genes in  P[t+1]
```

For t in (0 … m):

Display P[t] in a column of nodes

Display Regulatory interactions as connecting edges

Shift one step to the right

5.2.6 Relevant Metrics and Definitions

The following simple metrics could be used to detect the relevant time points in the dataset and to evaluate different cascades:

- **PS:** The peak strength of a gene is the Pearson correlation of its expression pattern to the TPP of the stage *t* to which it is associated.

$$PS_g = cor(D_g, A_t)$$

- **PSS:** The peak strength of a stage *t* is the average of the PSs of the regulators in that stage. Stages with a higher PSS contain regulators with a stronger stage-specific pattern.

$$PSS_t = \frac{\sum_{g=0}^{g=n} PS_g}{n}$$

- **PSC:** Peak strength of the cascade is the average of peak strengths of the stages in a cascade. This metric could be used to evaluate the whole cascade, especially when different versions of this cascade are calculated using different parameters. The parameter set that leads to the highest PSC is considered as the optimal set.

$$PSC = \frac{\sum_{t=0}^{t=m} PSS_t}{m}$$

- **Intra-Regulatory network:** The subset of the cascade composed of a set of stage-specific genes and the edges between them. In the TRC, it corresponds to the nodes in one column and the vertical edges.

- **Inter-Regulatory network:** The subset of the cascade composed of a set of stage-specific genes and the edges outgoing from them to the gene set of the next stage. In the TRC, it corresponds to the nodes in two consecutive columns and the cross edges between them.

- **Intra-RI:** Intra-regulatory influence, the number of outgoing edges from a node in a column to other nodes in the same column.

- **Inter-RI:** Inter-regulatory influence, the number of outgoing edges from a node in a column to other nodes in the next column.

- **TV:** Target value, the number of incoming edges to a node.

- **RV:** Regulator value, the number of outgoing edges from a node.

5.2.7 Visual Representation

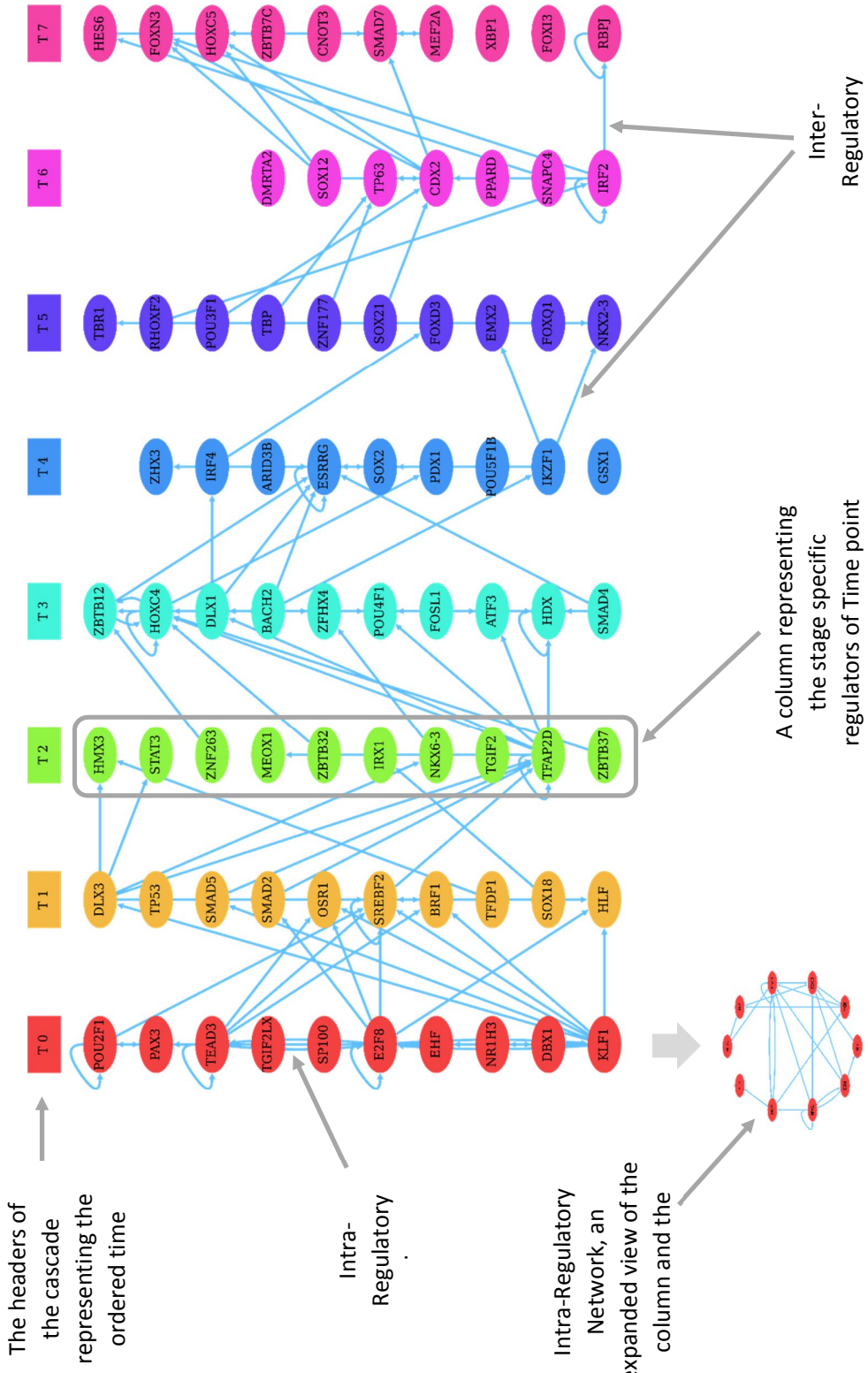

Figure 28. The basic architecture of the TRC. The different elements and lables that compose the TRC.

5.3 Web Tool

One of the main goals was to develop a tool that implements the TRC workflow as well as other workflows that utilize the power of the background regulatory network to investigate aspects of co-expression and co-regulation in a gene expression dataset. The decision on implementing the tool in the form of a web service was made as it is easier to reach a wider audience of experimentalists who are frequently reluctant to install a software package or use complex command lines and pieces of code. The web service offers 4 main workflows, as well as five other secondary workflows, each allowing the user to explore the data from a different angle (Figure 29). The full web service can be accessed currently at the temporary address:

http://tf-investigator.sybig.de/TRC

The workflows are interactive and interconnected, allowing the user to use the results of one workflow such as a co-expression cluster from the co-expression workflow and ask for correlated or enriched regulators that regulate this cluster using the correlated TFs workflow. Another example would be of a user using a set of stage-specific regulators of a certain time point from the TRC workflow, asking for the targets genes of these regulators using the correlated targets workflow, evaluating this target set using GO enrichment, filtering the target set for genes associated with a particular enriched GO term and then using this filtered set for a seed-based co-expression analysis.

This tool offers interactive results networks and sets, which is different from the typical black box results of many other methods. These results are usually distilled into a size that can facilitate visual inspection and allows the researcher to use biological intuition and his background knowledge to focus on particular genes and patterns in the results and investigate them deeper, rather than being overwhelmed by an information overload. The aim is to allow the researcher to guide the results by eliminating what is not compatible with the biological background of the experiment and keeping what makes biologically sense through incremental and iterative steps ending with biologically sensible results and new insights.

In what follows of this subsection is a detailed description of the main workflows as well as some snapshots that might give a glimpse of the inner workings of the web service.

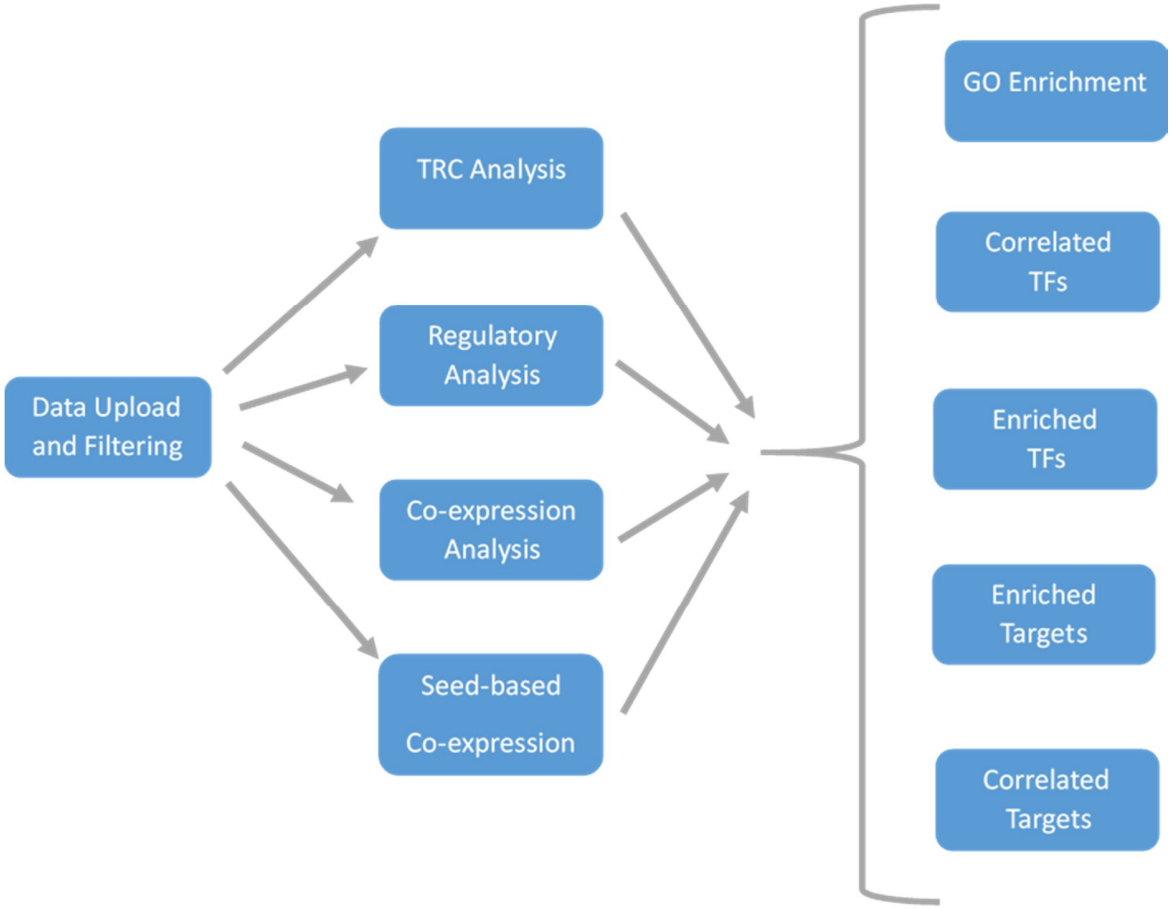

Figure 29. An overview of the main components of the web service and its general architecture.

5.3.1 Co-expression Workflow

This workflow applies a general analysis assuming no prior knowledge of genes of interest or patterns. It provides a platform to identify co-expressed genes and to investigate the regulatory forces that drive these highly correlated genes to be expressed in a similar pattern. It is based on the hypothesis that genes that are highly co-expressed are likely to be controlled by a set of common regulators and/or be involved in the same biological process.

Step 1: The expression data is uploaded and filtered according to different options and parameters. The correlation threshold indicates the minimum correlation required for a pair of genes to qualify as co-expressed. The expression threshold indicates the minimum expression level a gene should have, at least in one of the time points, to qualify for the next step of the analysis; this is used to eliminate lowly expressed genes. The user could also specifically use only the regulatory genes for the analysis, filter invariant genes, or those with many zero counts, as well as those that have no records in the background regulatory network.

Step 2: A co-expression network is constructed using a modified Pearson coefficient based on the methodology used in WGCNA [93], with the correlation threshold.

Step 3: Then, the n largest clusters in the previous network are chosen, and each is reduced to the top m hubs. This condenses the co-expression network into its significant components, making it easy to analyze visually. By default, in the web workflow, m and n are set to be 10, as this produces a relatively concise network that can be visually analyzed.

Step 4: The user selects a cluster from the displayed reduced network, or a collection of genes that might be interesting, then moves to the next workflows to investigate the regulators or the targets of this group of genes.

The co-expression workflow is interactive, where the user can toggle through different clusters viewing the general expression pattern of the genes in the side panel and using genes in the cluster as seeds for another workflow or a GO enrichment analysis (Figure 30).

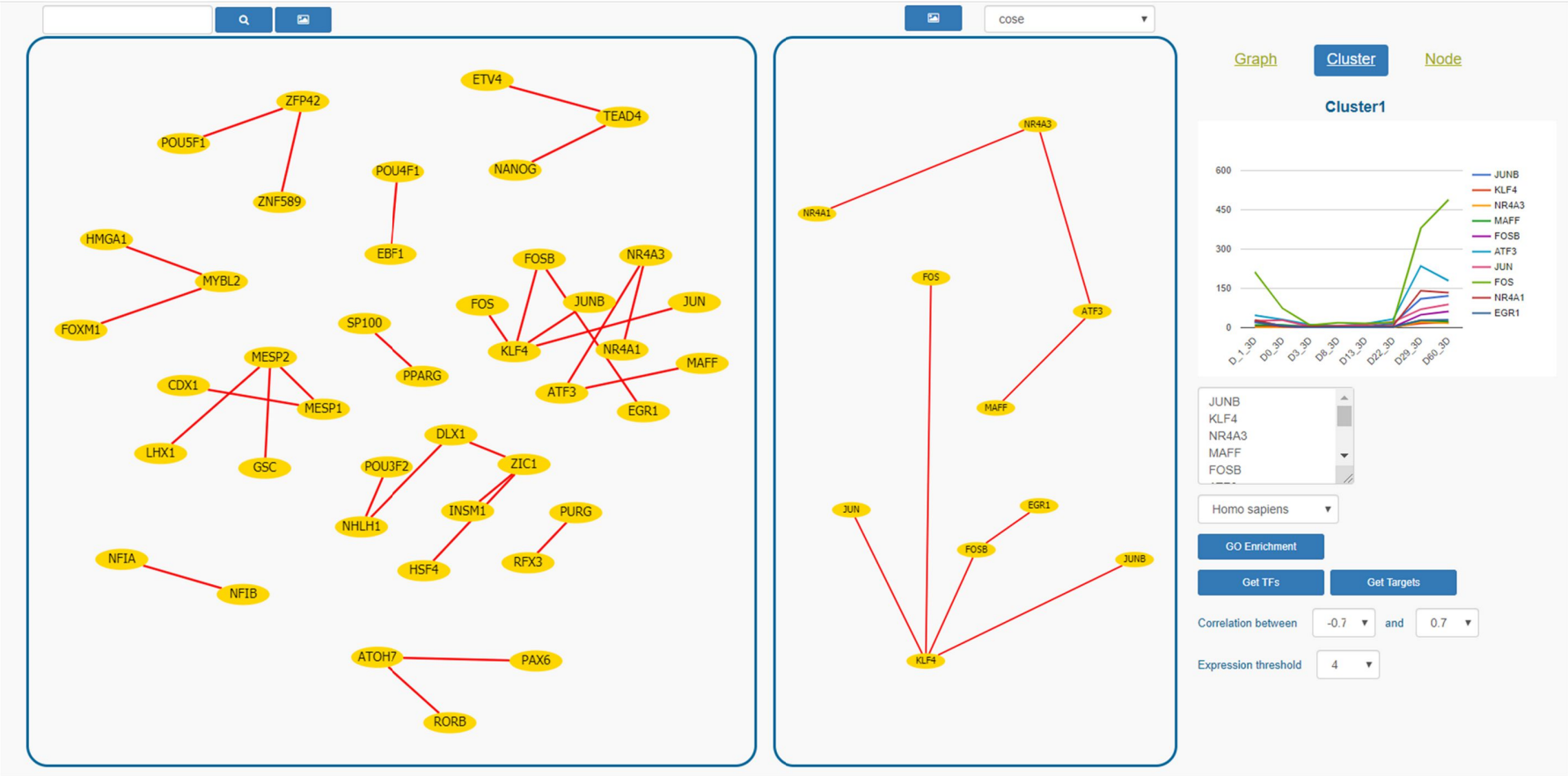

Figure 30. A snapshot of the co-expression workflow in action. The reduced co-expression clusters on the left, the zoomed-in cluster view in the middle, and on the right is the panel showing detailed information about the cluster selected, such as the expression pattern of the genes in the cluster and external links to other workflows.

5.3.2 Seed-Based Co-expression Analysis Workflow

This workflow is ideal when a small set of genes are previously known to be relevant to the experiment. It offers a platform to find genes that are related to the genes of interest in terms of pattern, by creating a co-expression network based on the input genes as a seed (Figure 31). It can be used efficiently to expand on previous knowledge of key genes and find new significant candidates.

Step 1: The expression data is uploaded and filtered according to different options and parameters. The correlation threshold indicates the minimum correlation required for a gene to be correlated with one of the seed genes to qualify as co-expressed. The expression threshold indicates the minimum expression level a gene should have, at least in one of the time points, to qualify for the analysis; this is used to eliminate lowly expressed genes. The cluster size represents the maximum number of correlated genes to be calculated for each seed gene. The user could also specifically use only the regulatory genes for the analysis, filter invariant genes, or those with many zero counts, as well as those that have no records in the background regulatory network.

Step 2: Each of the top correlated genes for each seed gene will be included and connected to the seed gene to form a cluster around the seed.

Step 3: The user selects a cluster from the displayed reduced network, or a collection of genes that might be interesting, then moves to the next workflows to investigate the regulators or the targets of this group of genes.

The co-expression workflow is interactive, where the user can toggle through different clusters viewing the general expression pattern of the genes in the side panel and using genes in the cluster as seeds for another workflow or a GO enrichment analysis.

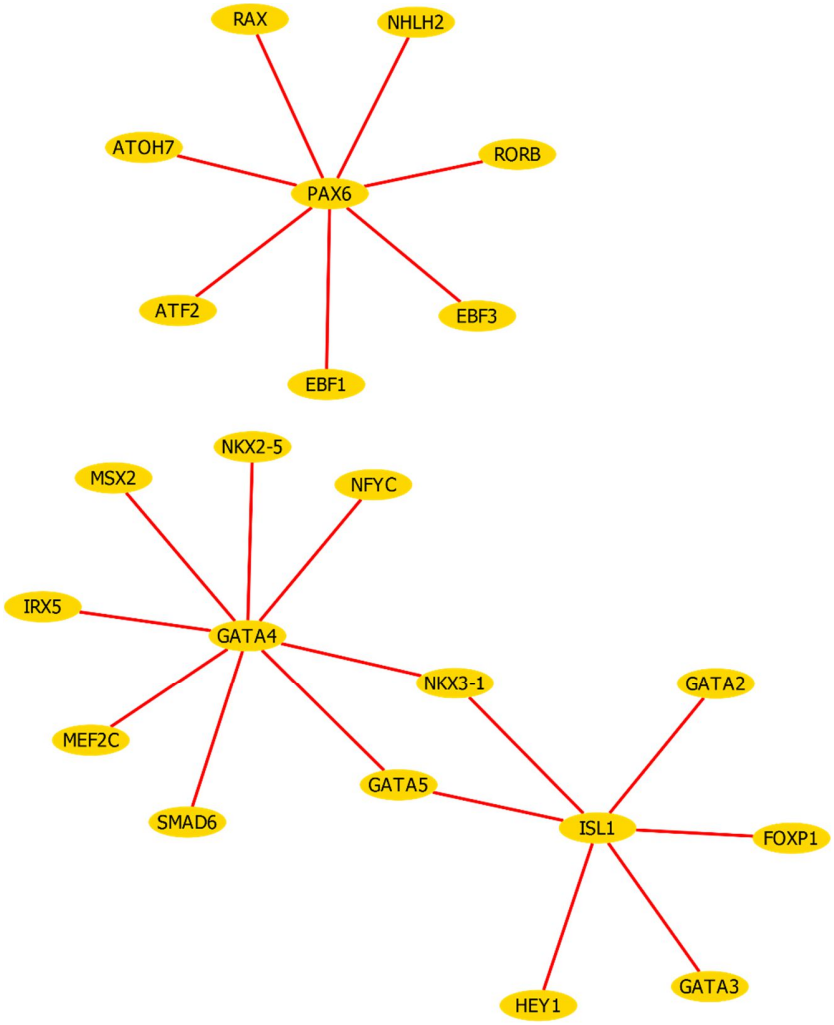

Figure 31. A seed-based co-expression network with the input seeds being ISL1, GATA4, and PAX6. The genes connected to each seed represent the top correlated gene; in this case, the network is filtered for only TF genes.

5.3.3 Regulatory Analysis Workflow

This workflow is dedicated to querying TFs and target genes based on the regulatory network with or without consideration of the relative expression patterns. The input is a list of genes, typically genes known to be involved in a certain pathway or biological process, co-expressed genes, or any genes of interest. The user can perform a GO enrichment analysis on this set or ask for TFs or target genes for this set. The workflow includes an enrichment of TFs that target the input list of genes based on a hypergeometric test and the background regulatory network, as well as an enrichment of the target genes of the input list in case the input list contained some regulators. The user can also ask for TFs or target genes of the input gene list based on the background regulatory network under the condition that these

TFs or target genes satisfy a certain correlation threshold to the associated gene in the list. This correlation threshold can be set by the user as a range from a negative correlation to a positive one. And similarly to the other workflows, the minimum expression threshold for the genes in the results can be adjusted as a parameter in the interface.

The result is displayed as a network where the input seed genes are displayed as green nodes in the center, and the corresponding TFs or targets are displayed as smaller yellow nodes around in a circular formation. The edges are outgoing from the seed gene nodes to the corresponding targets in case of analyzing enriched or correlated targets and incoming from the TF nodes to the seed nodes in the case of analyzing enriched or correlated TFs. In the case of an experimentally verified regulatory interaction, the edge is displayed in red color (Figure 32).

The side panel contains three tabs that display general information about the graph, about the edges and nodes, and other features. When clicking on a node, a graph displaying the expression pattern across the time points, where the mean of the replicates of a time point is used for the plot for simplicity, and a sortable table that summarizes all the nodes connected to this node as well as the correlation. Clicking on an edge displays a bar chart that shows when the regulatory interaction is active based on the expression levels of the regulator, the correlation between the source and the target, and a plot of the expression patterns of the source and the target of that edge. The third tab includes features such as analyzing the GO enrichment of the set of targets or the set of regulators of the seeds in the graph, as well as exporting these sets.

Options in the top panel allow the user to search for a node in the displayed network, export the current view as a high-quality image that can be used for posters, and change the layout into the other available ones in the dropdown list. Another option in the top panel is the "play" button, which animates the network where genes appear and fade away with across time, displaying the plastic nature and the dynamic aspect of the regulatory networks. The user can mark nodes and track them as they appear and disappear in different stages, which sometimes draws the attention to certain patterns that are not noticed otherwise. The nodes and edges don't change place as in many other dynamic network visualizers, which makes the tracking process practical. Another option "highlight stage" allows the user to compare two stages by highlighting one stage and then choosing another. The number and list of nodes and edges that appear and disappear between the two stages are displayed in the graph info panel.

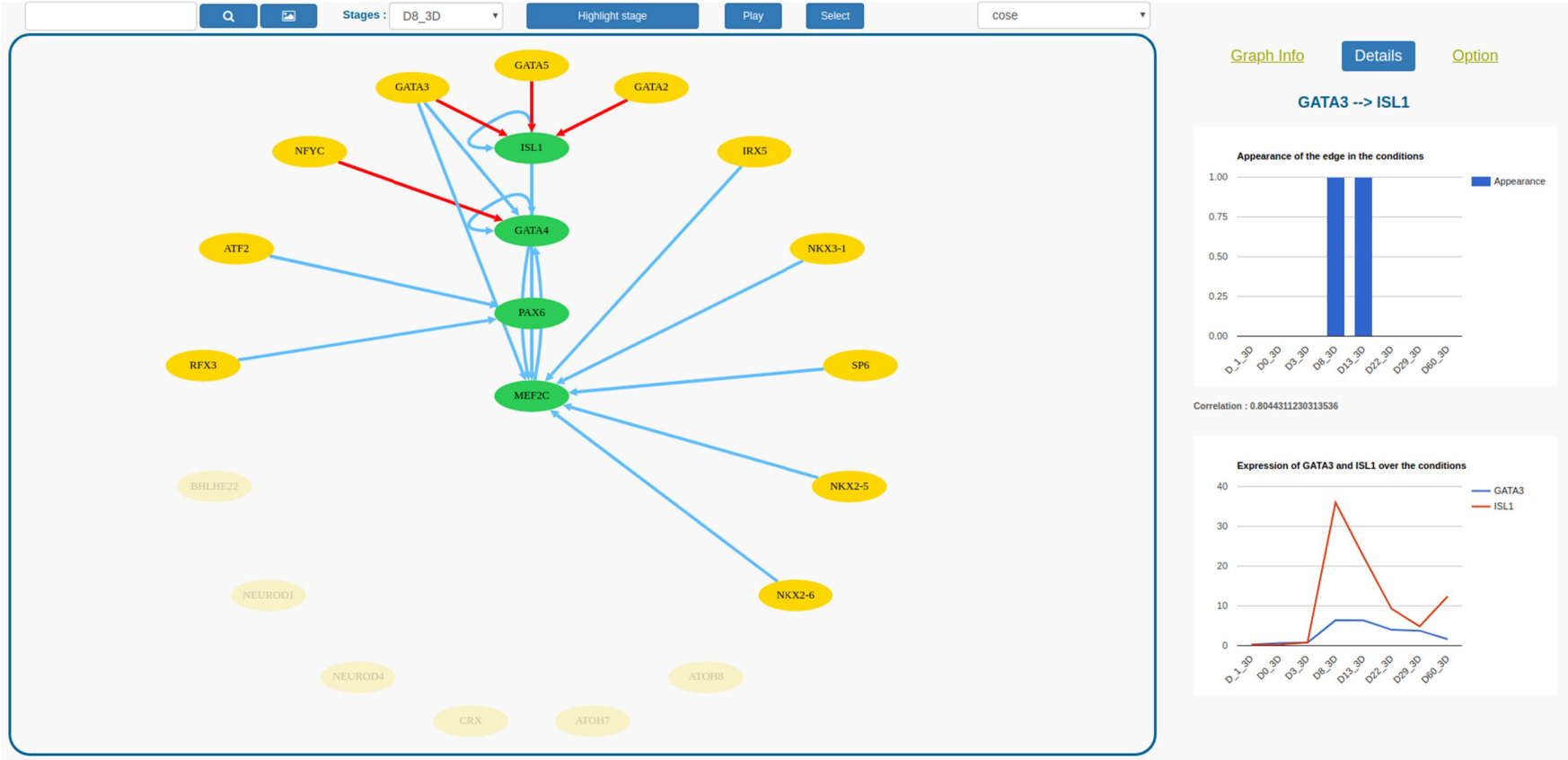

Figure 32. A snapshot of the regulatory network analysis workflow in action. Green nodes representing the seed genes and the yellow ones representing their potential TF. The faded nodes represent the unexpressed TFs at this particular stage as the network is animated. The panels to the side give information such as a plot about the activity of the edge and a comparison of the associated gene patterns.

5.3.4 The TRC Workflow

The opening page of the TRC workflow allows the adjustment of the 3 parameters, minE, minC, and maxS (see 5.2.4). The opening page also allows the choice of using only regulators, which is the default and part of the essence of the TRC method or removing that restriction to generate a cascade that contains non-regulators as well.

After the parameters are chosen, the TRC algorithm runs using the data file uploaded in the current session with the parameters and produces the resulting file on the server side. Next, the user is forwarded to the TRC interactive visualizer (Figure 33).

The cascade is animated at a slow pace, where the genes associated with the first stage are displayed as nodes in a column formation with the intra-regulatory interactions represented as vertical edges between them. The inter-regulatory interactions are animated next as edges coming out from the just animated genes, and the targets which are the genes from the next stage pop up next, and it goes on.

As the intra-regulatory edges are hard to see, being mostly hidden behind the nodes to give maximum visibility to the names of the genes in the nodes, the user can click on the stage header, which opens a window dedicated to the intra-regulatory network of this stage.

In the side panel, three tabs are accessible to provide information about the graph in general, the stage and the genes within, information around the node or edge clicked, and links for further analyzing gene sets in the other workflows or external ones.

The details panel, when clicking on a stage, offers a plot of the expression patterns of all the genes in that time point, the average peak strength of the stage and a sortable table that summarizes different useful metrics for the nodes belonging to that stage such as indegree and intra-regulatory influence. When a node is clicked, the details panel displays the expression pattern of the gene, the peak strength of the gene as well as a table that summarizes all the nodes connected to this node. When an edge is clicked, a plot that compares that expression patterns of the two connected genes of the edge is displayed, along with the correlation.

The last panel contains an editable text area that displays the list of genes contained in a certain time point that is clicked. The user can fill in, delete, or add any genes besides the default stage-specific ones that are automatically filled. This set can be used as a seed for any of the other secondary workflows, where the parameters to that workflow can be adjusted accordingly in the panel.

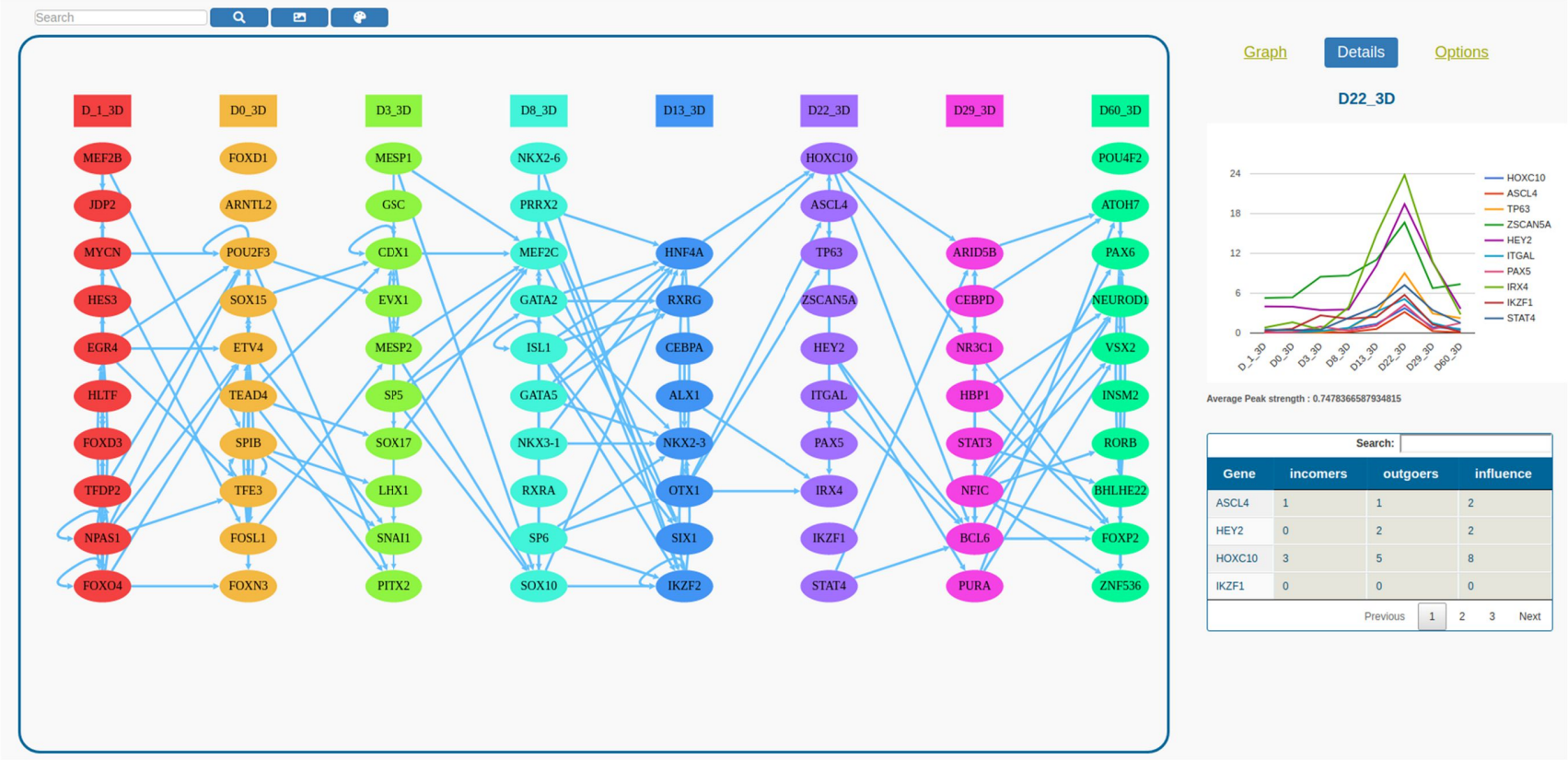

Figure 33. A snapshot of the TRC workflow in action. The TRC represented in the main view and the panel on the right displaying information about the stage-specifc regulators of the clicked stage, such as the expression patterns and a table summary of some topological features.

5.3.5 Implementation

The analysis of time series datasets and the generation of the TRC in a file format was implemented in java, and available as a jar file that takes an expression file, a regulatory network and the main TRC parameters and outputs a file containing the TRC in a format that is directly usable by the web visualizer.

The generation of the network utilized C as a fast programming language for the initial steps and SQL for storing the binding site predictions and relevant information. After that, java was used for organizing the network and transforming it into a graph database as well as file formats that can be used in the different workflows.

The web tool utilized PHP for the server end, and a combination of javascript, jQuery, and HTML for the client end. The visualization process of the plastic network and the TRCs relied heavily on javascript. The library cytoscape.js was used for visualizing nodes and edges and styling them.

For the visualization of the TRC model, we needed to create a custom layout to organize the nodes and edges in the required architecture and display it in an animated manner. I created a simple layout algorithm to organize the nodes in a specific architecture then display the nodes and edges as the time points unfold. The algorithm determining the position of each node goes as follows:

For each node in the layout, the following attributes are created:

Column: index of the stage *2 . This means each stage gets an even column for its nodes, and the odd columns remain empty. This firstly creates a clearer layout in the cascade with enough space between the full columns. Secondly, it creates distance, so the edges between the time points can be drawn effectively.

Row: The number of rows to be drawn is determined by the column with the highest number of nodes. After which the nodes are filled bottom up according to their sorted indices.

Color: the color of the node is determined by the index of the stage it belongs to. These colors are preprogrammed, where each index from 0 till the max columns allowed contains a color that follows the natural color spectrum and yet is contrasting enough with the column before and after.

Once each node is given a position defined by a particular row and column in the grid, these positions are transformed into pixel coordinates that vary according to the screen size and

zoom in levels. An adapter included within the cytoscape.js library is used to transform grid indices into physical locations.

Animation:

1. All nodes and edges are hidden.
2. For each time point or column:
   - Display the nodes associated with the column: nodes are unhidden, and their size is progressively grown from zero to the full size.
   - Display the edges associated with this column: reveal all the edges that are within that time point and the edges outgoing from the nodes in that column.
   - Wait for a few seconds to allow the nodes to fully animate, then lunch the next iteration.

For the regulatory network, we wanted a layout that displays nodes and as they fade in and out across the time without changing their locations as most of the tools do when displaying a dynamic network. This allows the user to keep track of the important nodes and be able to compare them as they change their expression through the stages.

Placement of the nodes: The seed genes, whether they are genes querying their regulators or genes querying their target genes, are placed inside a circle, and the queried regulators or targets are displayed in the parameter of this circle. This is achieved through a concentric layout based on the in or out-degree. Edges go outwards in the case of querying the targets and inwards into the seeds in the case of querying the regulators of the seeds.

Animation: The network can be animated, activating the genes that are expressed in the first condition, while the other nodes, though visible, are faded to indicate their lack of expression at that time point. Edges that originate from a faded node disappear while those that originate from an active node are activated. After a small time delay to allow the user to skim through the time point snapshot of the network, the animation proceeds to activate the nodes of the next time point and fading those that are not expressed next. During the animation, information about the transitions can be obtained, such as the number and lists of nodes that disappeared during the transition, the number of appearing and disappearing edges, and other information that can help the user assess the information held within this transition.

5.4 Heat Development Dataset Analysis

This subsection covers the analysis of the heart development dataset from different angles, aspects, and approaches.

5.4.1 TRC Analysis

The TRC was generated for the heart development dataset, and stage-specific key regulators and their interactions were detected. As a general overview, each time point in the TRC has at least 8 regulators associated with it, and most of the stages had 10, summing up to 76 regulators overall (Figure 34). The TRC contained no microRNAs, so all the regulators in it were TFs. All the stages were connected and contained intra-regulatory interactions. We performed a detailed analysis of each stage, focusing on the following:

1) The GO enrichment of the stage-specific regulators and the relevance of the terms to the differentiation events observed experimentally at that time point.
2) The identity of each of the regulators in that stage and a literature search on their general and specific roles and any experimental evidence of their involvement in cardiac development.
3) The regulatory interactions between the stages.
4) The intra-regulatory network of the stage and potential master regulators.
5) Potential co-regulators that might control the expression of the stage-specific regulators set.
6) Targets of the stage-specific regulators and their GO enrichment.
7) Potentially collaborating TF pairs that control the regulators set, and potential proximal binding sites used for these collaborations in the promoters of common targets.
8) Potential new regulatory candidates that might be important for cardiac differentiation among the set of regulators.

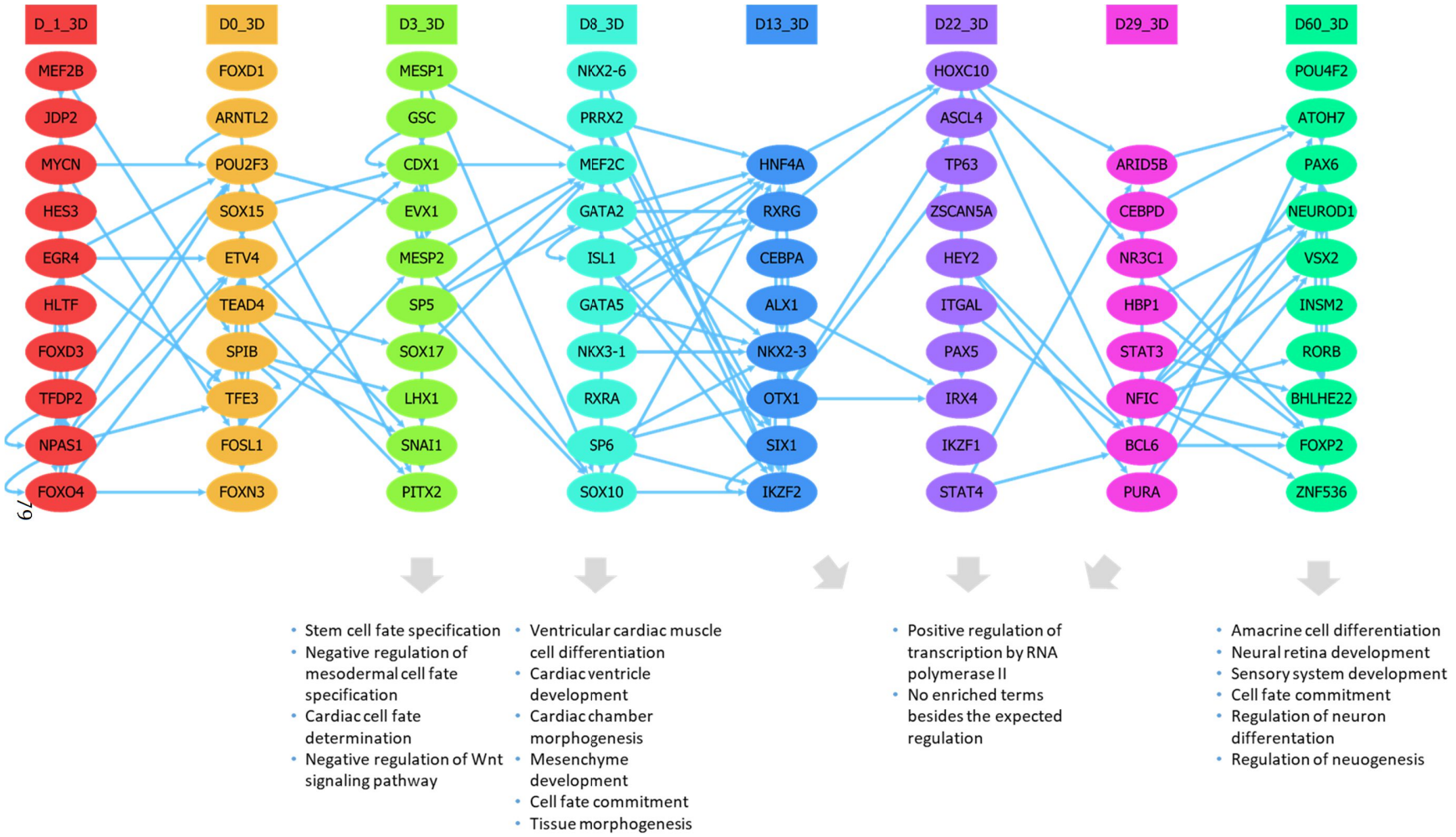

Figure 34. The TRC based on the heart development dataset. The relevant GO terms of the main stages are displayed at the bottom.

Day -1

Regulators of Day -1 (Figure 35) showed enrichment of general GO terms like regulation of transcription, which is biased in our case due to the restriction of the list to regulators; however, no specific terms appeared (Table 3). This is expected since, at this time point, the differentiation has not started yet, and the time point is a mere control, so we didn't expect to see any terms related to cardiac development or specific processes. Analyzing this stage was important for the cascade serving as a control, where the detected regulators in the other stages have to satisfy the condition of being lowly or non- expressed at this stage as a natural consequence of the TRC method.

Table 3. The top GO terms enriched for the regulators specific to Day -1.

GO Term	# Ref.	#	# Ex.	Fold	+ -	P - value
positive regulation of transcription by RNA polymerase II	1214	9	.58	15.57	+	6.32E-07
positive regulation of transcription, DNA-templated	1547	9	.74	12.21	+	5.47E-06
positive regulation of nucleic acid-templated transcription	1634	9	.78	11.56	+	8.90E-06
positive regulation of RNA biosynthetic process	1635	9	.78	11.56	+	8.95E-06
positive regulation of RNA metabolic process	1719	9	.82	10.99	+	1.40E-05
positive regulation of nucleobase-containing compound metabolic process	1881	9	.90	10.05	+	3.11E-05
positive regulation of macromolecule biosynthetic process	1899	9	.90	9.95	+	3.39E-05
positive regulation of cellular biosynthetic process	1985	9	.95	9.52	+	5.02E-05
positive regulation of gene expression	1998	9	.95	9.46	+	5.32E-05
positive regulation of biosynthetic process	2018	9	.96	9.36	+	5.81E-05
negative regulation of gene expression	1716	7	.82	8.56	+	2.11E-02
regulation of transcription by RNA polymerase II	2699	10	1.29	7.78	+	1.12E-05
positive regulation of nitrogen compound metabolic process	3173	10	1.51	6.62	+	5.62E-05
positive regulation of cellular metabolic process	3314	10	1.58	6.34	+	8.67E-05
positive regulation of macromolecule metabolic process	3343	10	1.59	6.28	+	9.46E-05

Figure 35. The intra-regulatory network corresponding to Day -1 the control stage.

Day 0

Similarly, Regulators of Day 0 (Figure 36) showed enrichment of general GO terms like regulation of transcription, which is biased in our case due to the restriction of the list to regulators, and no specific terms appeared (Table 4). This is also expected since at this time point the seeding of the stem cells has just finished, the mesoderm induction is about to start, and the differentiation has not started yet, so we didn't expect to see any terms related to cardiac development or specific processes yet. Analyzing this stage was not of importance, but it was also serving as a control in the TRC, where the detected regulators in the other stages have to satisfy the condition of being lowly or non-expressed at this stage as a natural consequence of the TRC method.

Table 4. The top GO terms enriched for the regulators specific to Day 0.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/-	P - value
positive regulation of transcription by RNA polymerase II	1214	9	.58	15.57	+	6.32E-07
positive regulation of transcription, DNA-templated	1547	9	.74	12.21	+	5.47E-06
positive regulation of nucleic acid-templated transcription	1634	9	.78	11.56	+	8.90E-06
positive regulation of RNA biosynthetic process	1635	9	.78	11.56	+	8.95E-06
positive regulation of RNA metabolic process	1719	9	.82	10.99	+	1.40E-05
positive regulation of nucleobase-containing compound metabolic process	1881	9	.90	10.05	+	3.11E-05
positive regulation of macromolecule biosynthetic process	1899	9	.90	9.95	+	3.39E-05
positive regulation of cellular biosynthetic process	1985	9	.95	9.52	+	5.02E-05
positive regulation of gene expression	1998	9	.95	9.46	+	5.32E-05
positive regulation of biosynthetic process	2018	9	.96	9.36	+	5.81E-05
regulation of transcription by RNA polymerase II	2699	10	1.29	7.78	+	1.12E-05
regulation of transcription, DNA-templated	3516	10	1.67	5.97	+	1.57E-04
positive regulation of nitrogen compound metabolic process	3173	9	1.51	5.96	+	3.20E-03
regulation of nucleic acid-templated transcription	3574	10	1.70	5.87	+	1.84E-04
regulation of RNA biosynthetic process	3579	10	1.70	5.87	+	1.87E-04

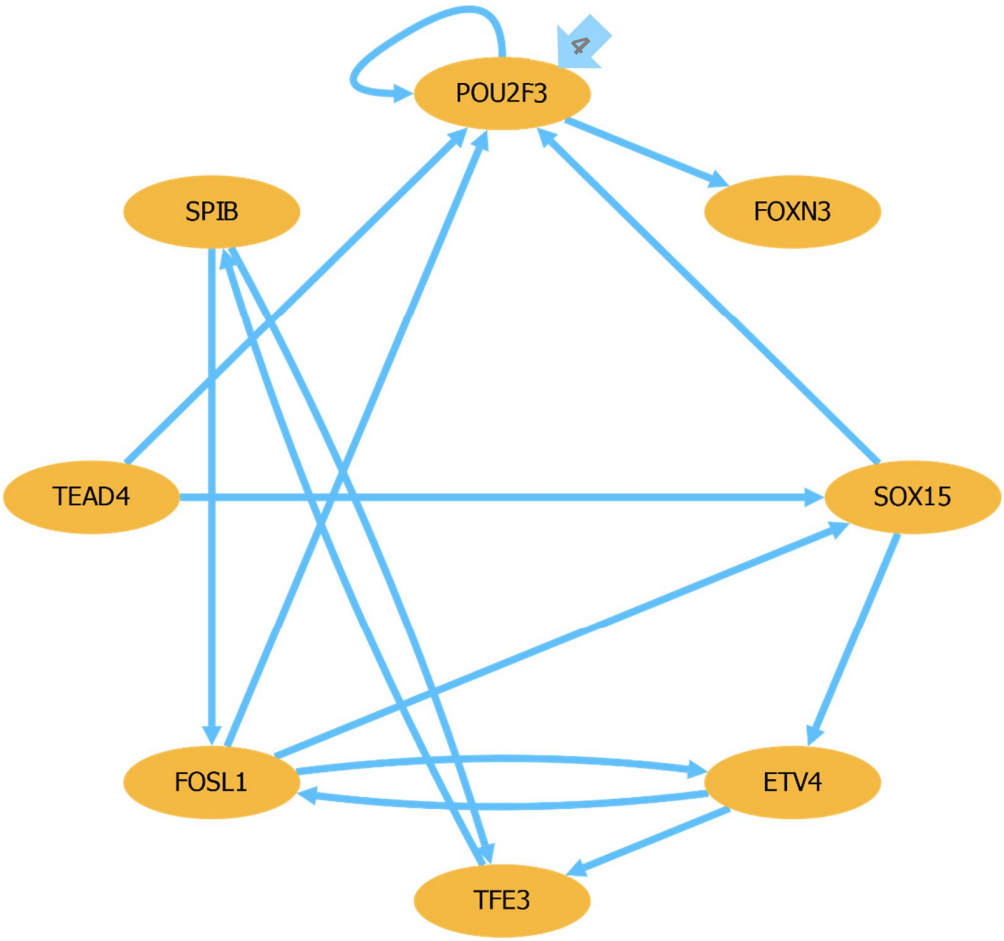

Figure 36. The intra-regulatory network corresponding to Day 0. POU2F3 being heavily regulated from the TFs of the previous stage with 4 incoming inter-regulatory edges.

Day 3

Day 3 marks the end of the mesoderm induction face, where CHIR99021, FGF-2, BMP4, Activin-A have been added for three days. The GO enrichment showed very specific and biologically consistent terms (Table 5). The top terms sorted by fold enrichment and satisfying a p-value >0.05 were overwhelmingly related to mesoderm formation, stem cell specification, cardiac cell fate determination, embryonic development, primary germ layer formation, and heart morphogenesis. This enrichment is very consistent with the fact that the stem cells have differentiated into mesodermal cells by Day 3. These mesodermal cells provide the basis for the cardiac differentiation that follows. This enrichment gives a certain level of confidence that these 10 regulators specific for Day 3, detected by the TRC, are indeed relevant and provide a good basis for further analysis.

Table 5. The top GO terms enriched for the regulators specific to Day 3.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/-	P-value
stem cell fate specification	4	2	.00	> 100	+	3.33E-02
negative regulation of mesodermal cell differentiation	4	2	.00	> 100	+	3.33E-02
negative regulation of mesodermal cell fate specification	4	2	.00	> 100	+	3.33E-02
cardiac cell fate determination	4	2	.00	> 100	+	3.33E-02
negative regulation of Wnt signaling pathway involved in heart development	5	2	.00	> 100	+	4.66E-02
signal transduction involved in regulation of gene expression	20	3	.01	> 100	+	1.68E-03
negative regulation of embryonic development	26	3	.01	> 100	+	3.45E-03
regulation of gastrulation	37	3	.02	> 100	+	9.30E-03
somitogenesis	66	3	.03	86.76	+	4.89E-02
formation of primary germ layer	113	5	.06	84.46	+	2.06E-05
regulation of embryonic development	129	5	.07	73.98	+	3.91E-05
gastrulation	159	6	.08	72.03	+	8.54E-07
mesoderm development	126	4	.07	60.59	+	3.98E-03
regionalization	334	9	.17	51.43	+	3.54E-11
anterior/posterior pattern specification	218	5	.11	43.78	+	5.04E-04
pattern specification process	433	9	.23	39.67	+	3.52E-10
heart morphogenesis	250	5	.13	38.17	+	9.83E-04

The regulators of Day 3 were MESP1, GSC, CDX1, EVX1, MESP2, SP5, SOX17, LHX1, SNAI1, PITX2 (Figure 37). We examined the function of each regulator. MESP1 (Mesoderm Posterior BHLH Transcription Factor 1) and MESP2 (Mesoderm Posterior BHLH Transcription Factor 2) are TFs known to be essential for the formation of the cardiac mesoderm [94]. MESP1 is one of the earliest TFs that gives rise to a set of cardiac-specific TFs [95]–[99]. This fact fits with the intra-regulatory network; MESP1 appears to be the master regulator of the set,

regulating 3 TFs CDX1, EVX1, SNAI1 and at the same time, not being regulated by any. Although MESP1 and MESP2 share a lot in common in terms of their structure and transcriptional targets, their roles are not completely redundant [100]–[104]. Experiments have shown that the knockout of one of them is not covered for by the other, and the effects are fatal for the heart. ChIP-seq experiments have shown that MESP1 has unique and specific targets such as RASGRP3 and PRICKLE, and controls the speed and direction of cell migration [105]. GSC is another TF observed peaking in Day 3 and known to be one of the markers of mesendoderm [106], and known as well to be regulated by MESP1, however the ChIP-seq analysis showed that MESP1 binds around 5 Kb upstream of GSC and have been hypothesized to act more of an enhancer of GSC [107]. However, In the intra-network of Day3 we can show that MESP1 is also able to regulate the expression of GSC indirectly via CDX1which might also explain its influence on its expression in an indirect way. CDX1 is the most active node in the intra-network of Day 3, regulated by four Day 3 regulators as well as regulating 2 others and itself. This predicted auto-regulation of CDX1 is also experimentally verified [108]. CDX1 is also known to be involved in the differentiation and proliferation of different cell types, including cardiac cells [109]–[112], which indicates that its role is more general in proliferation rather than being responsible for the cardiac specification. SP5 is a known marker for mesoderm [113] and a target of the Wnt/β-catenin signaling pathway [114], however the regulatory analysis in the Day 3 network shows that its role might be indirectly regulating the critical TFs CDX1 and SOX17. SOX17 has been shown to be essential for the specification of cardiac mesoderm in mice [115], and here we can hypothesize that this also applies to human cardiac cells. SOX17 might function as one of the direct regulators of CDX1 and potentially collaborating with some other Day 3 regulators such as MESP1, MESP2, and SP5 in regulating it, or contributes to potential redundancy in the regulation of this essential TFs providing more robustness to the network by protecting the central node. SNAI1 is a zinc finger transcription factor that has been shown to promote the exit from the pluripotency by direct repression of self-renewal genes [116] [117]. The regulatory network shows that MESP1 and MESP2 might be responsible for the induction of SNAI1 and the resulting repression cascade. Some studies have shown that PITX2, another TF peaking in Day 3, regulates the left-right asymmetry by patterning second cardiac lineage-derived myocardium. EVX1 is another Day 3 TF that is required for patterning and gastrulation [118]–[120]. LHX1 does not appear in the intra-network due to the lack of its involvement in any intra-regulatory interactions; however, it is known for its involvement in the formation of the epithelium [121], in this case, we can see its potential involvement in the mesoderm genesis as well.

Figure 37. The intra-regulatory network corresponding to Day 3.

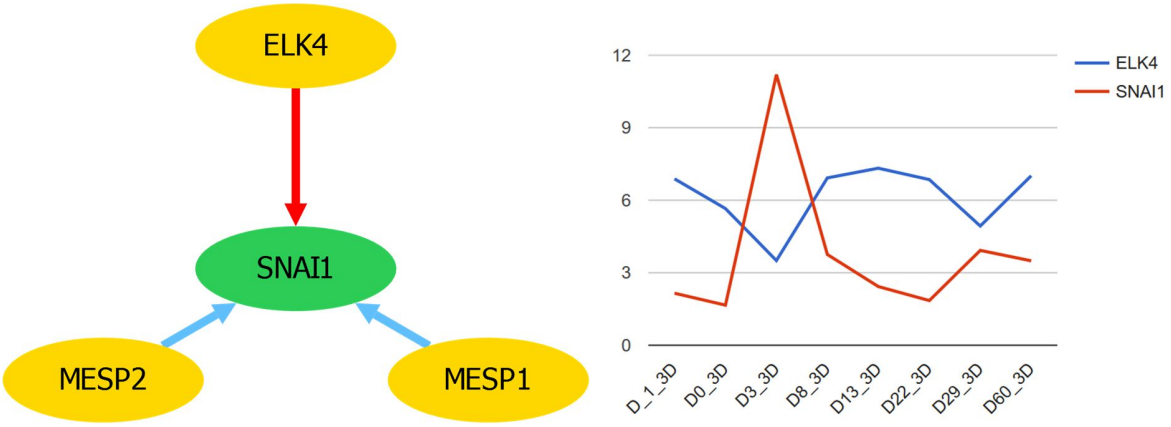

Figure 38. (Left) The top correlated regulators of SNAI1. (Right) The anti-correlated patterns of ELK4 and SNAI1.

Interestingly many TFs in this set are involved in cancer and tumor genesis, such as SNAI1, which marks highly resistant tumors [122]. It is not clear whether the expression of SNAI1 is assisting the tumor growth or is a response to fight the tumor growth. We investigated the potential TFs for SNAI1 in the whole dataset and not only in the TRC, and a new TF ELK4

appeared to be regulating it, besides MESP1 and MESP2 (Figure 38). However, what is interesting about ELK4 is its strong anti-correlating pattern, which might indicate its repressive influence on SNAI1 and might be a potential drug target for combating tumors by reducing the expression of SNAI1.

Studying the targets of the Day 3 TFs, we queried the regulatory network for targets of these TFs that have a correlation between -0.7 and 0.7 to its regulator and resulted in 94 target genes. The GO enrichment of the target gene set was related to gastrulation, regionalization, pattern specification, chromatin assembly and disassembly, and the formation of a primary germ layer (Table 6). These terms are perfectly consistent with the biology underlying the events at Day 3, where the chromatin is being opened for a wave of cardiac-specific regulatory events, and the mesodermal layer is formed providing the basis for the upcoming more specific layer of cardiac cells and shows that the targets of these TFs are also specific and constitute part of the wave.

Table 6. The GO enrichment of the targets of Day 3 TFs.

GO Term	***Nb. in Reference***	***Nb. in upload***	***Nb. Expected***	***Fold Enrich.***	***+/ -***	***P - value***
signal transduction involved in regulation of gene expression	20	4	.09	44.67	+	3.29E-02
formation of primary germ layer	112	7	.50	13.96	+	9.30E-03
nucleosome assembly	116	7	.52	13.48	+	1.16E-02
gastrulation	158	9	.71	12.72	+	5.27E-04
chromatin assembly	133	7	.60	11.76	+	2.77E-02
chromatin assembly or disassembly	154	8	.69	11.60	+	5.80E-03
regionalization	337	13	1.51	8.62	+	4.66E-05
pattern specification process	433	14	1.94	7.22	+	1.00E-04
Unclassified	3156	3	14.13	.21	-	0.00E00

Furthermore, we investigated whether the TFs of Day 3 collaborate in the regulation of the target set and for that, we utilized the PC-TraFF algorithm. We ran the extended version of the PC-raff algorithm, searching for potential TF collaborations that utilize the PWMs associated with the TFs of Day 3 in the promoter regions of the correlated target set of these regulators. The maximal distance between the pairs was chosen as 20, and the z-score cutoff was 3. One collaborating pair was detected to be specific for this set, and that was V$CDXA_02 - V$CDX1_01 with a z-score of 4.22. Both PWMs in the pair were associated with

CDX1, which suggests the potential utilization of CDX1 of proximal binding sites in the promoters of its targets to create a bigger protein complex.

Day8

Regulators of Day 8 show very specific cardiac terms in the GO enrichment, such as ventricular cardiac muscle differentiation, cardiocyte differentiation, and heart morphogenesis (Table 7). These terms go very well in accordance with the biological events happening in the experiment at this stage. At Day 8, early cardiac specification is occurring, and cells are starting to differentiate into not-yet mature cardiomyocytes.

Table 7. The GO enrichment of the TFs of Day 8.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/ -	P - value
ventricular cardiac muscle cell differentiation	18	3	.01	> 100	+	9.20E-04
pharyngeal system development	26	3	.01	> 100	+	2.52E-03
cell fate determination	42	3	.02	> 100	+	9.76E-03
cardiac ventricle morphogenesis	72	3	.03	87.48	+	4.61E-02
cardiocyte differentiation	116	4	.06	72.40	+	1.85E-03
cardiac muscle tissue development	159	5	.08	66.03	+	5.96E-05
cellular response to steroid hormone stimulus	187	4	.09	44.91	+	1.19E-02
mesenchyme development	216	4	.10	38.88	+	2.09E-02
striated muscle tissue development	285	5	.14	36.84	+	1.03E-03
muscle tissue development	298	5	.14	35.23	+	1.28E-03
heart morphogenesis	249	4	.12	33.73	+	3.65E-02
regulation of animal organ morphogenesis	256	4	.12	32.81	+	4.06E-02
gland development	417	5	.20	25.18	+	6.63E-03
heart development	528	6	.25	23.86	+	4.52E-04
chordate embryonic development	640	7	.30	22.96	+	2.52E-05

tissue morphogenesis	560	6	.27	22.50	+	6.39E-04

The main regulator that stands out at this stage is MEF2C, a heavily regulated TF by both regulators of the previous stage and the same stage, with 13 incoming edges (Figure 39). MEF2C is essentially known to be one of, if not the most essential TF for cardiomyocyte development and specification, cooperating and interacting with other cardiac factors and forming the core of the cardiac differentiation regulatory network [123]–[129]. The observed heavy regulation of MEF2C can be either due to a precise simultaneous corporation between the TFs binding to its promoter which leads to its activation exactly at that time point, or redundancy in the function of these TFs in activating MEF2C providing the robustness needed in case one of the activators was eliminated. SP6 regulates 3 TFs, SOX10, MEF2C, and GATA2 in the network and is not regulated by any other TF, making it a potential master regulator of the Day 8 regulatory network. Although there is no particular evidence in the literature of the involvement of SP6 in cardiac development and differentiation, we hypothesize that its role might be crucial in the specification stage, and provides a good candidate for further experimental investigation. PRRX2 is another regulator in this network, regulating MEF2C and GATA2, and is involved in vascular smooth muscle differentiation [130]. Observing it coexpressed and peaking at this stage of cardiac specification hints to either its additional role in cardiac muscle cell development or, more probably, its role as a key regulator that initiates the development of a vascular network that supplies blood and is integrated within the developing heart. This indicates that knocking out PRRX2 might enhance the purity of the cardiomyocyte culture in the experiment by eliminating the development of vascular smooth muscle cells within. GATA2 appears to be peaking in Day 8 as well , however little is known about its role in cardiac differentiation. As it belongs to the same family like GATA4 and GATA5 which are important in cardiac differentiation, we hypothesize that its role here is redundant. GATA5, a TF required of cardiac differentiation [131]–[134], appears specifically on Day 8. SOX10 is another active regulator in Day 8, has little evidence for its involvement in cardiac differentiation however provides a good candidate for experimental investigation due to involvement of other SOX family members such as SOX4 and SOX11 that are heavily involved in cardiac differentiation [135]–[137] thus it could be taking over some of their roles. NKX2-6 is an understudied TF in the literature; however, it is very specific Day 8 expression coupled with the fact that NKX2-5 another family member of the NK-2 homeobox family is one of the most essential TFs for cardiac development [131][132]opens the question about its potential involvement in cardiac differentiation. Worth to notice that NKX2-5 and NKX2-6 are both homologs of the of drosophila homeobox-containing protein called 'tinman', which has been shown to be essential for the development of the heart-like dorsal vessel [140]. ISL1, which regulates itself as well as regulating MEF2C on Day 8 , is known to be another crucial TF for cardiac specification [95], [141]–[146]. Strong experimental evidence exists for MEF2C being a direct transcriptional target of ISL1 and GATA factors in the embryonic heart [147], which supports the observed regulatory prediction on Day 8.

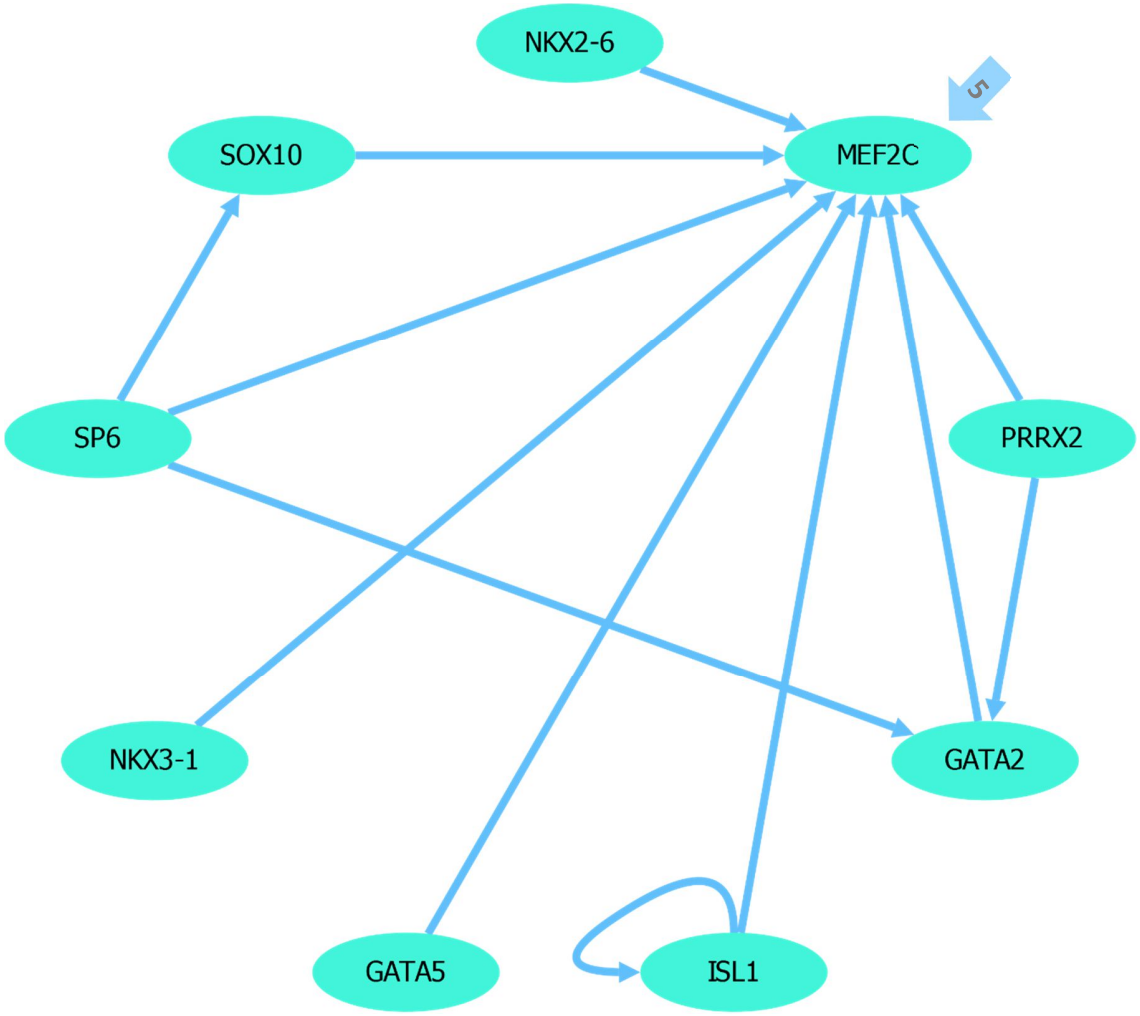

Figure 39. The intra-regulatory network corresponding to Day 8. MEF2C being heavily regulated by the TFs of the previous stage with 5 incoming inter-regulatory edges.

Studying the targets of the Day 8 TFs, we queried the regulatory network for targets of these TFs that have a correlation between -0.7 and 0.7 to its regulator and resulted in 154 target genes, a much wider set of correlated targets than in Day 3. The GO enrichment of the target gene set was just like the TFs very cardiac-specific. Terms ranged from ventricle formation, valve morphogenesis, to cardiac differentiation. These terms are perfectly supported by the events going on in Day 8, where the cardiac specification is happening, and immature cardiomyocytes are appearing. The complete list of the Day 8 specific target set can be found in the Appendix.

Table 8 . The GO enrichment of the targets of Day 8 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/ -	P - value
cardiac right ventricle morphogenesis	19	7	.14	50.23	+	4.96E-06
cardiac chamber formation	11	4	.08	49.58	+	3.10E-02
pulmonary valve morphogenesis	17	5	.12	40.10	+	4.09E-03
aortic valve morphogenesis	28	7	.21	34.08	+	4.80E-05
pulmonary valve development	21	5	.15	32.46	+	9.97E-03
atrioventricular valve development	26	6	.19	31.46	+	9.37E-04
aortic valve development	32	7	.23	29.82	+	1.07E-04
semi-lunar valve development	37	8	.27	29.48	+	1.01E-05
atrioventricular valve morphogenesis	24	5	.18	28.40	+	1.77E-02
heart valve morphogenesis	52	9	.38	23.60	+	5.11E-06
cardiac ventricle morphogenesis	72	12	.53	22.72	+	9.27E-09
cardiac atrium morphogenesis	30	5	.22	22.72	+	4.67E-02
heart valve development	61	10	.45	22.35	+	8.81E-07
endocardial cushion development	45	6	.33	18.18	+	1.66E-02
ventricular cardiac muscle tissue development	55	7	.40	17.35	+	2.98E-03
ventricular cardiac muscle tissue morphogenesis	48	6	.35	17.04	+	2.34E-02
cardiac ventricle development	128	15	.94	15.98	+	1.32E-09
myofibril assembly	61	7	.45	15.65	+	5.66E-03
cardiac chamber morphogenesis	127	14	.93	15.03	+	1.89E-08

cardiac muscle tissue morphogenesis	64	7	.47	14.91	+	7.62E-03
ventricular septum development	72	7	.53	13.26	+	1.58E-02
muscle tissue morphogenesis	75	7	.55	12.72	+	2.04E-02
muscle cell development	146	13	1.07	12.14	+	1.40E-06
cardiac chamber development	170	15	1.25	12.03	+	5.78E-08
cardiocyte differentiation	116	10	.85	11.75	+	2.61E-04

Furthermore, we investigated whether the TFs of Day 3 collaborate in the regulation of the target set, and for that, we utilized the PC-TraFF algorithm. We ran the extended version of the PC-raff algorithm, searching for potential TF collaborations that utilize the PWMs associated with the TFs of Day 8 in the promoter regions of the correlated target set of these regulators. The maximal distance between the pairs was chosen as 20, and the z-score cutoff was 3. The result was 10 potentially collaborating PWM pairs (Table 9), among which 9 were specific to this set. While most pairs were composed of PWMs associated with the same TF, with SP5 associated with 6 of these pairs, 3 pairs stood out where the PWMs belonged to different TFs: V$ISL2_05 - V$PRX2_Q2, V$ISL2_02 - V$PMX1_Q6, and V$GATA6_04 - V$PMX1_Q6.

V$ISL2_05 - V$PRX2_Q2 and V$ISL2_02 - V$PMX1_Q6 point to the potential collaboration of PRRX2 and ISL1 in regulating a set of targets. Furthermore, V$GATA6_04 - V$PMX1_Q6 points to the potential collaboration of GATA2 and/or GATA5 with PRRX2 to regulate a set of targets.

Table 9. The potentially collaborating PWM pairs found in the promoters of the target set of the Day 8 TFs.

Matrix 1	**Matrix 2**	**Z-score**	**Background Difference**
V$PRX2_Q2	V$PMX1_Q6	4.750444077	-0.001241656
V$SP1_Q6_01	V$SP1_Q2_01	3.594553147	0.002856102
V$SP1_Q6	V$SP1_Q4_01	3.113176652	0.002223333
V$GATA6_04	V$PMX1_Q6	3.001578304	0.001683787
V$SP1_Q2_01	V$SP1_03	3.941624451	0.004473108
V$ISL2_05	V$PRX2_Q2	3.692424705	0.003380307
V$ISL2_02	V$PMX1_Q6	3.300999227	0.001934722
V$SP1_Q2_01	V$SP1_02	5.099317264	0.005619215
V$SP1_02	V$SP1_03	6.730314615	0.005515271
V$SP1_Q6_01	V$SP1_03	3.669435411	0.003446728

Day 13

In contrast to the previous stage, regulators of Day 13 (Figure 40) showed enrichment of general GO terms like regulation of transcription, which is biased in our case due to the restriction of the list to regulators, and no specific terms appeared (Table 10). This indicates that no novel or unique activity happens on Day 13 that stands out from another stage. However, the expression profile of some important cardiac TFs such as NKX2-5 and GATA4 showed that these factors are present in Day13, but their expression starts on Day 8, after which they continue being expressed till Day 29, but that doesn't make them unique for Day 13. The target set of the TFs at this stage showed no significant enrichment at all.

Table 10. The GO terms enriched in the TFs of Day 13.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/-	P - value
positive regulation of transcription by RNA polymerase II	1212	8	.46	17.32	+	1.13E-06
positive regulation of transcription, DNA-templated	1553	8	.59	13.52	+	8.19E-06
positive regulation of nucleic acid-templated transcription	1651	8	.63	12.72	+	1.33E-05
positive regulation of RNA biosynthetic process	1652	8	.63	12.71	+	1.34E-05
positive regulation of RNA metabolic process	1736	8	.66	12.09	+	1.99E-05
positive regulation of nucleobase-containing compound metabolic process	1897	8	.72	11.07	+	4.04E-05
positive regulation of macromolecule biosynthetic process	1919	8	.73	10.94	+	4.43E-05
positive regulation of cellular biosynthetic process	2007	8	.76	10.46	+	6.34E-05
positive regulation of gene expression	2011	8	.77	10.44	+	6.44E-05
positive regulation of biosynthetic process	2039	8	.78	10.30	+	7.19E-05
regulation of transcription by RNA polymerase II	2294	8	.87	9.15	+	1.84E-04

positive regulation of nitrogen compound metabolic process	3203	8	1.22	6.56	+	2.65E-03
positive regulation of cellular metabolic process	3344	8	1.27	6.28	+	3.74E-03
positive regulation of macromolecule metabolic process	3375	8	1.29	6.22	+	4.02E-03
regulation of transcription, DNA-templated	3467	8	1.32	6.06	+	4.99E-03

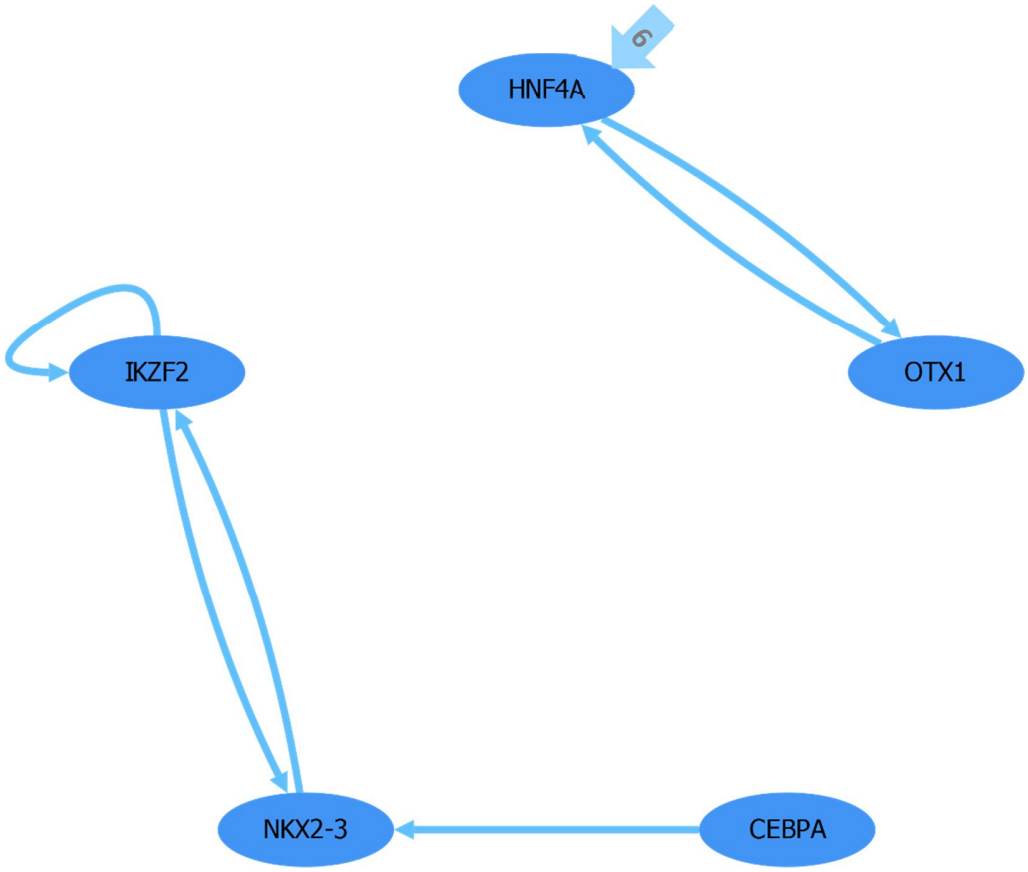

Figure 40. The intra-regulatory network corresponding to Day 13.

Day 22

Just like the previous stage, regulators of Day 22 showed enrichment of general GO terms like regulation of transcription, which is biased in our case due to the restriction of the list to regulators, and no specific terms appeared (Table 11). This indicates that no novel or unique activity happens on Day 22 that stands out from another stage. However, as mentioned before, the expression profile of a cluster of cardiac TFs showed that these factors in it are present on Day 22, but their expression spans from Day 8 till Day 29, but that but not particularly unique for Day 22. The target set of the TFs at this stage showed no significant enrichment at all.

Table 11. The GO terms enriched in the TFs of Day 22.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+ / -	P - value
regulation of transcription by RNA polymerase II	2294	8	1.09	7.32	+	6.75E-03
regulation of transcription, DNA-templated	3467	9	1.65	5.45	+	7.03E-03
regulation of nucleic acid-templated transcription	3534	9	1.68	5.35	+	8.32E-03
regulation of RNA biosynthetic process	3539	9	1.69	5.34	+	8.42E-03
regulation of RNA metabolic process	3785	9	1.80	4.99	+	1.52E-02
regulation of cellular macromolecule biosynthetic process	3908	9	1.86	4.84	+	2.02E-02
regulation of nucleobase-containing compound metabolic process	4040	9	1.92	4.68	+	2.70E-02
regulation of macromolecule biosynthetic process	4046	9	1.93	4.67	+	2.73E-02
regulation of cellular biosynthetic process	4185	9	1.99	4.52	+	3.68E-02
regulation of biosynthetic process	4264	9	2.03	4.43	+	4.33E-02

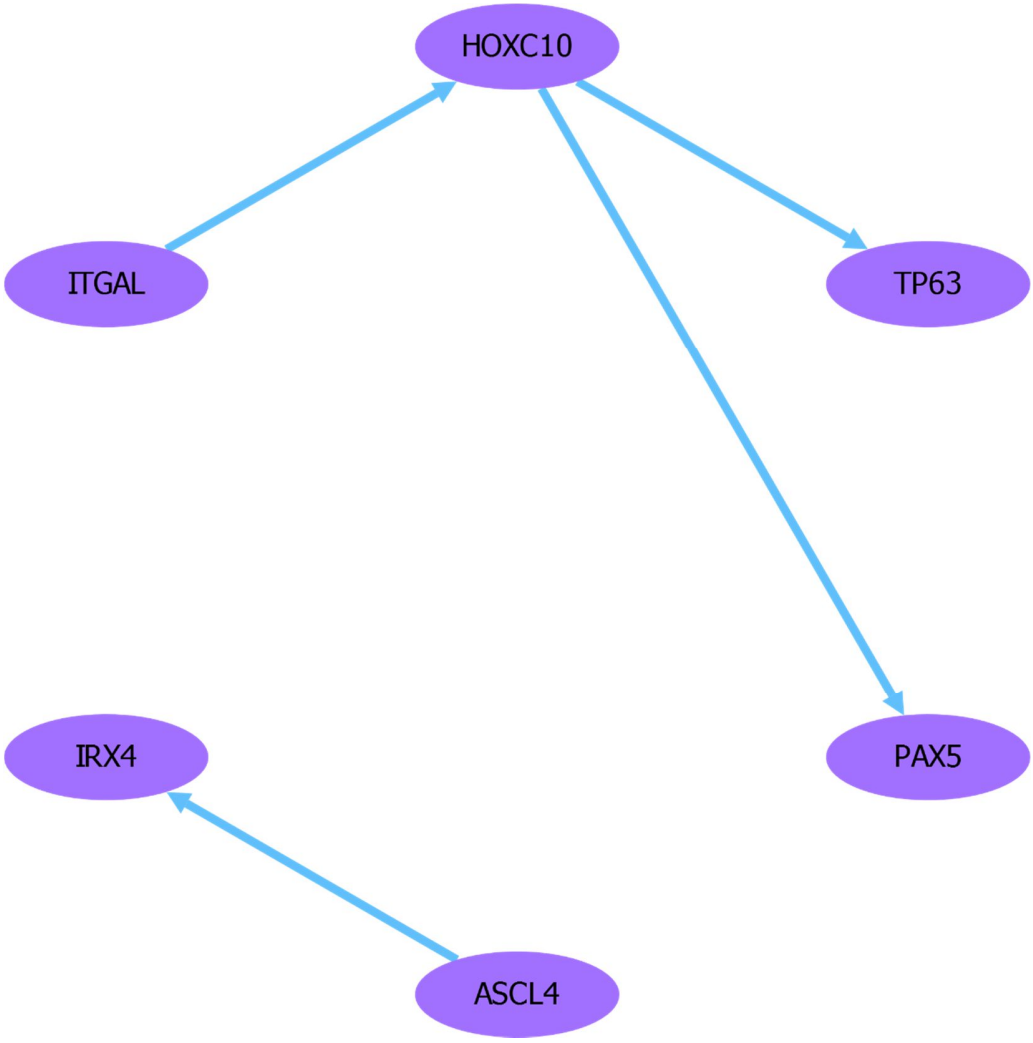

Figure 41. The intra-regulatory network corresponding to Day 22.

Day 29

Just like the previous stage, regulators of Day 29 (Figure 42) showed enrichment of general GO terms like regulation of transcription, which is biased in our case due to the restriction of the list to regulators, and no specific terms appeared. This indicates that no novel or unique activity happens on Day 29 that stands out from another stage. However, as mentioned before, the expression profile of a cluster of cardiac TFs showed that these factors are present on Day 29, but their expression spans from Day 8 till Day 29, but that but not particularly unique for Day 22. However, the target set of the TFs at this stage showed some significant enrichment.

Table 12. The GO terms enriched in the Day 29 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/ -	P - value
negative regulation of transcription by RNA polymerase II	880	6	.38	15.91	+	3.73E-03
negative regulation of transcription, DNA-templated	1208	7	.52	13.52	+	6.19E-04
negative regulation of nucleic acid-templated transcription	1261	7	.54	12.95	+	8.32E-04
negative regulation of RNA biosynthetic process	1263	7	.54	12.93	+	8.41E-04
negative regulation of RNA metabolic process	1354	7	.58	12.06	+	1.36E-03
negative regulation of cellular macromolecule biosynthetic process	1399	7	.60	11.67	+	1.70E-03
negative regulation of nucleobase-containing compound metabolic process	1456	7	.62	11.22	+	2.23E-03
negative regulation of macromolecule biosynthetic process	1481	7	.63	11.03	+	2.51E-03
negative regulation of cellular biosynthetic process	1539	7	.66	10.61	+	3.26E-03
negative regulation of biosynthetic process	1566	7	.67	10.43	+	3.67E-03
negative regulation of gene expression	1738	7	.74	9.40	+	7.49E-03
regulation of transcription by RNA polymerase II	2294	8	.98	8.14	+	1.50E-03
regulation of transcription, DNA-templated	3467	8	1.49	5.38	+	3.83E-02
regulation of nucleic acid-templated transcription	3534	8	1.51	5.28	+	4.45E-02
regulation of RNA biosynthetic process	3539	8	1.52	5.27	+	4.50E-02

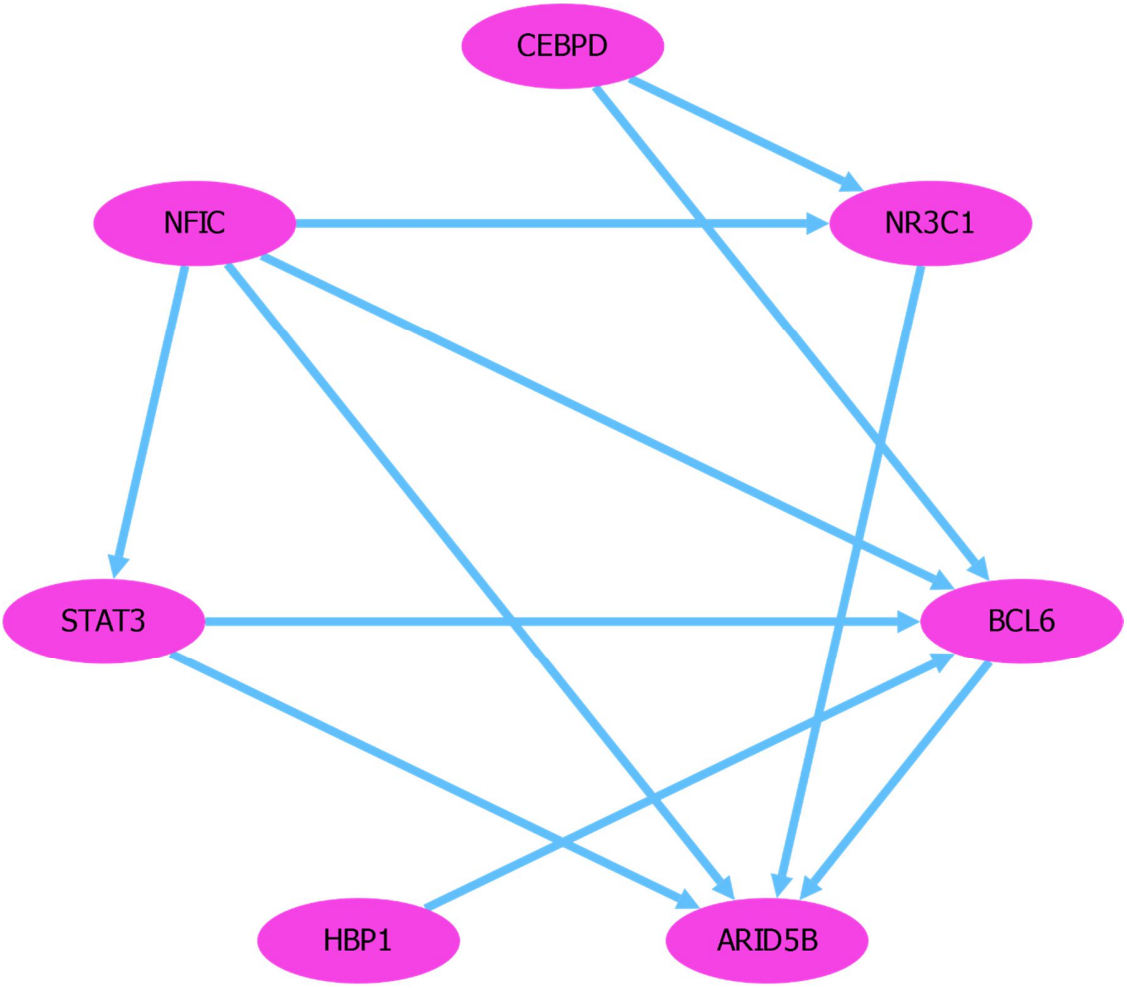

Figure 42. The intra-regulatory network corresponding to Day 29.

Studying the targets of the Day 29 TFs, we queried the regulatory network for targets of these TFs that have a correlation between -0.7 and 0.7 to its regulator and resulted in 466 target genes, a much wider set of correlated targets than in other time points. Despite the fact that the TF set had no specific enrichment terms, their target gene set was enriched for cardiac conduction and heart muscle contraction, a sign of mature cardiomyocytes. This particular set is of interest for the corresponding experiment, as one of the goals was to enhance the contraction strength of the resulting Cardiomyocytes; Thus, overexpressing these TFs might, in turn, raise the expression of the genes in this target set and result in a stronger contraction. However, further investigation needs to be done to distill the target set into those genes involved only in muscle contraction and identifying their regulators. The complete list of the Day 29 specific target set can be found in the Appendix.

Table 13. The GO enrichment of the targets of D29 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/ -	P - value
regulation of striated muscle contraction	94	13	2.09	6.23	+	5.12E-03
muscle organ morphogenesis	82	11	1.82	6.04	+	4.94E-02
muscle cell development	146	17	3.24	5.25	+	9.27E-04
striated muscle cell development	133	15	2.95	5.08	+	7.54E-03
regulation of muscle contraction	163	18	3.62	4.98	+	8.14E-04
regulation of muscle system process	236	24	5.24	4.58	+	2.61E-05
cardiac muscle tissue development	159	16	3.53	4.53	+	1.31E-02
regulation of heart contraction	243	23	5.39	4.26	+	1.92E-04
muscle tissue development	298	27	6.61	4.08	+	2.70E-05
muscle system process	294	26	6.53	3.98	+	8.42E-05
striated muscle tissue development	285	25	6.33	3.95	+	1.89E-04
regulation of blood circulation	290	25	6.44	3.88	+	2.59E-04
muscle cell differentiation	245	21	5.44	3.86	+	3.65E-03
muscle contraction	246	21	5.46	3.85	+	3.88E-03
muscle organ development	295	25	6.55	3.82	+	3.53E-04
muscle structure development	480	39	10.65	3.66	+	1.63E-07

Day 60

This stage has special properties, as its enrichment contains no cardiac relevant terms but neural development terms instead. Enriched terms contained terms such as amacrine cell (inhibitory neuron) differentiation, neurogenesis, and neural system development. This might seem like an error, a methodological problem, or a coincidence; however, the experimental observation showed otherwise. Experimentalists monitoring this particular experiment have observed that after the cardiac cells mature in Day 60, a group of neural cells start emerging in the culture and almost dominate the culture by Day 60. Thus this enrichment is well justified, and the TFs involved in kicking off the neurogenesis were investigated.

Table 14. The GO terms enriched in the D60 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+ / -	P - value
amacrine cell differentiation	5	2	.00	> 100	+	3.83E-02
positive regulation of transcription regulatory region DNA binding	25	3	.01	> 100	+	2.26E-03
regulation of transcription regulatory region DNA binding	53	3	.03	> 100	+	1.90E-02
positive regulation of DNA binding	56	3	.03	> 100	+	2.23E-02
neural retina development	65	3	.03	96.90	+	3.43E-02
retina development in camera-type eye	148	5	.07	70.93	+	4.20E-05
camera-type eye development	317	6	.15	39.74	+	2.25E-05
eye development	357	6	.17	35.29	+	4.53E-05
visual system development	361	6	.17	34.90	+	4.84E-05
sensory system development	366	6	.17	34.42	+	5.24E-05
sensory organ development	553	6	.26	22.78	+	5.93E-04
regulation of transcription by RNA polymerase II	2294	10	1.09	9.15	+	2.22E-06
neurogenesis	1649	7	.79	8.91	+	1.62E-02
nervous system development	2357	9	1.12	8.02	+	2.32E-04

A central node in the intra-regulatory network of Day 60 is PAX6, with three regulators and three targets in the network (Figure 43). PAX6 (Paired Box 6) is a TF that is crucial in the neural cell fate determination and sensory development [148]–[158]. This crucial experimentally verified role of PAX6 goes very well with its central location in the intra-network and well-peaked pattern specific for Day 60 . BHLHE22 is another active hub in the network, is hypothesized for its potential role in retinogenesis, and in this case, that fact that

it is coexpressed with a cluster of neural-specific TFs supports that theory, but still needs further investigation. ZNF536, a heavily regulated node in the network, is known to be expressed in the developing brain and is shown to be a manor repressor of neural development genes [159]–[161]. This indicates that its role in this network is more of a repressive role in order to control the neural development and pace it, as the levels of the known neural TF NEUROD1 increase, it promotes further the expression of ZNF536 which in its turn slows down neural differentiation. NEUROD1 is a TF known to be essential for the survival and maturation of neurons as well as having the ability to reprogram certain reactive cells into functional neurons [162]–[166] and regulates in the intra-network the expression of PAX6 and ZNF536. FOXP2 peaking in Day 60 is a TF known to be involved directly in speech and language skills as well as its involvement in directing neural development [167]–[170] and seems to be one of the direct targets of PAX6. VSX2, another target of PAX6 that co-expresses it is known to be specific for the development of the visual system [171]–[174]. ATOH7, just like NEUROD1 is a regulator of PAX6 and ZNF536, is known for its involvement in photoreceptor development and neurogenesis in the retina [175]–[178]. ATOH7 is a high candidate for being one of the master regulators of neurogenesis, but a further experimental investigation needs to be performed. Three other TFs of Day 60, POU4F2, INSM2, and RORB, don't appear in the intra-network due to their lack of involvement in intra-regulatory interactions. There exists evidence that POU4F2 is involved in fish retina development [179][180], and in this case, we might hypothesize that it is involved in human retina and neural development as well. INSM2 is a tumor repressor, and little is known about its role in neural development; however, a hypothesis can be made about its role being similar to ZNF536 in terms of control of the neurogenesis and pacing the cell growth. RORB is known for its role in photoreceptor development [181][182] and also interestingly associated with the bipolar disease, which we can hypothesize is due to its potential role in neurogenesis [183].

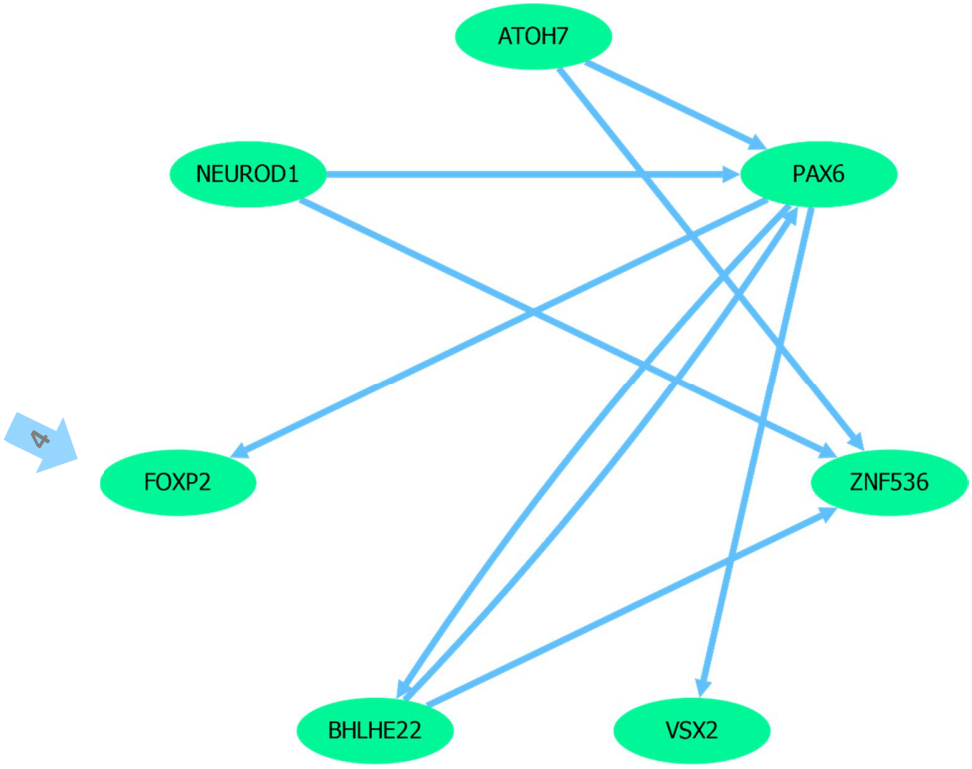

Figure 43. The intra-regulatory network corresponding to Day 60. FOXP2 is heavily regulated by the TFs of the previous stage, with 4 incoming inter-regulatory edges.

Studying the targets of the Day 60 TFs, we queried the regulatory network for targets of these TFs that have a correlation between -0.7 and 0.7 to its regulator and resulted in 193 target genes. These target genes were enriched for terms very specific to neural development, such as the central nervous system and neuron differentiation and neuron projection morphogenesis.

Table 15. The GO terms enriched in the targets of Day 60 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/ -	P - value
cerebellum development	107	9	1.02	8.79	+	1.47E-02
metencephalon development	116	9	1.11	8.10	+	2.75E-02
positive regulation of synaptic transmission	146	10	1.40	7.15	+	2.22E-02
central nervous system neuron differentiation	184	11	1.76	6.24	+	2.42E-02
regulation of synapse structure or activity	235	13	2.25	5.78	+	6.78E-03
neuron projection morphogenesis	494	27	4.73	5.71	+	6.92E-09
gliogenesis	221	12	2.12	5.67	+	2.19E-02
plasma membrane bounded cell projection morphogenesis	498	27	4.77	5.66	+	8.27E-09
cell projection morphogenesis	502	27	4.81	5.62	+	9.88E-09
synapse organization	280	15	2.68	5.60	+	1.39E-03
regulation of synapse organization	224	12	2.14	5.60	+	2.50E-02
axonogenesis	381	20	3.65	5.48	+	1.51E-05
cell part morphogenesis	524	27	5.02	5.38	+	2.55E-08
cell morphogenesis involved in neuron differentiation	443	22	4.24	5.19	+	5.93E-06
positive regulation of neuron projection development	290	14	2.78	5.04	+	1.19E-02

Furthermore, we investigated whether the TFs of Day 60 collaborate in the regulation of the target set, and for that, we utilized the PC-TraFF algorithm. We ran the extended version of the PC-raff algorithm, searching for potential TF collaborations that utilize the PWMs associated with the TFs of Day 60 in the promoter regions of the correlated target set of these regulators. The maximal distance between the pairs was chosen as 20, and the z-score cutoff was 3 in the first attempt, and subsequently the maximal distance was adjusted to 100, and the Z-score cutoff was 3 , yet neither of these attempts leads to the detection of any potentially collaborating PWM pairs.

5.4.2 Multi-Stage Regulators

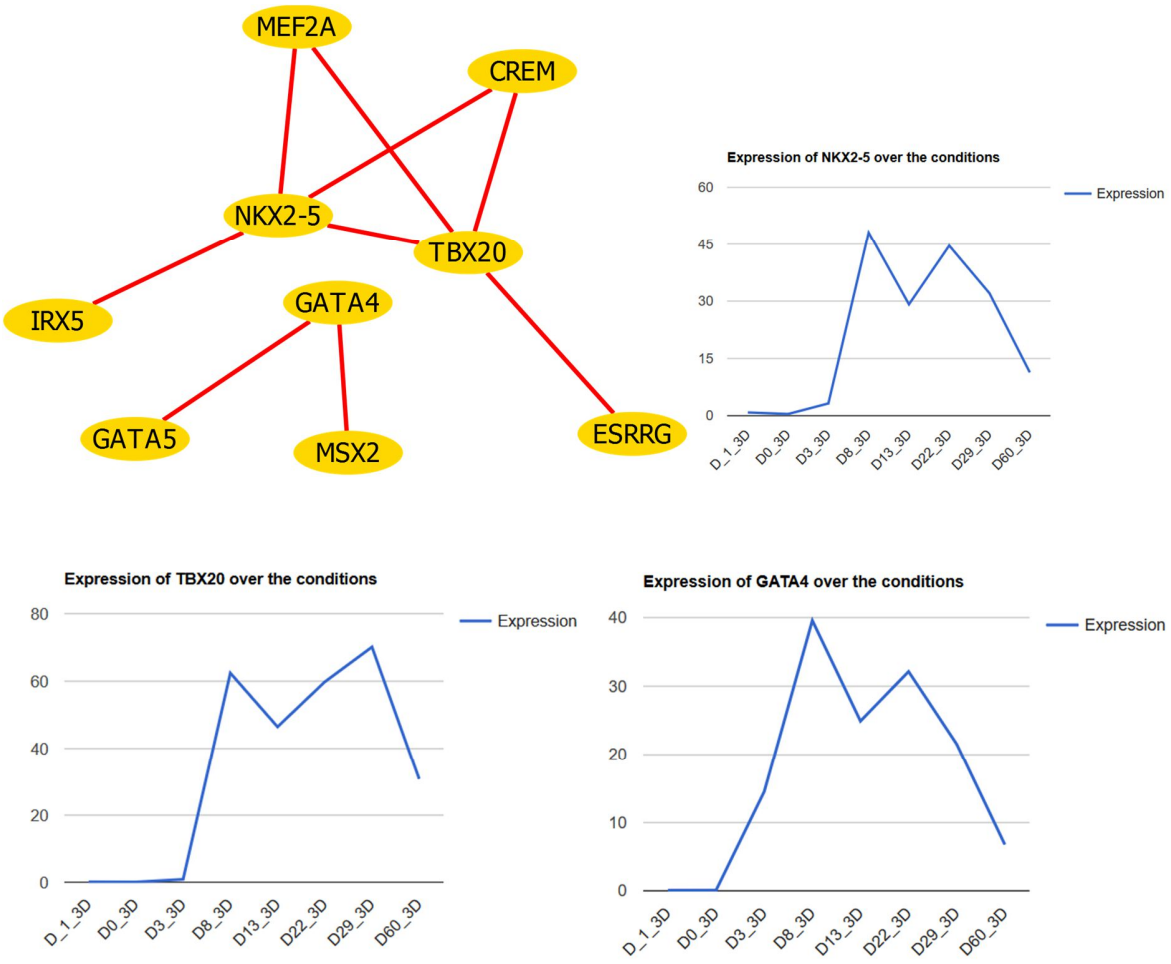

Figure 44. A co-expression cluster of TF genes that are active through the whole process of cardiac differentiation and the expression patterns of some of the important TFs in this cluster.

Using the co-expression workflow, a group of regulators was detected with an expression pattern in the form of a span starting from Day 3 or Day 8 and ending after Day 29 when cardiac maturation has happened. Regulators have different temporal spans during cardiac development in which they are active thus probably used (Figure 44). These temporal spans can give a hint about the type of role these regulators play. In contrary to the peaking TFs detected in the TRC, this group of TFs seems to be important throughout the whole cardiac differentiation process rather than just a specific stage in the differentiation. These TFs can

possibly be responsible for different kinds of processes and tasks that are more general and needed at every stage of the differentiation. For that, we did deeper literature research on the known roles of these TFs and analyzed their target sets.

Some of these regulators, such as TBX20, GATA4, and NKX2-5, are known to be among the handful core regulators that are essential for cardiac differentiation. TBX20 is a well-known conserved TF that plays a fundamental role in cell proliferation and the development of the heart, particularly the cardiac chamber and valve formation [184]–[199]. TBX20 is also proven to act as a dual repressor and activator during the development of the heart [186] [199] as well as directly interacting with NKX2-5, GATA4, and GATA5 [197]. From the expression pattern, TBX20 seems to be not involved in the early cardiac development or the formation of the mesoderm layer as it is not expressed yet in Day 3; however, its expression goes up in Day 8 when the cardiac specification starts and is maintained till the cardiomyocytes are mature. GATA4 is another essential cardiac TF that is required for the formation of the heart tube and the ventral morphogenesis and can cause congenital heart defects when mutated [124], [126], [200]–[207]. GATA4 seems to have an expression head start on the other TFs in this cluster as its expression level starts to increase already at Day 3 then continues to increase peaking at Day 8, maintaining its expression through the cardiac differentiation decreasing again towards the end of the differentiation. This makes GATA4 a potential master regulator and higher in the hierarchy among this group as it is expressed at stage Day 3 when the others are not expressed yet, revealing its independence from the other and even a potential role in activating them. NKX2-5, another core regulator of cardiac development, is detected in this group, is known to regulate early cardiac-specific transcription and other cardiac functions in the adult heart [208]–[212]. NKX2-5 is closer in its activity pattern to TBX20 than to GATA4, associating it to a similar set of processes that TBX20 might be involved in.

5.4.3 Chromatin Modification Analysis

I ran a co-expression analysis on the histone genes in the gene set, and one main cluster peaking at D3 appeared in the analysis (Figure 45). The cluster consisted of: HIST1H3G, HIST1H2BO, HIST1H1D, HIST1H1B, HIST1H2BG, HIST1H2BI, HIST1H3F, HIST1H4C, HIST1H2AH, HIST1H2BF. The GO enrichment of this set, as expected, showed terms related to chromatin accessibility and DNA packaging.

Day 3 is a time point that coincides with the mesoderm induction stage and precedes the early cardiac specification. Thus, a hypothesis can be made about the involvement of such genes in Day 3 in opening the chromatin, making the promoters of wider sets of TFs and target genes accessible for certain TFs to start a regulatory cascade leading to cardiac development. Another explanation of this early activity of these genes could be the necessary DNA packaging needed during cell proliferation.

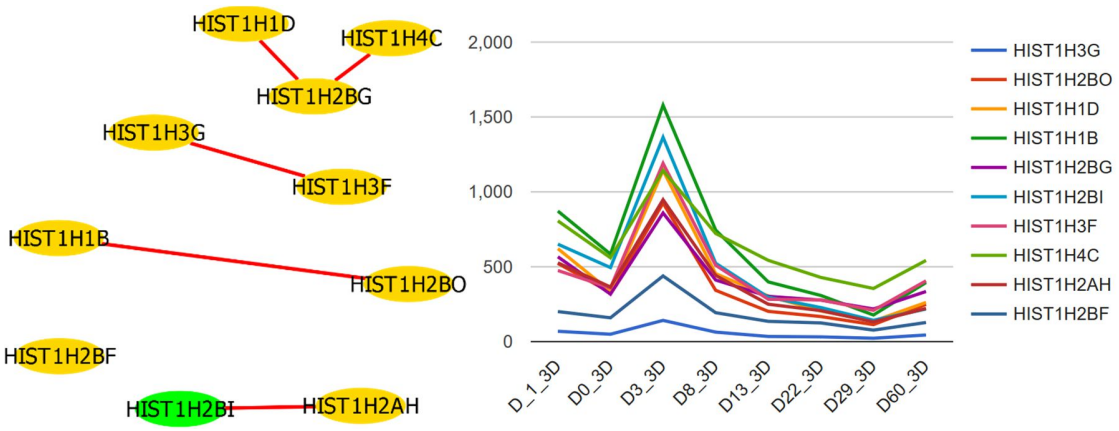

Figure 45. The cluster of histone genes with their expression patterns peaking at Day 3.

Table 16. The GO enrichment of the Histone genes detected in the Day3 cluster.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/-	P - value
nucleosome assembly	116	5	.03	> 100	+	3.12E-07
chromatin assembly	133	5	.04	> 100	+	6.08E-07
chromatin assembly or disassembly	154	5	.04	> 100	+	1.25E-06
nucleosome organization	157	5	.04	> 100	+	1.37E-06
DNA packaging	177	5	.05	98.85	+	2.47E-06
protein-DNA complex assembly	208	5	.06	84.12	+	5.45E-06
regulation of gene silencing	127	3	.04	82.66	+	4.09E-02
protein-DNA complex subunit organization	250	5	.07	69.99	+	1.35E-05
DNA conformation change	291	5	.08	60.13	+	2.85E-05
chromatin organization	695	6	.20	30.21	+	1.21E-05
cellular protein-containing complex assembly	793	5	.23	22.06	+	4.07E-03
chromosome organization	1055	6	.30	19.90	+	1.47E-04

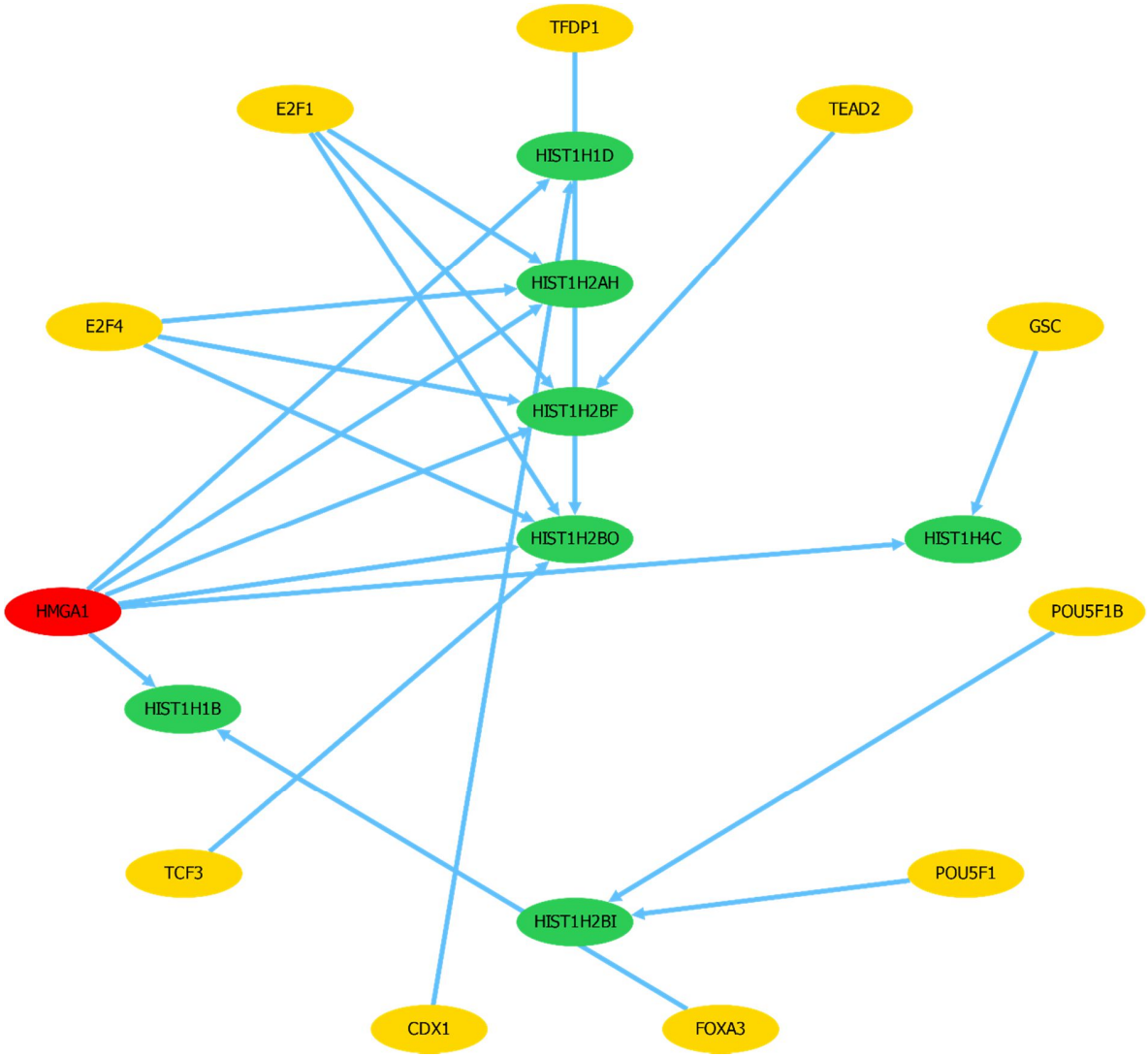

Figure 46. The main regulators that potentially regulate the expression of the histone genes. HMGA1 regulates most of these genes, making it a good potential master regulator.

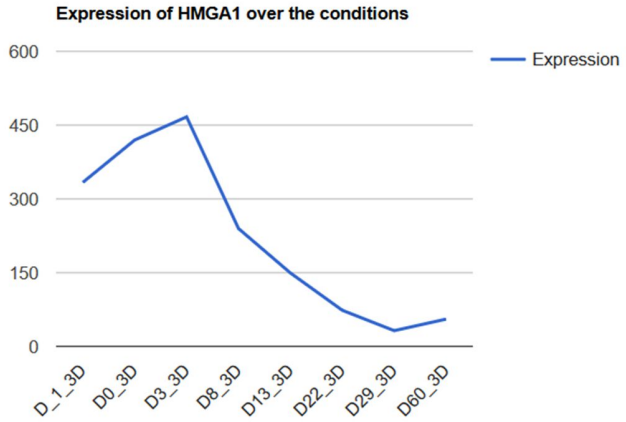

Figure 47. The expression of HMGA1, which is high before the differentiation starts drops as it starts and remains low.

We used the regulatory workflow on the Day 3 histone genes, asking for correlated TFs that regulate at least one gene in the set. HMGA1 stood out as a TF that is able to regulate 6 out of 7 of these histone genes (Figure 46), which can potentially indicate its role as a master regulator or one of the essential regulators necessary for chromatin remodeling via activating histone genes that code for chromatin unwrapping protein. The down-regulation HMGA1 is known to be crucial for chromatin composition in the context of myogenic differentiation [213]. The interaction of the HMGA1a protein with the chromatin is also well documented, as HMGA1 has been found to alter the chromatin structure by interacting with the transcription machinery, resulting in negative or positive regulation of the transcriptional activity of several genes [214]–[216]. Interestingly underexpression of HMGA1 (Figure 47) is also associated with cardiac hypertrophy, which is the abnormal enlargement of the heart muscle, resulting from increases in cardiomyocyte size [217].

5.5 Early cardiac differentiation

5.5.1 TRC analysis

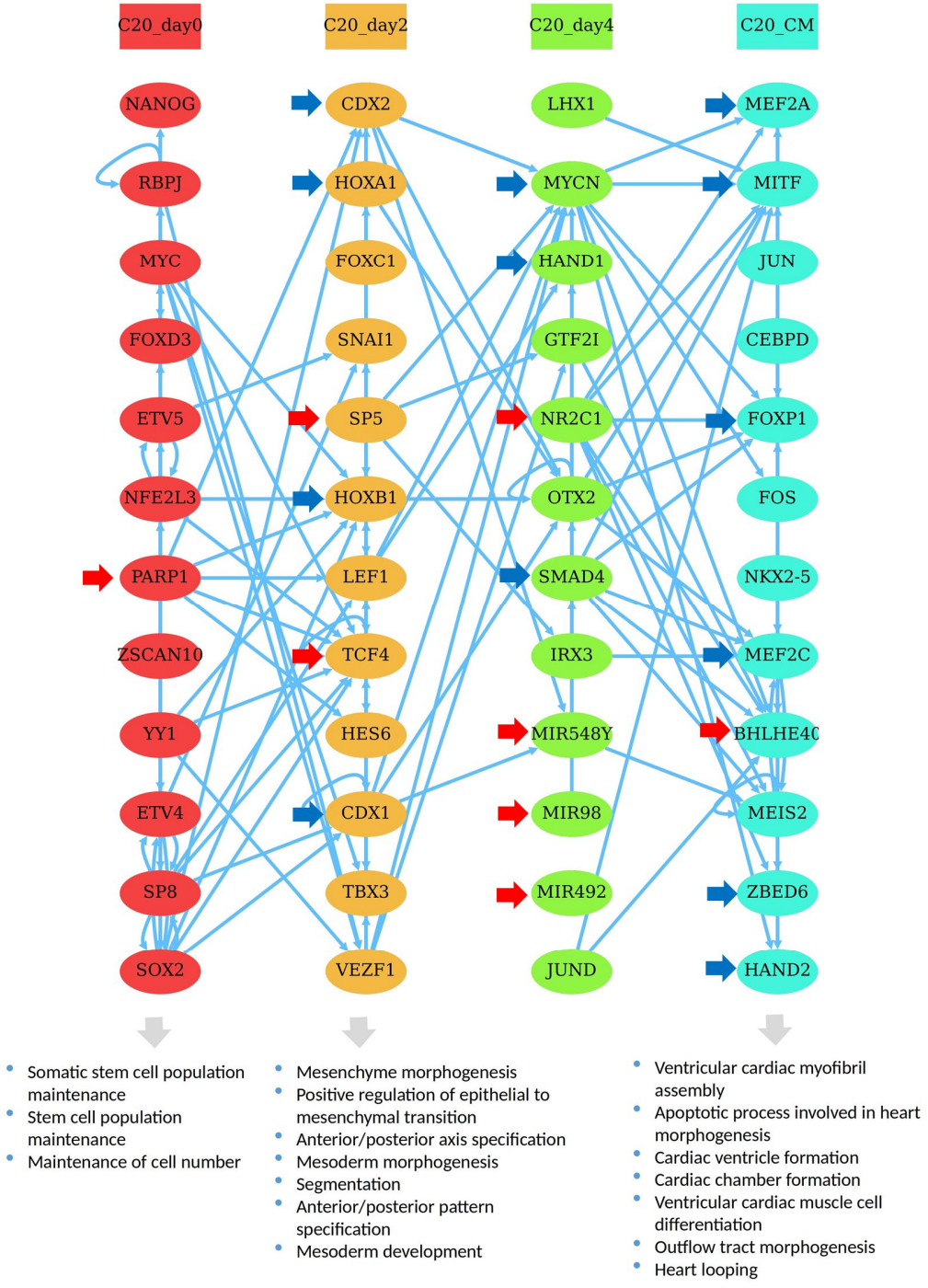

Figure 48. The TRC of the early cardiac differentiation based on the C20 cell line. Below are the main relevant GO terms for three of the stages.

We ran the TRC workflow on the C20 derived cardiomyocytes dataset, which included four time points with two replicates each. The time points were focused on early cardiac differentiation, thus provided a high temporal resolution for the days Day 0 to Day4 followed by a gap after which the mature cardiomyocytes were examined and sequenced. The resulting TRC consisted of four time points, where each time point had 12 associated regulators and the regulatory interactions between them we constructed (Figure 48).

First, we inspected the GO enrichment of the regulators of each stage, picking from the terms that had a p-value>0.05, those with the highest fold enrichment. Regulators of the first time point showed enrichment of terms related to stem cell maintenance, which goes naturally with the biological context since the process of differentiation had not started yet, and the cells are still maintained in the induced stem cell state. These regulators are probably essential for maintaining the pluripotency state and also could potentially be repressing differentiation. Regulators of Day 2 show enrichment of terms associated with mesenchymal and mesoderm morphogenesis, which are the initial steps to giving rise to cardiac cells. Regulators of the last stage, the Cardiomyocyte (CM) stage, show high enrichment of very specific terms that are related to heart development such as cardiac ventricle and chamber formation, ventricular cardiac muscle differentiation, heart looping, and outflow tract morphogenesis. These terms show a high consistency with the underlying stage of differentiation reported by the experiment.

Next, we dived in deeper by inspecting each regulator and regulatory interaction in the cascade looking for literature and experiments that study their involvement in cardiac development.

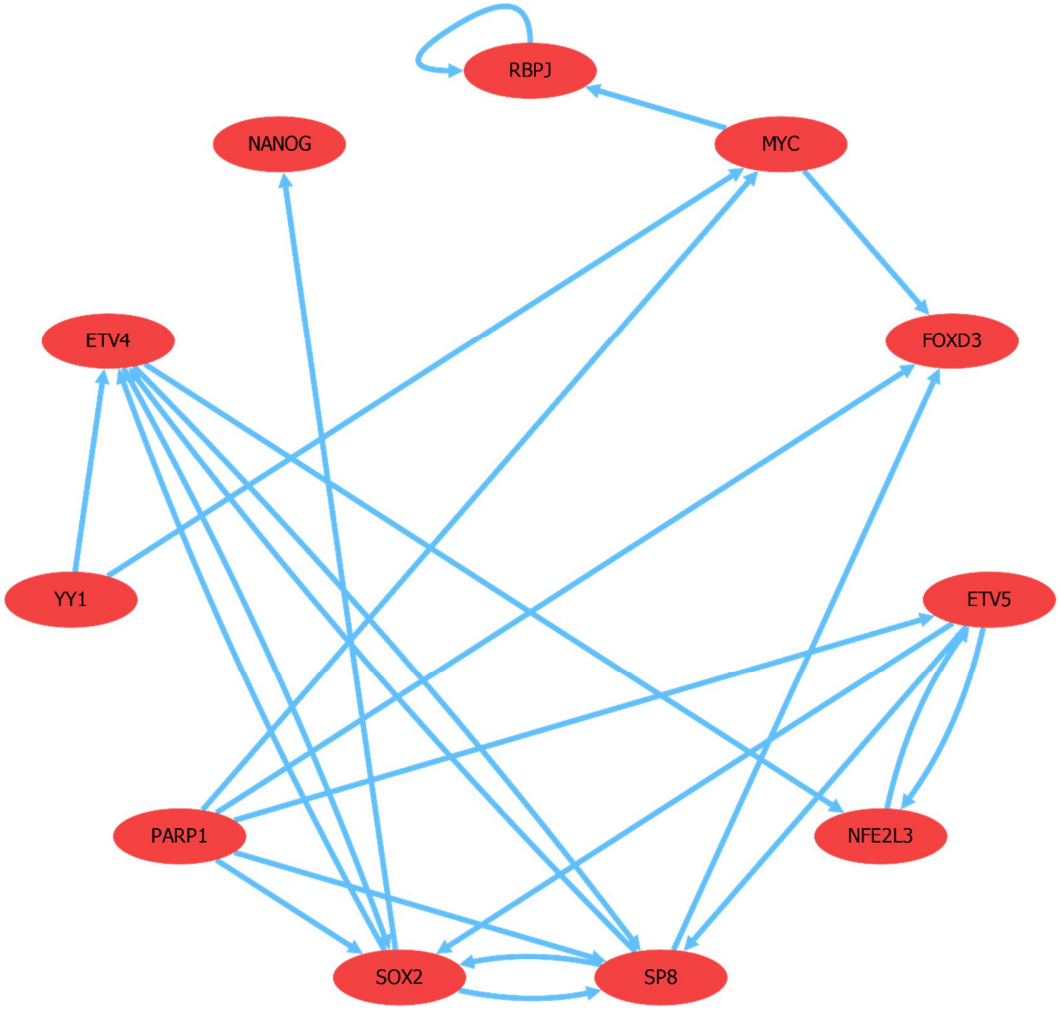

Figure 49. The intra-regulatory network corresponding to Day 0 in the C20 cell line of the early cardiac differentiation dataset.

In the first time point, TFs associated with maintaining the pluripotency state like NANOG[218], PARP1, SOX2 [219] [220], MYC [221], ETV4 and ETV5[222] appear.

CDX1 and CDX2, which are known to modulate early cardiogenesis peak at Day 2, alongside some potentially important early cardiac regulators such as TCF4 and LEF1. On Day 4, MYCN stands out with a high outdegree and indegree confirming its known role in heart development [223] alongside some potential candidate regulators such as LHX1, OTX2, NR2C1, MIR548Y. The last stage where the cardiomyocytes have already matured, features core regulators essential for cardiac development such as MEF2C, HAND2 [224]–[227], NKX2-5 , MEIS2 [228], MITF [229], FOXP1 [230] and some new candidate regulators that could be significant in the cardiac maturation such as MEF2A and BHLHE40.

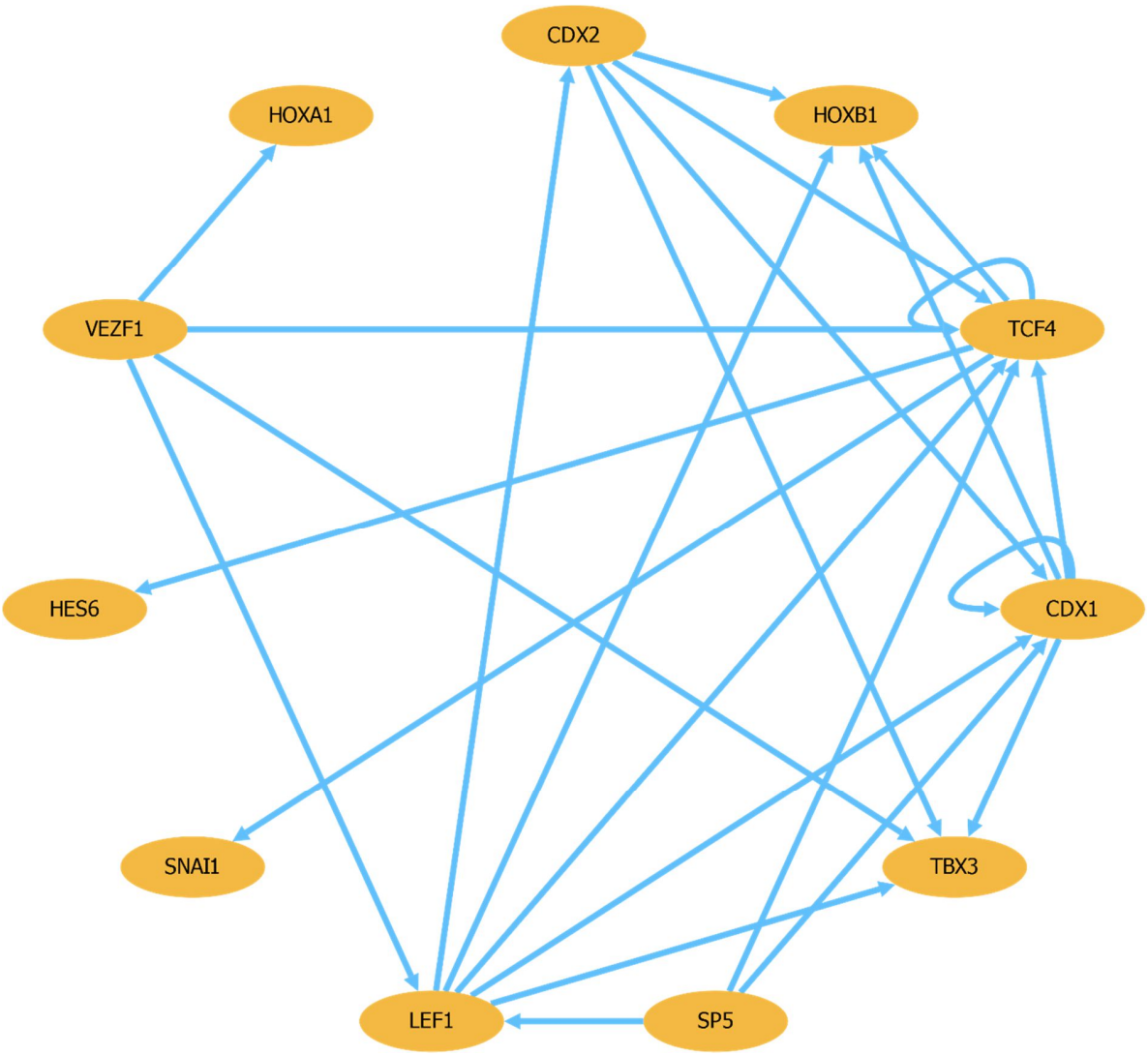

Figure 50. The intra-regulatory network corresponding to Day 2 in the C20 cell line of the early cardiac differentiation dataset.

The intra-network of Day 2 is heavily connected (Figure 50). On the peripherals, SP5 regulates three essential regulators TCF4 CDX1 and LEF1, and VEZF1 regulates four regulators. These 2 factors might be the candidate factors that have the potential to kick off the activation process of the rest of the regulators in the stage. TCF4 is a central node in this network, being heavily regulated and regulating multiple TFs as well as regulating itself. LEF1 also the main hub in this network possibly playing an important role in the regulation of CDX1 and CDX2 s

as well stand out as heavy regulators of this network, conforming with their essential role in early heart development

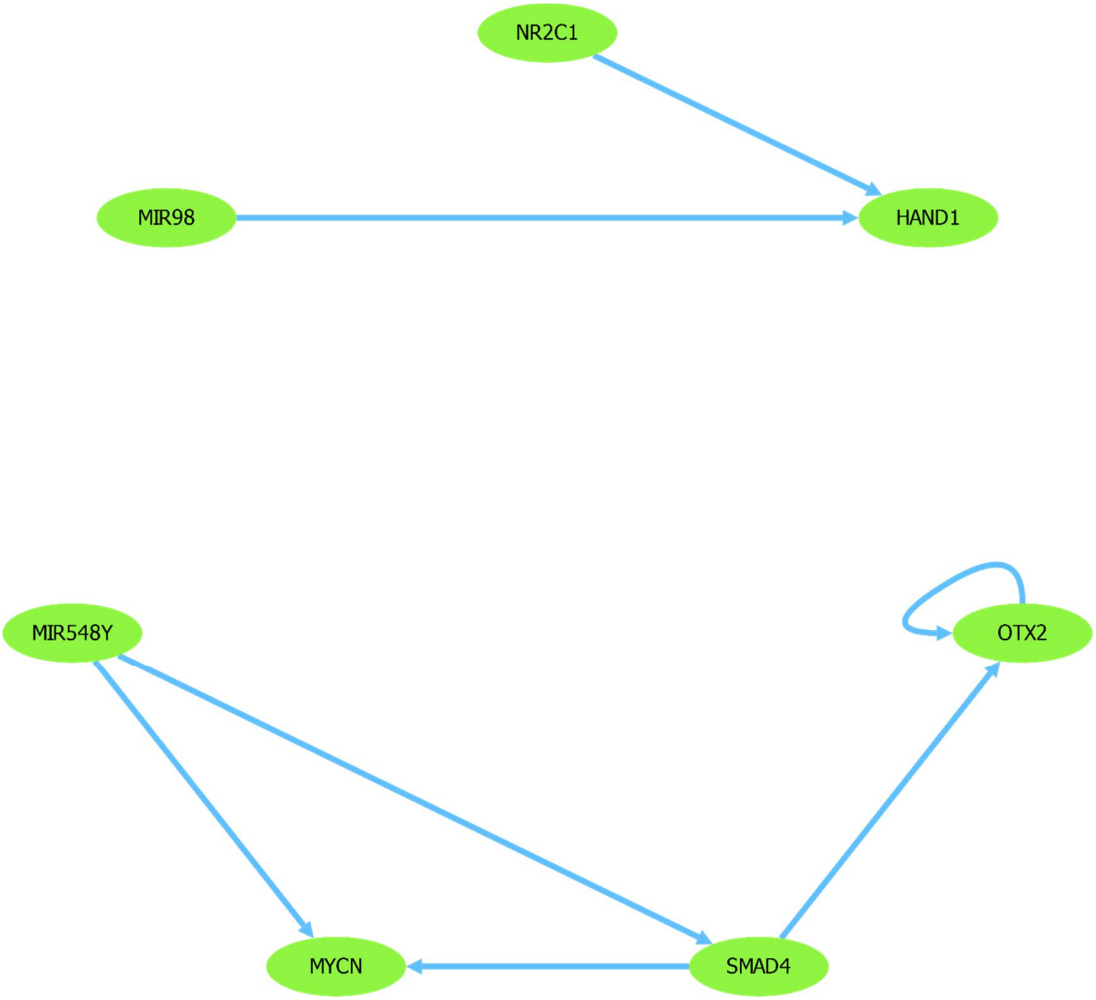

Figure 51. The intra-regulatory network corresponding to Day 4 in the C20 cell line of the early cardiac differentiation dataset.

The intra-network of Day 4 shows a relatively small number of intra-regulations compared to the other stages (Figure 51). Two microRNAs, MIR548Y and MIR98, are predicted to regulate the expression of the essential TFs, MYCN, SMAD4, and HAND1.

On the peripherals of the CM network, we observe NKX2-5, JUN, FOS, CEBPD as regulators that regulate the other same stage regulators but yet not regulated by any of them (Figure 52). We can hypothesize that these TFs are expressed slightly earlier than the others in this stage and might be responsible for activating the remaining stage-specific regulators in the CM

stage. FOXP1 is heavily regulated by the other TFs, and so is HAND2 without regulating any of the other TFs. MEF2A and MEF2C have high betweenness centrality in this subnetwork and tend to regulate many of the same target TFs, suggesting either the redundancy of their regulatory role and the robustness of the network against the deletion of one of such essential central nodes. However, another hypothesis can be made about the potential collaboration between the two TFs and the potential necessity of them both binding to the promoters of the common target TFs for the activation to happen. MEIS2 stands out as a self-regulating TF gene that regulates and is regulated by multiple TFs reinforcing its already know role in development [228].

Figure 52. The intra-regulatory network corresponding to the cardiomyocytes maturation stage in the C20 cell line of the early cardiac differentiation dataset.

5.5.2 MicroRNAs Analysis

We ran a co-expression analysis using only the microRNAs present in the data, to analyze the expression behavior dominant in the microRNAs, which is usually dominated by the expression patterns of other genes and TFs due to the significantly different small number of microRNAs compared to other genes. The co-expression network showed the existence of 4 clusters (Figure 53) with different expression patterns, three stage-specific, and one unspecific (Figure 54). However, the main focus was on the most significant cluster, a cluster that contained 10 microRNAs that are unexpressed except in the end at the CM stage. This cluster we refer to as the CM specific microRNA cluster. While not a lot of information is known about the 10 microRNAs, we hypothesize that they might be involved in cardiac development possibly by controlling the proliferation in the end by repressing the involved genes. Their role might be similar to the roles of some microRNAs that control cardiac hypotrophy such as MIR218 and MIR133 [231][232]. Cardiac hypotrophy is the result of the abnormal enlargement of the heart muscle cells, and those microRNAs control negatively the expression of genes associated with the growth of these cells. Since the microRNAs observed in the cluster are only expressed after the differentiation is finished and knowing that the control on the growth of the cells is exerted by microRNAs, they can be potentially among those microRNAs.

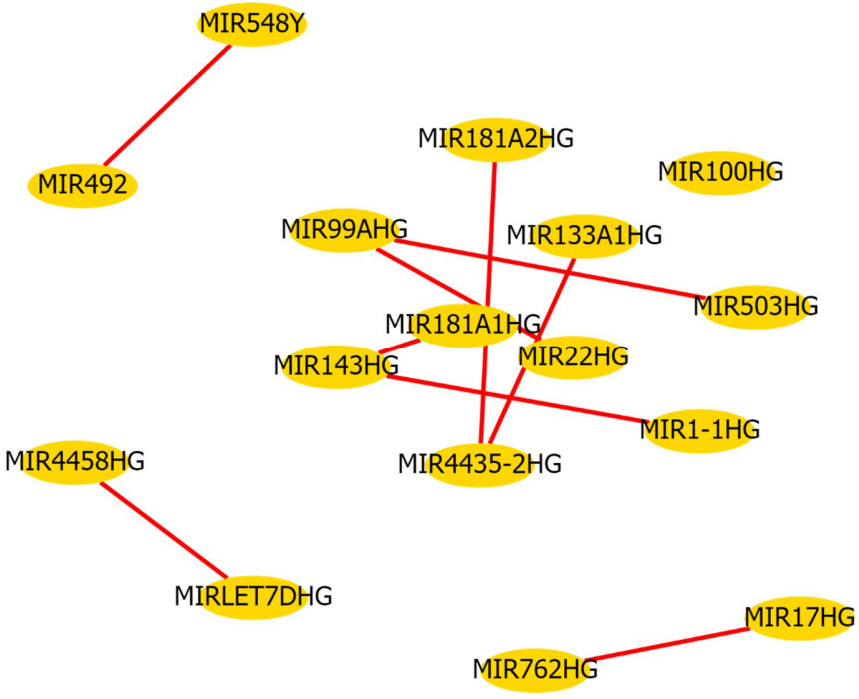

Figure 53. The co-expression network based on the microRNAs present in the early cardiac differentiation dataset. The largest cluster with 10 hubs dominates the network.

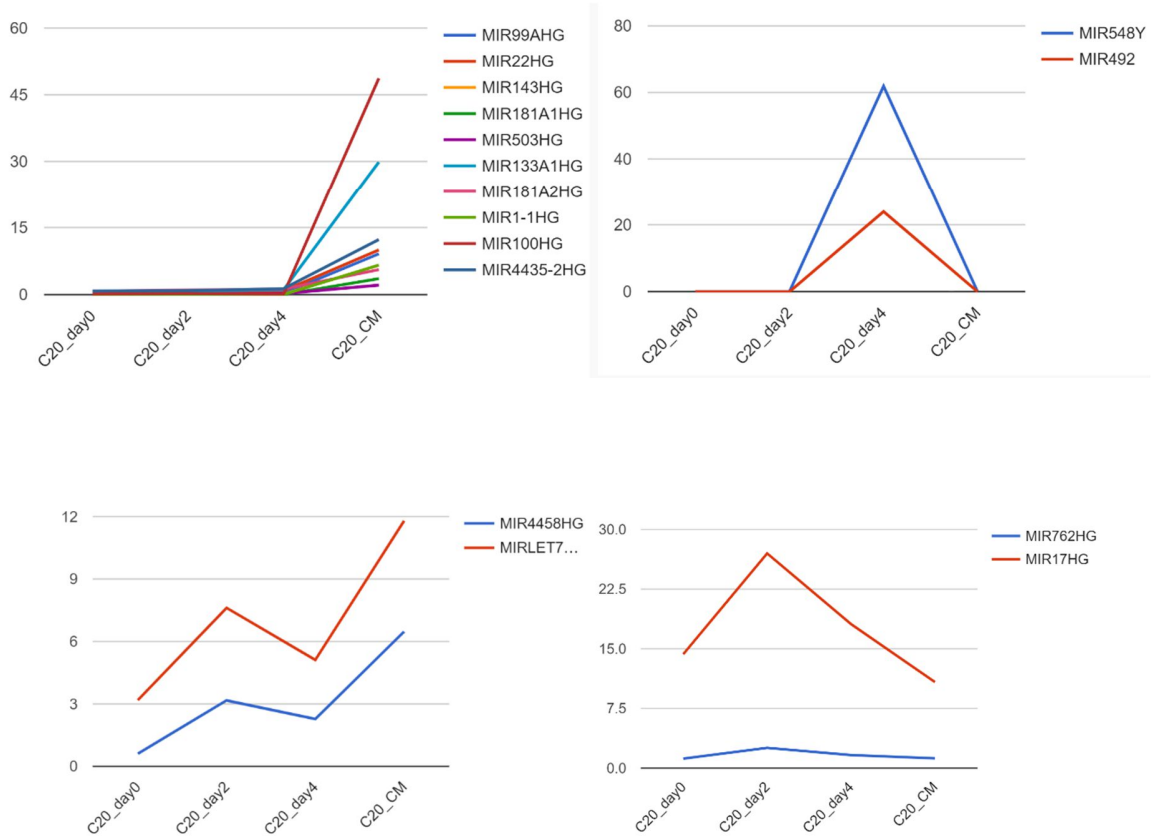

Figure 54. The expression patterns of the microRNAs in the different co-expression clusters.

5.5.3 Collaborating TFs

We ran the extended version of the PC-raff algorithm, searching for potential TF collaborations that utilize the PWMs associated with the TFs of the CM stage in the promoter regions of the correlated target set of the CM regulators. The maximal distance between the pairs was chosen as 100, and the z-score cutoff was 3. Six significantly collaborating PWM pairs were detected (Table 17), among which one pair, the V$CEBP_C - V$CEBPB_Q6, was specific for this gene set. These pairs were either associated with CEBPD collaborating with each other or associated with NKX2-5 and NKX2-6 and collaborating with each other. This indicates the potential usage of CEBPD for proximal binding sites for bigger complexes or the collaboration between NKX2-5 and NKX2-6 to regulate common target sets.

Table 17. The potentially collaborating PWM pairs in the promoters of the target set of the TFs of the CM stage.

Matrix 1	Matrix 2	Z-score	Background Difference
V$CETS1P54_02	V$CETS168_Q6	3.368573889	-0.000133862
V$NKX25_01	V$NKX25_Q6	3.009133379	-0.000393057
V$CETS168_Q6	V$CETS1P54_03	5.451870726	-0.001390073
V$CEBPB_01	V$CEBPB_Q6	3.427644453	-0.001031102
V$NKX25_01	V$NKX25_Q5	3.64831424	-0.000235006
V$CEBP_C	V$CEBPB_Q6	6.791328278	0.00323746

5.6 Neural precursors

5.6.1 TRC Analysis

Figure 55. The TRC based on the neural progenitors' temporal dataset. The main GO terms associated with the two relevant stages are displayed at the bottom

.

Upon the visual inspection of the cascade, we observe a missing time point that is Day 2, indicating that this time point does not have any peak strength or any genes that exceed the certain correlation threshold to the TPP, suggesting that Day 2 might be a time point that doesn't underly any unique stage-specific activity (Figure 55).

Examining the GO enrichment of each time point reveals high enrichment of relevant terms among the regulators active at Day 1 and Day 11. Regulators of Day 1 showed enrichment for specific terms such as cell and neuron fate commitment, neuron differentiation, and cell differentiation in the spinal cord. Regulators of Day 11 showed high enrichment of even more specific terms such as spinal cord association neuron differentiation, dorsal spinal cord development, cell fate determination, cell differentiation in the spinal cord, hindbrain development. On the other hand, examining the GO enrichment based on the DEG analysis publicly available for the same dataset, differentially expressed genes in Day 0 vs. Day 1 and Day 0 vs. Day 11 showed no significant enrichment of specific terms associated with neural development but rather more general terms. Interestingly the expression patterns show that most genes that are peaking in one of these two stages have a slight peak also in the second.

Table 18. The GO enrichment of Day 1 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/-	P - value
neuron fate commitment		68	4 .03	> 100	+	1.41E-04
cell differentiation in spinal cord		55	3 .02	> 100	+	1.48E-02
cell fate commitment		248	5 .11	47.03	+	2.64E-04
neuron differentiation		1012	6 .43	13.83	+	8.46E-03
regulation of transcription by RNA polymerase II		2294	8 .98	8.14	+	1.50E-03
regulation of transcription, DNA-templated		3467	9 1.49	6.06	+	8.26E-04
regulation of nucleic acid-templated transcription		3534	9 1.51	5.94	+	9.80E-04
regulation of RNA biosynthetic process		3539	9 1.52	5.93	+	9.93E-04
animal organ development		3197	8 1.37	5.84	+	2.03E-02
regulation of RNA metabolic process		3785	9 1.62	5.55	+	1.82E-03

regulation of cellular macromolecule biosynthetic process	3908	9	1.68	5.37	+	2.42E-03
regulation of nucleobase-containing compound metabolic process	4040	9	1.73	5.20	+	3.26E-03
regulation of macromolecule biosynthetic process	4046	9	1.73	5.19	+	3.31E-03
regulation of cellular biosynthetic process	4185	9	1.79	5.02	+	4.48E-03
regulation of biosynthetic process	4264	9	1.83	4.92	+	5.30E-03
system development	4424	9	1.90	4.75	+	7.38E-03

Table 19. The GO enrichment of Day 11 TFs.

GO Term	Nb. in Reference	Nb. in upload	Nb. Expected	Fold Enrich.	+/-	P - value
spinal cord association neuron differentiation	14	3	.01	> 100	+	4.71E-04
dorsal spinal cord development	22	3	.01	> 100	+	1.59E-03
cell fate determination	42	3	.02	> 100	+	9.76E-03
cell differentiation in spinal cord	55	3	.03	> 100	+	2.12E-02
neuron fate commitment	68	3	.03	92.63	+	3.91E-02
retina development in camera-type eye	148	4	.07	56.75	+	4.79E-03
eye morphogenesis	151	4	.07	55.62	+	5.18E-03
hindbrain development	153	4	.07	54.89	+	5.45E-03
central nervous system neuron differentiation	184	4	.09	45.64	+	1.12E-02
cell fate commitment	248	5	.12	42.33	+	5.23E-04
visual perception	220	4	.10	38.17	+	2.25E-02
sensory perception of light stimulus	223	4	.11	37.66	+	2.37E-02
camera-type eye development	317	5	.15	33.12	+	1.74E-03

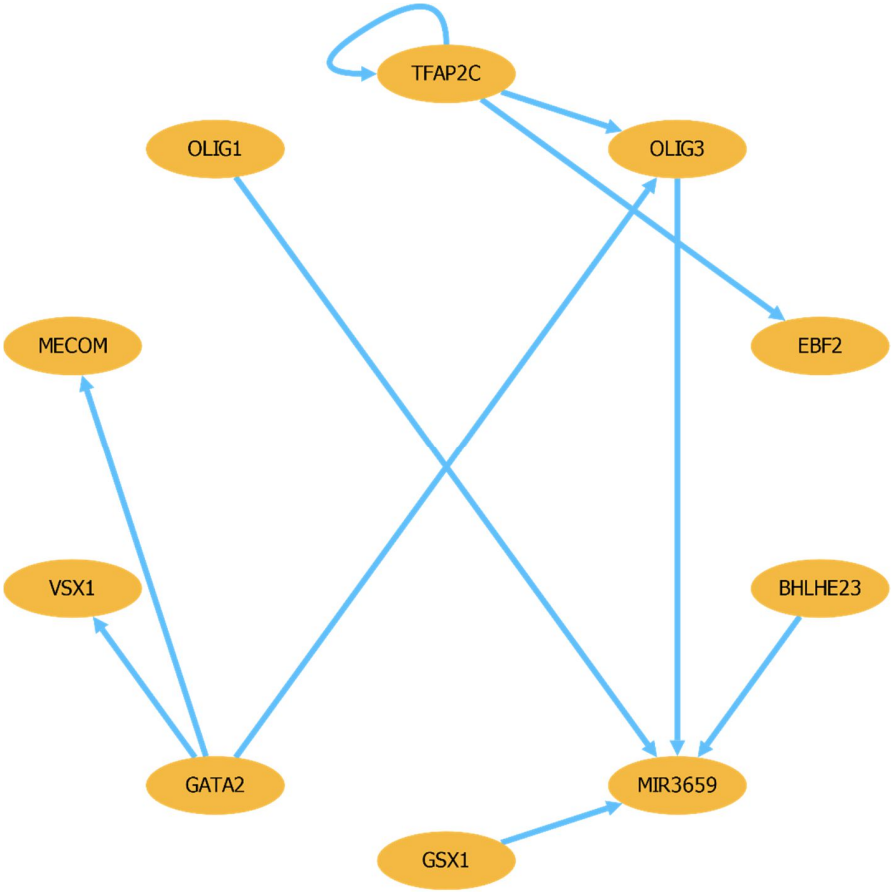

Figure 56. The intra-regulatory network of Day 1.

A deeper look into the intra-regulatory network of Day 1 (Figure 56) shows OLIG1 and OLIG3, which are known for their importance in neural and spinal development [233]–[235], suggesting that their importance lies in the earlier part of the differentiation. A microRNA, MIR3659, peaking at Day 1 is heavily regulated by four TFs in the intra-regulatory network, which raises the question of the nature of its involvement in neural differentiation, which needs to be further investigated. GATA2 is a well-known neural TF proven to control the proliferation of the neural progenitors and drives them into differentiation [236]–[241]. GATA2 regulates three other TFs in the network VSX1, MECOM, and OLIG3, which makes it a candidate to be a master regulator of that stage. VSX1 is known to have an essential role in the retina spinal cord and brain development [242]–[245], and here we observe its importance in the earlier stages of neural development, which is consistent with the evidence of its role in regulating prechordal mesendoderm. MECOM, BHLHE23, GSX1and EBF2 are not well studied or annotated in the context of neural development; however, their pattern and association with the other TFs of this stage suggest its candidacy for being important neural TFs and can be a good set to test experimentally in the context of neural differentiation. TFAP2C, an active node in the Day 1 network regulating 2 other TFs as well as itself, is a known neural and developmental TF that regulates transcription by opening enhancers [137], [246]–[248].

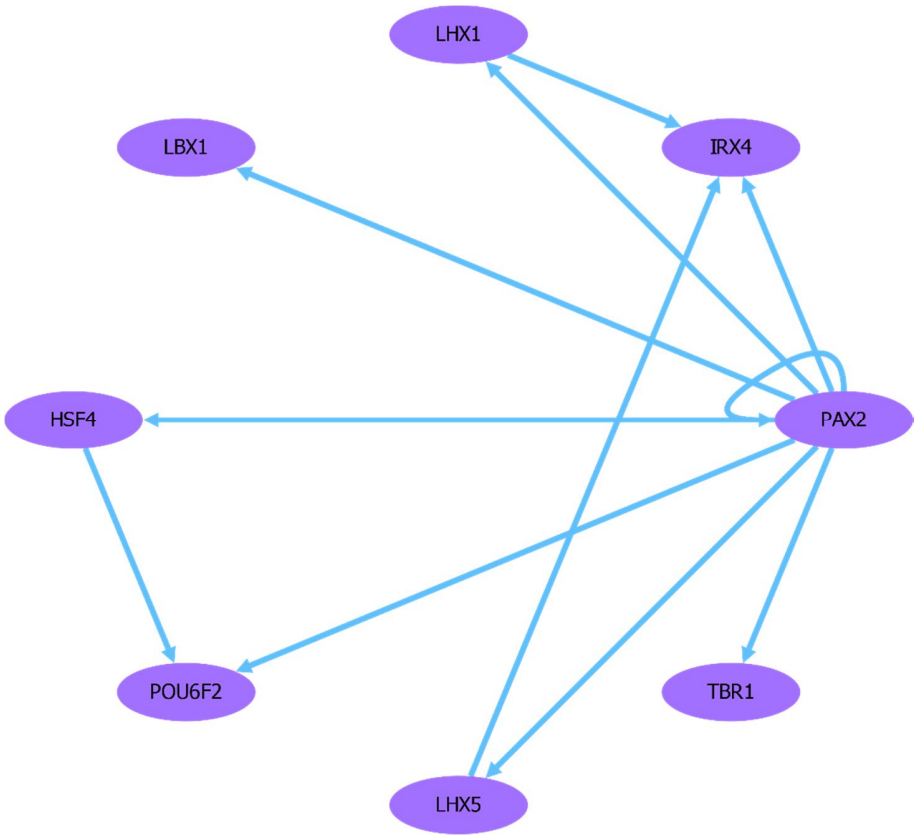

Figure 57. The intra-regulatory network of Day 11.

PAX2 in Day 11, with the highest outdegree, regulates 13 different regulators in the same and next time point which hints that its known essential role in neural development [249][250][251]–[255] is due to its regulatory impact on a big set of neural regulators (Figure 57). LHX5 and LHX1 are both associated with neuronal differentiation, particularly in the forebrain [256]–[263]. POU6F2 is associated with renal and neuronal development [264]. HSF4 is a regulator involved in the mammalian lens development [265]; however, its existence in this neural network suggests that its role in the lens development might be in developing the neural connections needed for the lens. TBR1 is another neural regulator that is associated with diseases such as autism and proven to regulate neural stem cell fate, which explains the fact that it is heavily regulated by the other TFs. LBX1 is known to be important in the neural patterning process [266]–[269].

KLF2 in Day 5 stands out as a significant potential regulator of the Day 11's regulatory wave due to its potential ability to regulate a big portion of Day 11 regulators.

The TRC shows an overall same-stage presence of certain TFs that belong to the same family or subfamily according to the classification of TFs in TFClass, such as OLIG1, OLIG3, and

BHLHE23 in Day 1, STAT1 and STAT6 in Day 5, the LHX1 and LHX5 in Day 11 , DBX1 and DBX2 in Day 18 . A hypothesis can be made that these TFs are part of the redundancy that leads to the robustness of such regulatory programs, or that these families and subfamilies of TFs collaborate in certain regulatory stages.

5.6.2 MicroRNA Analysis

A co-expression analysis was performed on the microRNAs in the dataset, and a large cluster along with some smaller clusters was observed, but they all had comparable patterns Figure 58. Almost all microRNAs peaked up at Day 1 and Day 11, which coincides exactly with the two stages where the stage-specific TFs showed a significant neural enrichment. The main cluster has two sharp peaks, while the others had a pattern that showed an increase that started already on Day 5 in the second wave Figure 59.

Again in this dataset, as we observed in the early cardiac differentiation dataset, a wave of microRNAs accompanies and goes in parallel with the wave of differentiation TFs, possible to titrate and control the TF activity.

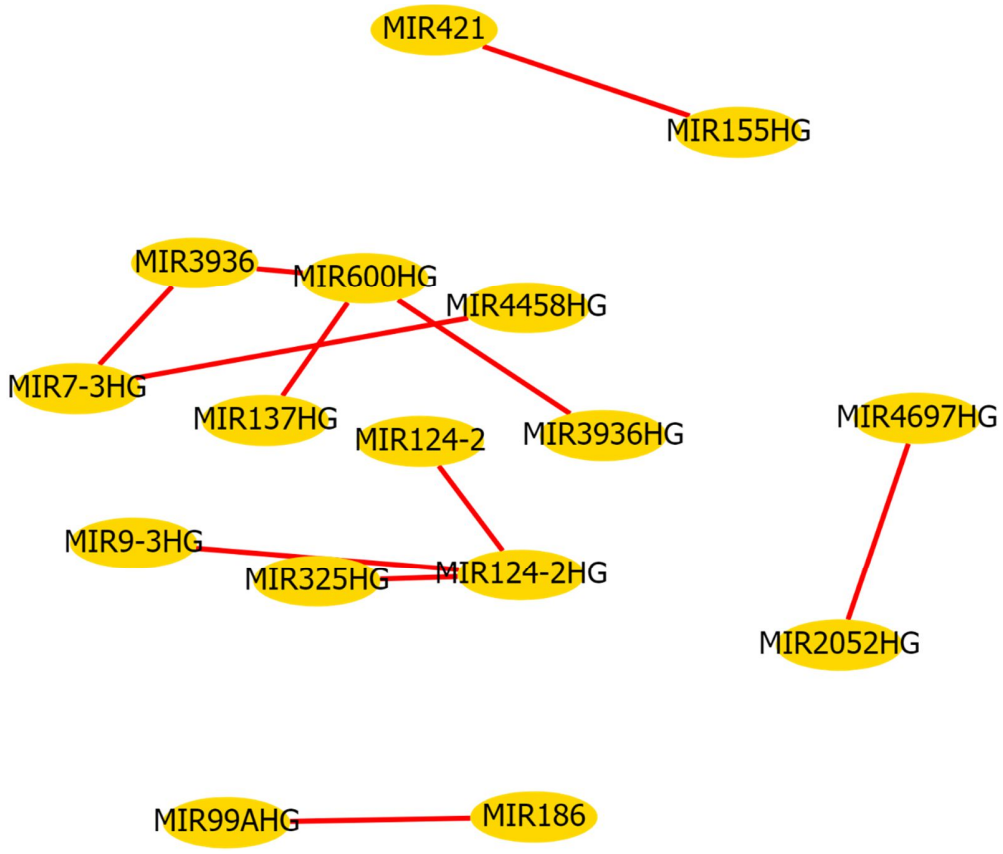

Figure 58. The co-expression network based on the microRNAs in the dataset.

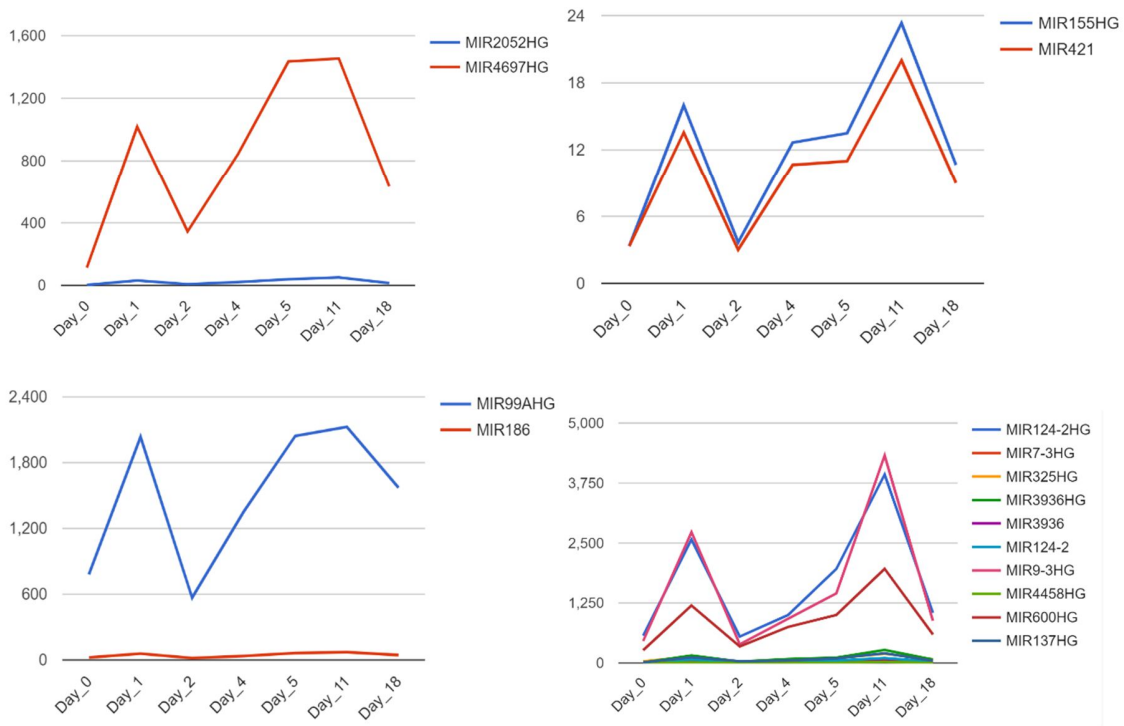

Figure 59. The expression patterns of the microRNAs in the different co-expression clusters.

6 Discussion

6.1 Sampling flaws and solutions

The TRC model is heavily dependent on the choice of time points in the experiment. The model performs optimally and as intended when each underlying stage, for example, a differentiation stage, is represented by one and only one time point in the data (Figure 60). In many good experimental designs, this condition is satisfied; however, some other experiments generate temporal datasets that suffer from over or under-sampling. These datasets, if untreated and analyzed through the TRC workflow, can generate a TRC that is flawed, or missing information that could have been otherwise recovered. This section tackles the problems associated with these datasets and devises methods and modifications to the original method that can be adapted for these cases.

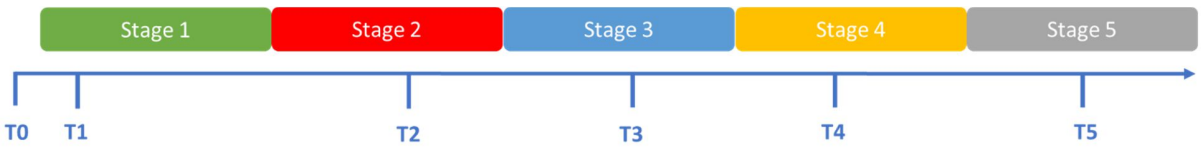

Figure 60. A good choice of sampling which generates an optimal temporal dataset for the TRC analysis. Each biological stage is represented by one and only one time point.

In the under-sampling scenario, the TRC model is suspectable to generate some false negatives and some false positives in the stage-specific regulators' sets. False negatives appear in the form of genes that were stage-specific however are not identified and false positives appear in the form of genes that are not stage-specific in reality; however, they appear in a stage-specific gene set. Figure 61 illustrates this problem, where the expression of a gene in the dataset doesn't reflect the complete expression pattern of the gene in reality. Stages that were not sampled and no time points are associated with, will be missed. If a gene is highly expressed or peaking in that missing stage, it will be missed by the TRC model leading to false negatives. False positives occur when a gene peaks at the missing stage and another stage, however only one peak is captured as the other peak is missing from the data, thus identified as stage-specific when it is not.

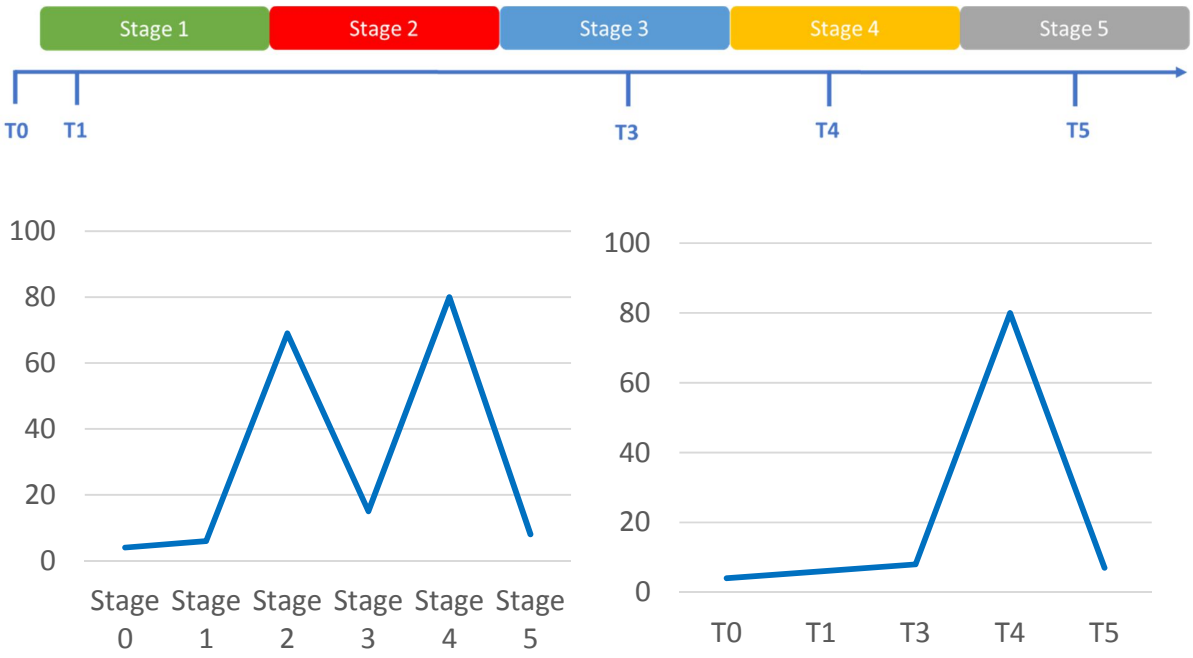

Figure 61. An example of the under-sampling problem. Although Stage 2 is a distinct biological stage, it has no associated time point. The pattern on the left shows the real expression pattern of a gene that is not stage-specific. In reality, it is peaking in both stage 2 and stage 4; however, due to the under-sampling, only its peak in Stage 4 is captured as T4 and is assumed to be stage-specific.

Although the intuition says that more data points mean more information and better results, this is not the case in the TRC model, particularly when the samples are not annotated properly. In the over-sampling scenario, the TRC model is suspectable to generate some false negatives. Figure 62 illustrates an example of this problem, where two samples are taken from the same stage and yet annotated as two distinct time points and not replicates. A gene can be stage-specific and have a peaking pattern in reality; however this peak is destroyed by redundant time points as it will peak across two consecutive stages, rendering its peak strength too weak to pass the threshold for a stage-specific set

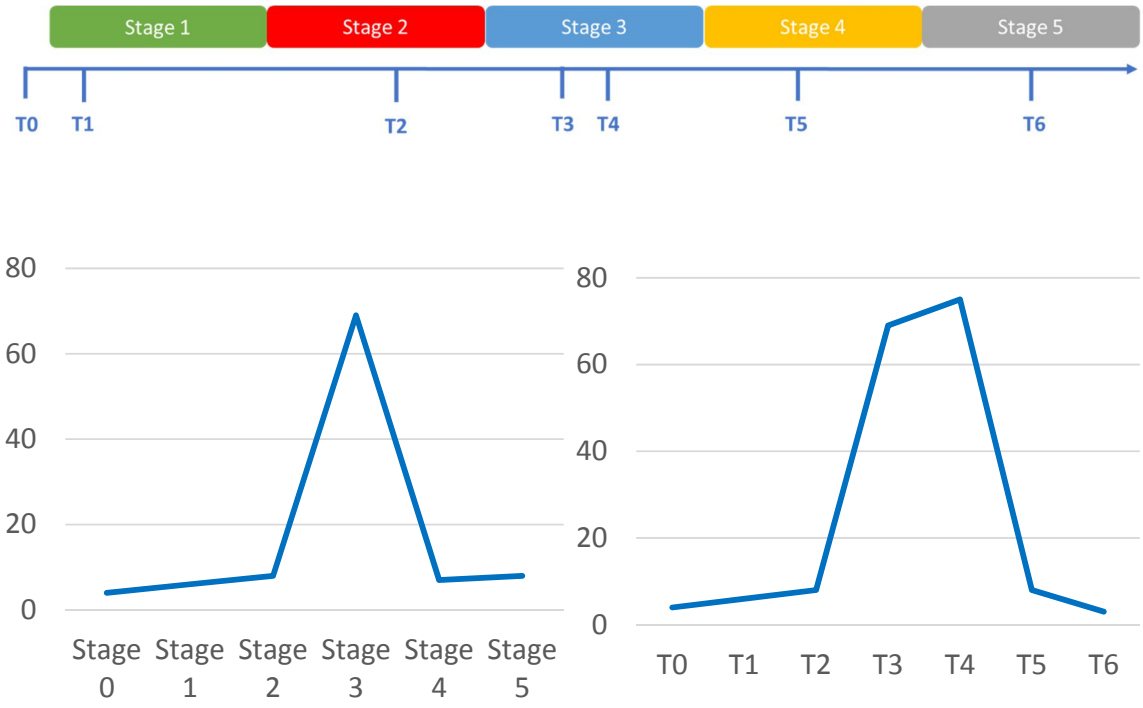

Figure 62. An example of the over-sampling problem. Two redundant time points are sampled from the same biological stage and not denoted as replicates. The pattern on the left shows the real expression pattern of a gene that is stage-specific, associated specifically with stage 3 in reality. However, due to the redundancy between T3 and T4, its pattern appears to be non-peaking and is assumed by the TRC model to be not stage-specific.

To deal with this particular problem, we devised several solutions:

1. **Time point deletion:** Keeping one of these time points and deleting the other redundant ones, can eliminate the noise caused by such redundancy, but the challenge lies in detecting which time points are the redundant ones and which one to keep out of them. These time points can be decided based on two approaches:

First Approach:

a. Delete time points one by one and calculate the TRC.
b. Calculate the PSC for each round of deletion.
c. The time point that leads to the max PSC when deleted is the potential redundant one.

Second Approach:

a. Calculate the correlation of the gene expression vector between each time point and the other.
b. Time points that show a high correlation are often redundant in their gene expression vector, and one of them is to be eliminated.
c. The time point that shows the higher average correlation with the rest of the time points is the candidate to be removed among the two.

Although several time points can be detected and deleted with these approaches, in practice, a time series dataset typically contains an average of five time points, and the deletion of more than one can lead to scarcity in the time points decreasing the information value of the resulting TRC. It is advisable to restrict the elimination to only one time point in medium-sized time-series datasets (58 time points) and increase the elimination size as the size increases.

2. **Using a multiple-time point peak TPP:** Instead of using the conventional one-peak pattern, using a TPP where the peak extends through several consecutive time points that represent one stage can overcome the redundancy issue (Figure 63).

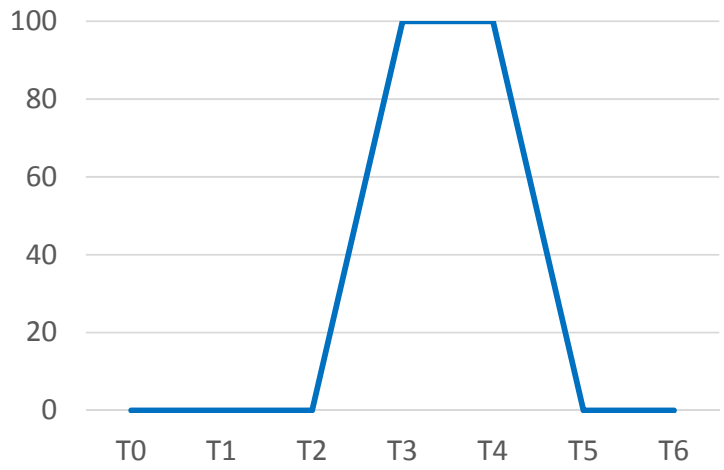

Figure 63. A multiple-time point TPP. The peak spans across T3 and T4 which are two time points corresponding to the same stage.

Determining which time points the peak should span across can be determined by two approaches:

First Approach:

a. Create a library with one TPP peaking at two consecutive time points, and the other TPPs are single peaks for the other time points.
b. Calculate the resulting TRC based on these libraries.
c. The Library that generates the TRC with the highest PSC is the one to use.

Second Approach:

a. Similar to the second approach in the time point elimination method.
b. The correlation is calculated between each pair of consecutive time points, and those with the highest correlation create a subsequent multiple peak TPP.

3. **Analyzing the experimental procedure:** While this approach is not computational, it takes into consideration the underlying case under study. A scientist can judge based on the observations and markers detected in each time point whether there is a redundancy in any time point and if there is a need to remove any of it. This approach can always be coupled with any of the above approaches to enhance the decision making leading to the generation of a more meaningful cascade.

6.2 Emerging properties and patterns

Specific regulatory patterns emerge in the TRC model due to the nature of the methodology. These patterns reflect particular regulatory modes, and a glimpse of the relative positions of connected nodes can give an idea of the nature of the underlying regulatory interaction.

Unlike some of the classic regulatory models, the TRC model takes advantage of the sequential order of the time series data to allow more intricate interpretations of regulatory interactions.

One property emerging from the peaking patterns is that each node in the cascade is positively correlated in its expression pattern to the other nodes in the same stage as they share a strong positive correlation to the same TPP, making their peaking profiles correlated as well. This means that each intra-regulatory edge is coupled with a positive correlation. Figure 64 is an example of such a regulatory interaction indicating that X is potentially one of the activators of Y and contributes to its peaking pattern, where Y is inactive where X is inactive and activated when X is activated (stage i), coupled with the fact that X can bind to the promoter of Y, this hypothesis of the regulatory influence of X on Y is strongly enforced.

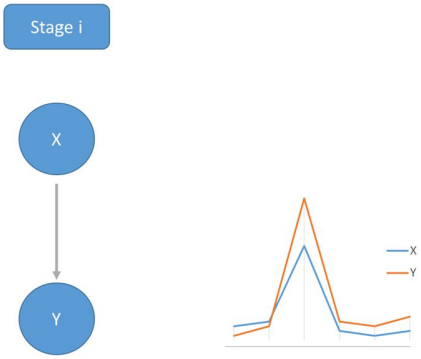

Figure 64. A one-way regulatory prediction within one stage coupled with a high positive correlation.

Figure 64 has a one-direction property that supports the causality, whereas cases such as the double edge displayed in **Error! Reference source not found.** cannot decisively assert whether X is an activator of Y or the other way around due to the non-causal nature of correlation and the double potential of these regulators to bind in each other's promoters.

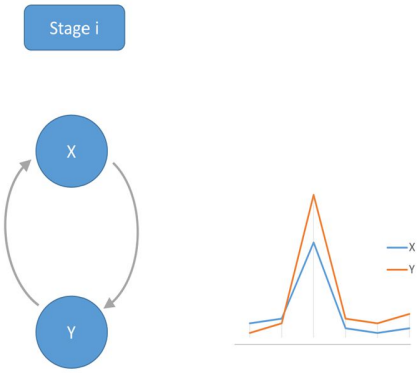

Figure 65. A two-way regulatory prediction within one stage coupled with a high positive correlation.

Another property emerges from the fact that each TPP in a library is time-lagged correlated with every other TPP in the same library. This makes every node in the cascade, which is based on a TPP, correlated via a time-lagged correlation to the nodes in the other stages. This means that every inter-regulatory edge represents a highly conserved binding site prediction of the source in the promoter of the target, coupled with a time-lagged correlation where the source regulator needs a certain time to reach a certain threshold of expression before it can activate the target. Figure 66 is an example of where a regulator in a certain stage potentially needs more time to activate the target; thus, the target is activated after a time lag and captured in the next stage. This time delay in the regulation process is studied in some cases and fits in many biological contexts including cell differentiation [270].

Thus, each edge in the cascade is always coupled with a correlation between the expression pattern of the regulator and its target. This coupling can be viewed as a reinforcement of the regulatory interaction predictive quality and gives it an edge over interactions based solely on the binding site analysis or solely derived from gene expression data. From another view, the binding site prediction behind the edge can explain the perceived correlation in the expression patterns between the target and the source.

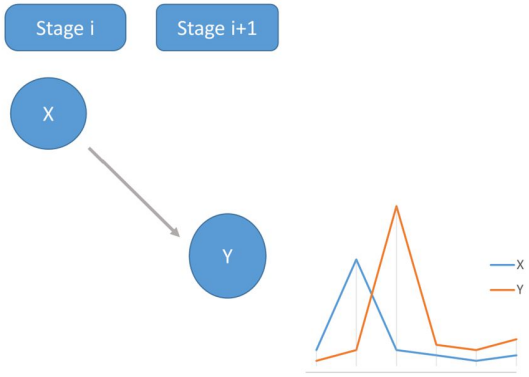

Figure 66. A regulatory interaction from one stage to the next, coupled with a high positive time-lagged correlation.

Another common hypothesis that surrounds co-expressed genes is that they might be co-regulated by a master regulator or a set of master regulators. Some of these master regulators can be captured through configurations in the cascade where a regulator emerging in a stage single-handedly has the potential to activate a wide set of correlated regulators, whether in the same stage as in the case of Figure 67 or a set of targets in the next stage via a time-lagged regulation as shown in Figure 68.

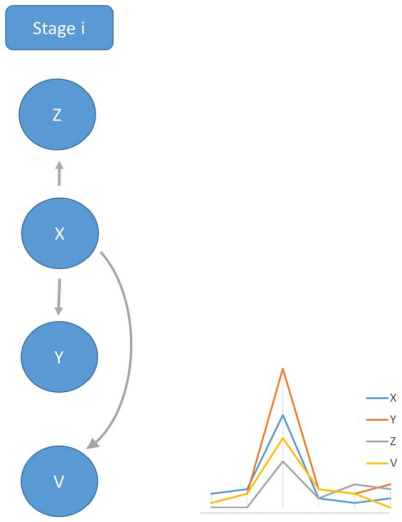

Figure 67. X a potential master regulator of Y, Z and V ,coupled with a high positive correlation to each of its targets

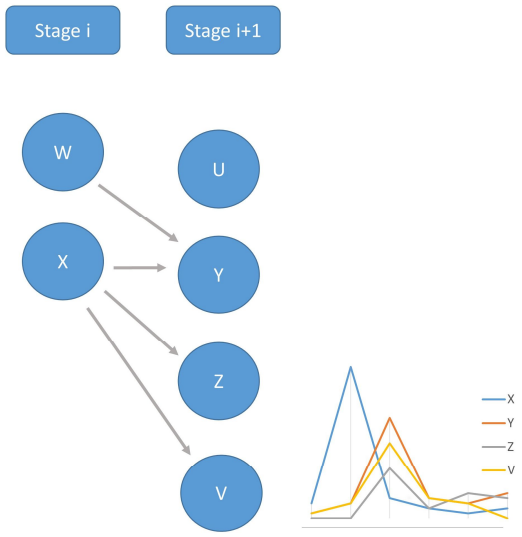

Figure 68. X a potential master regulator activating Y, Z, and V coupled with a high positive time-lagged correlation to each of its targets

6.3 Parameter adjustment

Adjusting the parameters can be a big factor in generating a better TRC. This adjustment should be based on a good understanding of the data and the questions in scope. Deciding on these parameters could be based on the quality of the dataset, inspecting the average expression levels, minimum and maximum of the genes in the dataset, and the number of time points and replicates in the dataset.

A very low minE would lead to noise in the cascade caused by very lowly expressed regulators, and a high minE could lead to missing on important regulators. While it has always been a question of debate, that has no definitive answer to which expression level threshold constitutes an expressed gene. One way we used in our work is looking at the mean and median of the expression values in a dataset to pick an expression threshold, typically close to the mean. However, another way would be looking at specific markers that are known to be expressed at a certain time point and basing the minE threshold on their average expression value or using the accompanying proteomic data, if available, to deduce what expression levels correspond to a gene being translated.

A low minC could lead to certain stages appearing to be significant when in reality they don't have any significantly peaking regulators, and a high minC could miss out on regulators that are peaking but not perfectly. This parameter has the least impact on the resulting TRC as the workflow will always rank the stage-specific regulators based on their correlation to the TPP, so even if the threshold was low, and many genes with low peak strength made it to the stage-specific list, they will be filtered out and the top minS ones are chosen. However, when the minS is large, this threshold becomes more significant. In datasets with a large number of replicates per time points, minC can be loosened up as the perfect correlation will be harder to achieve as all the replicates have to satisfy the required peak and sometimes the small variability between the replicates can produce some noise.

A low minS will also miss some important regulators with a not-perfectly-peaking pattern and could lead to a loss of significant GO terms due to the small data set size per stage, and a high maxS could overwhelm the cascade with insignificant regulators and make the cascade harder to inspect visually. A recommended way to go about this parameter is to increase it with datasets of a low number of time points to capture more regulators and increase it with those that have a larger number of time points to limit the number of regulators in the TRC. However, any minS greater than 20 is not recommended as the TRC becomes harder to inspect visually.

6.4 TRC comparative analysis

TRCs derived from distinct datasets can be compared to derive insights on the similarities and differences in the stage-specific regulators' sets. For this purpose, I developed a workflow that facilitates the visual comparison of two TRCs.

The comparative workflow takes two TRCs and compares the node of the second to the nodes of the first, producing a hybrid TRC that highlights the shared genes and their stage color code. The colors of the nodes in the first TRC are neutralized by changing their color into one unified color, typically not among the node colors in the second TRC to be compared with. The node colors in the second TRC are left as they are. The neutralized first TRC is used as the basis for the result, then the common nodes between the two TRCs are colored based on their color in the second TRC ().

Figure 70 shows the TRC resulting from the overlap of a TRC based on the H9 early cardiac differentiation dataset with the primary cascade based on the heart development dataset described in the results. Upon examining the resulting overlap, it is apparent that the colors are not scattered randomly across the TRC, and the temporal order of the colors is rather preserved. For example, the overlapping genes that are specific to Day 3 in the primary cascade are captured either on Day 2 or Day 4 in the H9 cascade. This indicates the robustness of the methodology against different time point selections in different experiments, enabling it to still capture the same important stage-specific regulators despite these differences.

Another observation we can derive from this comparison is that the overlapping genes from Day 3 in the primary cascade show four genes that appear earlier than the other two. This gives us a hint to the regulatory precedence of genes like CDX1 in activating the other Day 3 genes in the primary cascade.

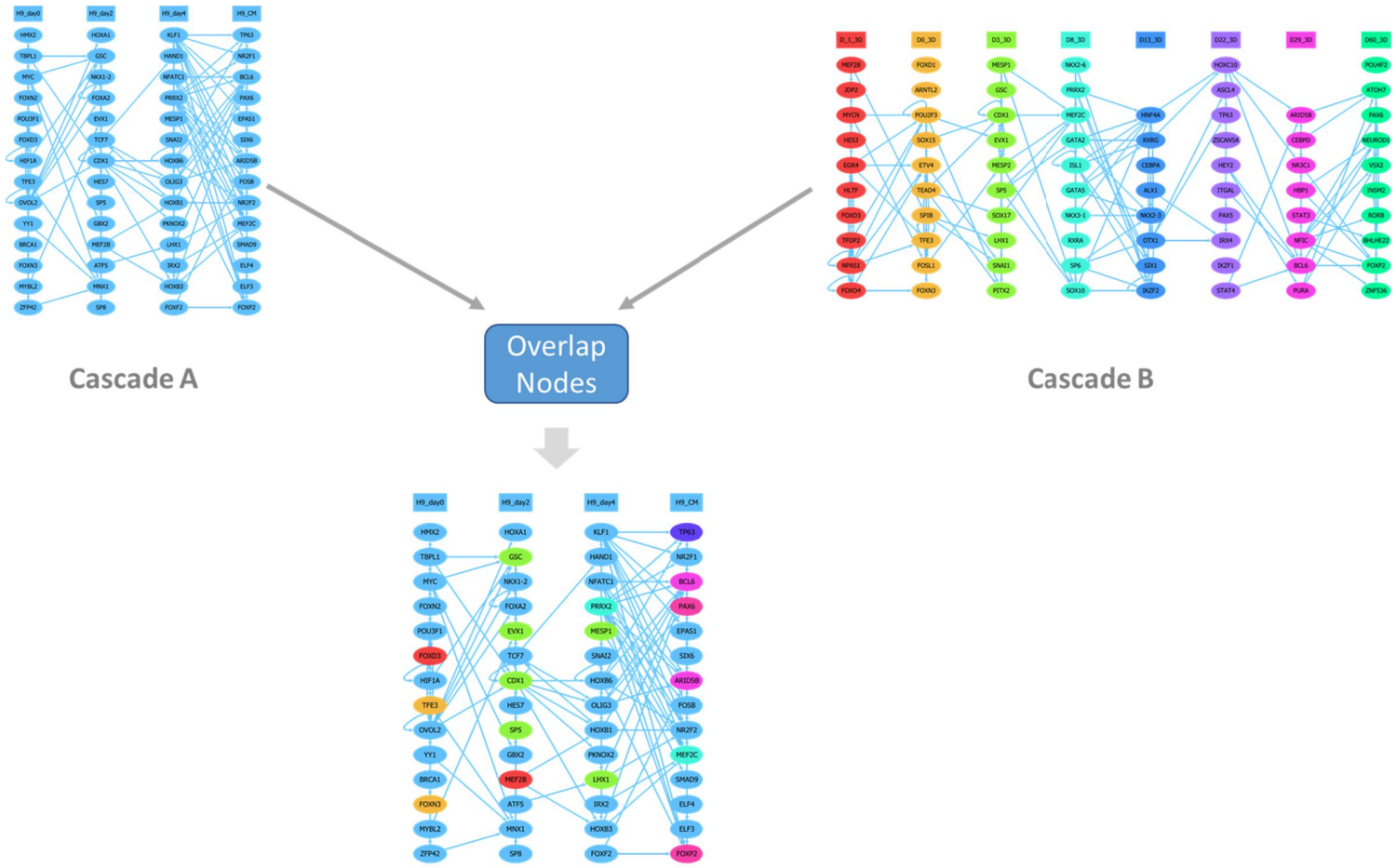

Figure 69. The overlapping proccess of two TRCs for comparative purposes. The colors in the first are neutralized and the overlapping nodes are colored according to their color in the second TRC.

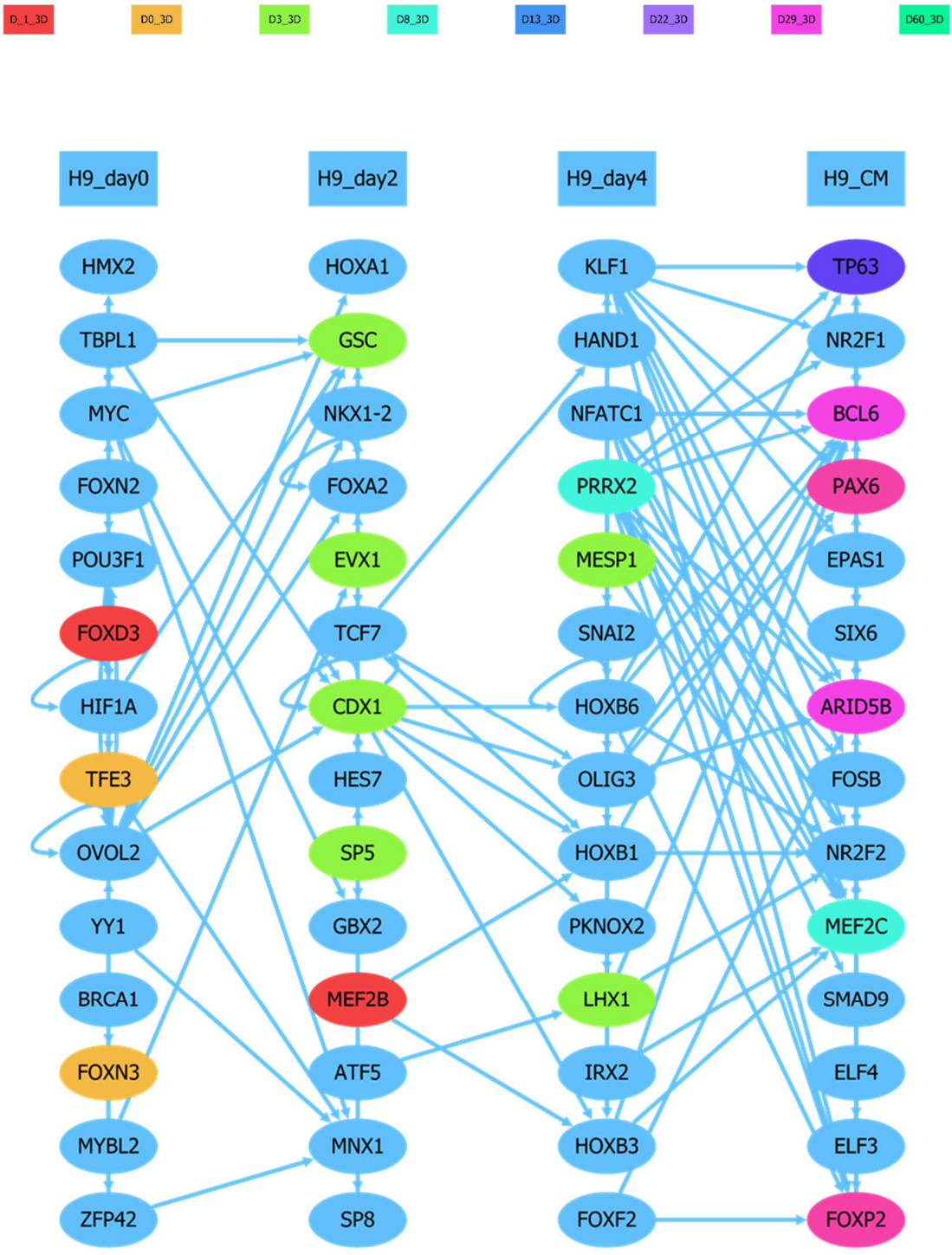

Figure 70. The resulting TRC from comparing the H9 early cardiac differentiation dataset with the heart development dataset.The stage color codes in the primary cascade are displayed on top. A conservation of temporal order and consistency between the two TRCs is observed.

6.5 Enrichment vs. correlation

Figure 71. Different relative TF-Target expression patterns of some of the TFs and targets found based on the enrichment in the regulatory network. The expression patterns are indicators that some of these predicted regulatory interactions are false positives.

The web service provides two methods to filter the queried potential regulators or targets of a set of genes, enrichment, and correlation.

The enrichment analysis uses a hypergeometric test to check for statistically significant regulators or targets, in the background network, that regulate or are regulated by a large percentage of the genes in the input list, but it ignores their expression value.

The correlation-based filtering, on the other hand, looks for those regulators and targets in the network that are also correlated to the input genes in the query, thus taking into consideration the relative expression values.

A statistically significant predicted TF could be completely unexpressed in the case of HOXD10 in Figure 71. , thus it cannot have activated its predicted target GSG. Or the predicted target can be expressed in far later stages that the regulation prediction would have been meaningless, like in the case of HOXC10 and its target MESP2. Or the expression patterns of the TF could be completely fluctuating with respect to the target that a direct regulatory interaction or strong influence might be highly unlikely despite the binding site possibility of the prediction.

While not all regulators and targets are correlated, those that are correlated can be indicators of an active regulatory interaction, particularly that it is coupled by binding site prediction. Thus, the correlation-based filtering can lead to potential false negatives but potentially to a larger percentage of true positives and lower false positives. On the other hand, the enrichment-based filtering can lead to many false-positive context-based regulatory interactions as it is only based on the regulatory network. As our focus is on providing smaller, context-dependent networks with lower false-positive regulatory interactions, we recommend by default the correlation-based filtering of the regulatory networks. However, the choice of the filtering can be dependent on the different questions and hypothesis the researcher is trying to investigate.

6.6 Other template libraries

Different biological processes use different regulatory programs and modes of regulation. For cell differentiation, the single peak pattern works compatibly, but for biological contexts such as diseases, immune response, stress response, and others, different template patterns might lead to capturing the significant regulators particular to that context. In this section, we devise some libraries and template patterns that can be used for different contexts.

One template pattern library that can be used for disease contexts and stress responses is represented in Figure 72. The library is composed of template patterns that capture genes that start being expressed after a certain time of exposure to stress or for example, a viral infection. This would help expose the temporal regulatory waves that follow up as a stress response.

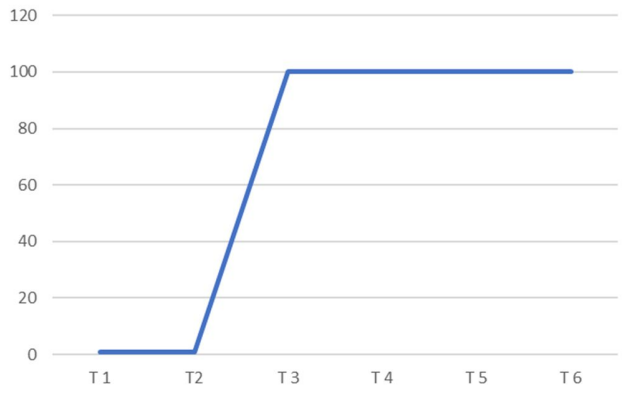

Figure 72. A TPP where the expression goes up and stays up after a the associated time point.

Another template pattern that can be used to construct a library is a multiple-stage span pattern. This is particularly useful when looking for more general regulators that govern the regulatory programs across multiple stages and time points.

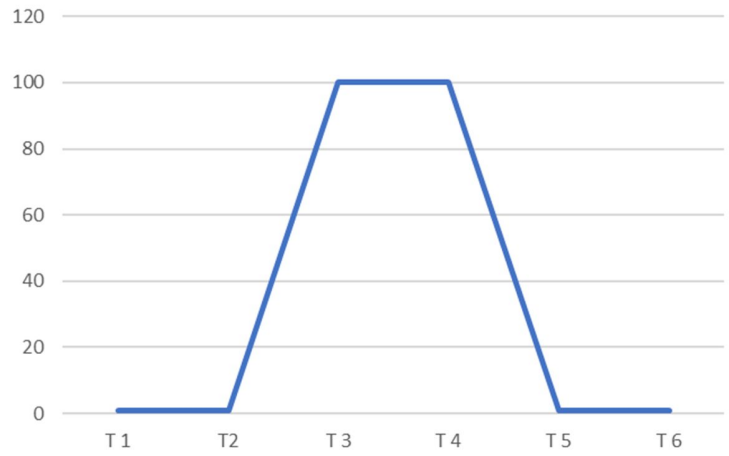

Figure 73. A multi- time point pattern for detecting more general TFs.

The anti-peak pattern is a pattern that is exactly opposite to the default TPP (Figure 74). The anti-peak pattern associated to time point t is an expression pattern constructed such that the expression is:

- 100 at time point t.
- 0 at every other time point.

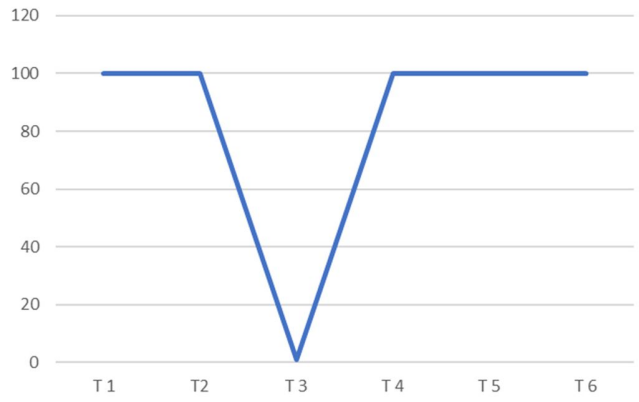

Figure 74. An anti peak template pattern associated with T3.

This pattern can provide the basis to generate a TRC that can help capture the stage-specific repressors. While the regulatory interactions, in this case, will be meaningless as the repressors within the same stage are not expressed to regulate each other or to regulate the upcoming stage, the list of repressors and their association with each stage can be very informative. The repressor lists can be used via the connected workflows to query the

correlated target genes and uncover insights about the repressive regulatory forces and their temporal role.

Other libraries such as shut down genes after a certain stage lead to capturing the patterns where certain genes are shut off after a certain time point of being exposed to a stress or a disease, leading to a TRC that has a different interpretation than the single peak one but can hold more useful information for the context.

6.7 TRCs and proteomics

The mRNA serves as a blueprint for protein synthesis, however not all mRNA fragments get translated to proteins, and a single mRNA fragment can be used more than once by different ribosomes to produce multiple polypeptides. All these variables and factors make the relationship between the mRNA and protein levels a non-direct one.

Although it is mostly applied to RNA-Seq, the model itself is not exclusive to data based on mRNA levels. Theoretically, the same workflow can be applied to good quality proteomic time-series data when available, to detect spikes in the protein levels, corresponding to certain genes and detect their stage specificity. However, these spikes in the protein levels are not as sharp or frequently occurring as those observed for the mRNA due to their low turnover rate.

Some studies suggested that though not linear, the relationship between the protein levels and the mRNA exists at a certain level. Genes with high mRNA levels will have a higher protein to mRNA ratio than lowly expressed mRNAs [271][272]. Other studies have even attempted to calculate these relationships for each gene individually [273]. These relationships can be used to estimate the protein levels of the genes in the cascade or for filtering out those that have no definite relationship between their mRNA and protein levels from the stage-specific regulator's sets.

Moreover, the cascade can provide small candidate sets for the proteomic investigation. The protein levels of the stage-specific regulators of the relevant time points or of the whole cascade can be measured, as a second step after the RNA-Seq measurement, using the same samples, thus reducing the costs of the proteomic analysis by testing a relatively small candidate set.

6.8 Shuffling and randomization

The TRC analyses on various datasets, including the ones analyzed in the results section, showed clear stage-specific regulatory waves and a GO enrichment that is highly consistent with the biological context of the experiment and, even more specifically, the context in the particular time points of the experiment.

The question arises whether these peaking profiles and case-specific GO enrichments are statistically significant, and constitute a characteristic of developmental gene expression datasets in particular, or if they randomly occur in any dataset.

While applying the TRC model thousands of times to a large number of random and shuffled datasets and evaluating the resulting TRCs would be optimal to proof the statistical significance of the results, it is merely unfeasible due to the manual process of assessing the resulting TRCs. Alternatively, we applied the model to randomly generated and shuffled gene expression datasets aiming towards a comparative analysis rather than a statistical proof of significance. We examined the resulting TRCs in terms of the GO enrichment of the stages to evaluate their relevance compared to a TRC generated from a real experimental dataset.

The first test involved shuffling the heart development dataset by re-assigning genes to other expression profiles to check whether any set of peaking regulators will show a specific GO enrichment, and none of the stages did lead to any relevant terms. The test was repeated by shuffling the regulator's profiles only, and the enrichment was again insignificant. The previous test showed that the identity of the peaking genes is essential, precise, and specific.

In the second test, the workflow was applied to the heart development dataset without restricting the stage-specific sets to only regulator genes (Figure 75). The generated cascade was overwhelmed by non-regulatory genes, and the GO enrichment showed the enrichment of very few cardiac-related significant terms and only in one of the stages. The cardiac-specificity of the GO terms in this TRC is much lower than the original regulators-only TRC. This could be due to the fact that non-regulator genes are less annotated in the GO than regulatory genes, or simply that the effect of a handful of non-regulatory genes is not as impactful as a handful of regulators in determining the fate of cell differentiation. This observation supports the choice in the TRC model of limiting the cascade to regulators where less relevant non-regulatory genes do not dilute the small stage-specific gene sets.

In the third test, the heart development dataset was shuffled by permuting all the values in the expression matrix. The result was again a lack of significance in GO the enrichment terms.

The last test was applying the TRC workflow to a randomly generated gene expression dataset, using the gene names and the time points from the heart development dataset combined with randomly generated expression values. The GO enrichment showed an absence of any relevant significant terms again.

Figure 75. TRC generated without the restriction to regulators. The absence of the regulatory connections is due to the peaking non-regulatory genes overwhelming the stage-specific gene sets.

6.9 TF families in TRCs

Through the analysis of more than 30 different datasets using the TRC model, an interesting pattern of behavior of TF families became apparent. We observed the relative patterns and associated stages of TFs belonging to the same family in each TRC and categorized it into two main modes of behavior.

The first mode is represented in a heavy same-stage appearance of multiple TFs belonging to the same family. An example of this would be CDX1, CDX2, HOXA1, and HOXB1 which belong to the HOX family and appear to peak in the same stage in the early cardiac differentiation TRC. We hypothesize that these TFs have rather redundant roles, competing for the same binding sites and adding robustness for the regulatory programs at that stage. However, a further investigation of the similarities of the protein structure of these TFs can confirm or lead to different hypothesis.

The second mode is represented in the rather successive appearance of members of the same family across stages. We hypothesize that these TFs have distinct roles where each of them, at a certain stage, occupy a set of binding sites that is reused by another TF of the same family in the next stage to contribute to a different regulatory program and biological processes.

6.10 Application to non-temporal datasets

Although the TRC model is designed to be applied for temporal gene expression datasets, it can be adapted for any gene expression dataset with multiple conditions. The TRC model can detect condition-specific regulators and regulatory networks the same way it detects stage-specific ones (Figure 76). Regulators are identified based on their peak in a specific condition and their low expression in all the others. The inter-regulatory edges in this case are meaningless, however the intra-regulatory edges and condition specific-regulators can be very helpful.

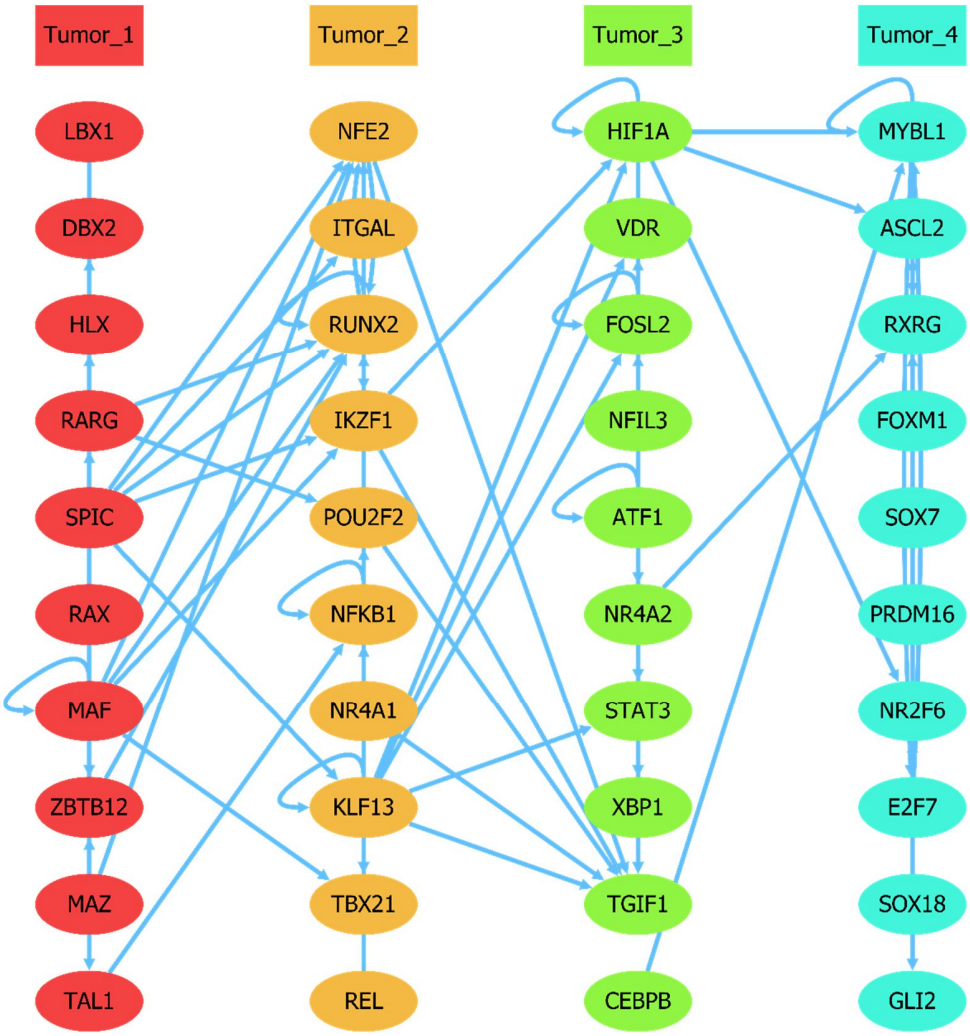

Figure 76. A TRC applied to a simulated multiple-conditions dataset. The regulators that are peaking in one tumor type and not the others are detected as well as the regulatory interactions between them. The cross edges between the columns are meaningless in this case as they are only meaningful in temporal datasets.

6.11 Comparison with other tools

Although the model is different in its methodology and output than many of the current popular methods, we attempted to compare the results of applying 3 popular methods STEM, iDREM and DEG analysis to the TRC generated by our model. We used the same dataset, the heart development dataset for this comparison.

6.11.1 STEM

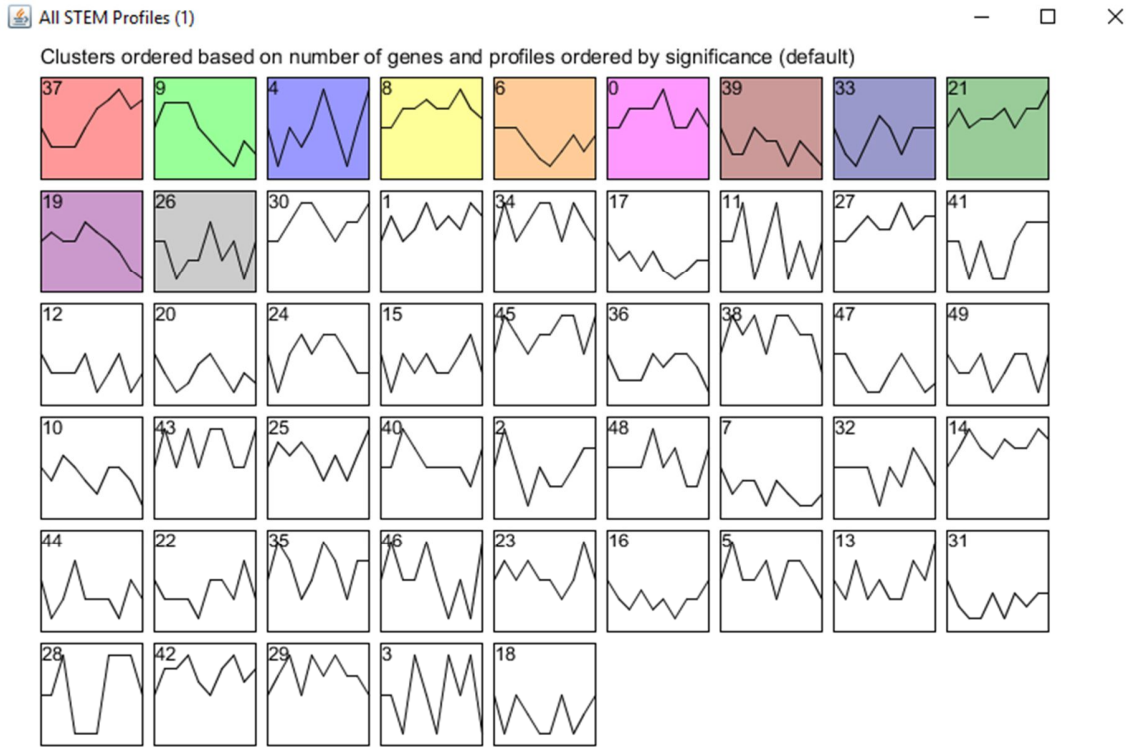

Figure 77. The top significant gene expression patterns predicted by STEM in the heart development dataset.

We applied the STEM tool to the heart development dataset using the default parameters in the STEM interface. The result is a list of statistically significant gene expression profiles and the associated gene list with each profile (Figure 77).

Upon examining the top 10 profiles, we cannot see any stage-specific pattern as most of the profiles have a rather fluctuating pattern making a conclusion about the role of the genes associated with these profiles unclear.

The GO terms enriched for each of the associated gene lists in the significant profiles were examined as well. An absence of cardiac-related terms or even cell differentiation contexts was common through the GO enrichment analysis of all of the top 10 sets.

6.11.2 iDREM

In the first run of iDREM, the human_predicted_1000 reference network provided by the tool was used as a background regulatory network, and the HMM model was generated and displayed (Figure 78). The model contained over 100 predicted bifurcation points. Each bifurcation point had a list of associated TFs. We examined the GO enrichment and the identity of the regulators at the critical bifurcation points. While main stage-specific cardiac regulators such as MEF2C, ISL1, and other important cardiac TFs were not identified in any of the bifurcation point TF sets generated by iDREM, it managed to identify some cardiac TFs that were not identified by the TRC model such as GATA4 and NKX2-5 , as important TFs in a bifurcation point in Day 8. iDREM seems to have an advantage on the TRC model in capturing general TFs whose expression span over multiple time points, but lacks in capturing stage-specific TFs in comparison to the TRC. When it comes to performance, iDREM took several hours per run compared to few seconds that TRC workflow required to generate its cascade.

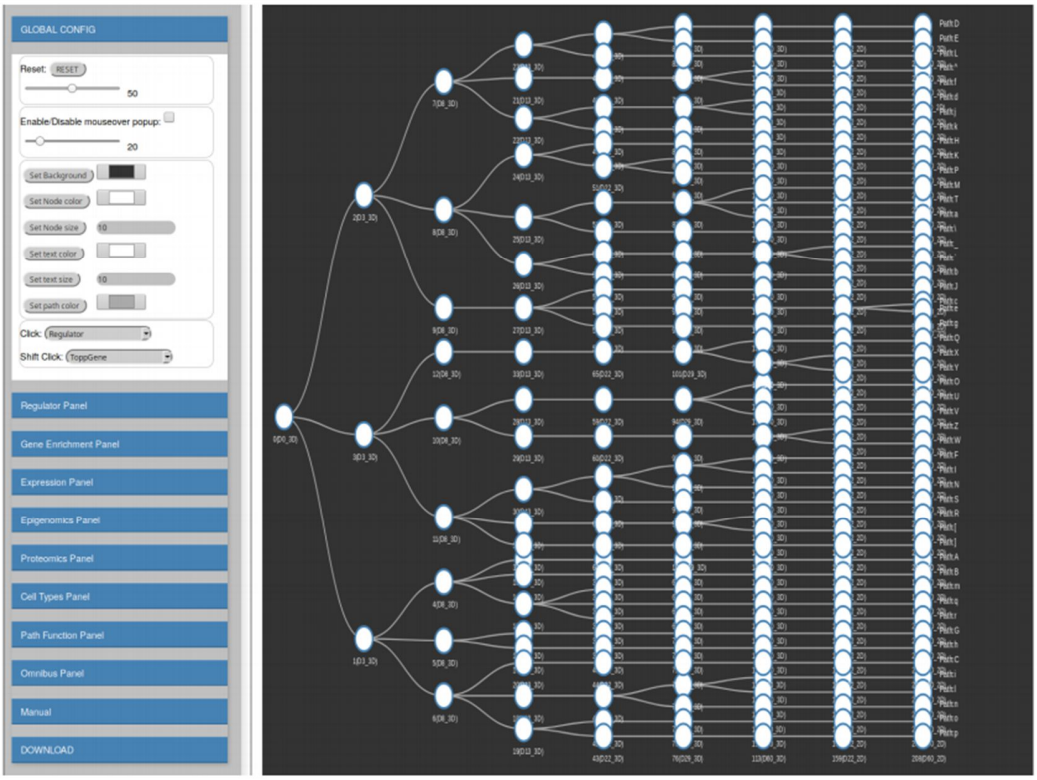

Figure 78. The HMM output of the analysis of the heart development dataset using iDREM. Each node in the model represents a bifurcation point, and not a gene as in the TRC. Each bifurcation node contains a list of genes.

TF	Num Total	Num Parent	Num Path	Expect Overall	Diff. Overall	Score Overall	Expect Split	Diff. Split	Score Split	% Split
RELA	1448	238	85	45.80	39.20	2.06e-8	67.56	17.44	5.60e-3	35.71
PRDM1	768	117	46	24.29	21.71	2.71e-5	33.21	12.79	5.76e-3	39.32
RREB1	1191	174	63	37.67	25.33	4.30e-5	49.39	13.61	0.012	36.21
ARNT	2869	489	159	90.75	68.25	2.39e-13	138.80	20.20	0.013	32.52
ALX4	772	127	47	24.42	22.58	1.47e-5	36.05	10.95	0.019	37.01
TEAD1	1378	233	80	43.59	36.41	9.05e-8	66.14	13.86	0.021	34.33
GATA4	1312	236	81	41.50	39.50	5.02e-9	66.99	14.01	0.021	34.32
REST	1694	195	68	53.58	14.42	0.024	55.35	12.65	0.023	34.87
GATA3	1662	295	98	52.57	45.43	1.11e-9	83.73	14.27	0.029	33.22
ZEB1	1953	306	101	61.77	39.23	4.55e-7	86.86	14.14	0.032	33.01
SPI1	1136	182	63	35.93	27.07	1.03e-5	51.66	11.34	0.033	34.62
MAFG	762	128	46	24.10	21.90	2.24e-5	36.33	9.67	0.034	35.94
MAFB	762	128	46	24.10	21.90	2.24e-5	36.33	9.67	0.034	35.94
MAFF	762	128	46	24.10	21.90	2.24e-5	36.33	9.67	0.034	35.94
MAFK	762	128	46	24.10	21.90	2.24e-5	36.33	9.67	0.034	35.94
NFE2L3	762	128	46	24.10	21.90	2.24e-5	36.33	9.67	0.034	35.94
MAF	762	128	46	24.10	21.90	2.24e-5	36.33	9.67	0.034	35.94
ETV4	975	144	51	30.84	20.16	3.06e-4	40.87	10.13	0.035	35.42
HNF4G	1330	207	70	42.07	27.93	1.86e-5	58.76	11.24	0.042	33.82
JUN	2281	444	141	72.15	68.85	1.76e-15	126.03	14.97	0.045	31.76

Figure 79. A snapshot of the gene list associated with one of the bifurcation points in the iDREM generated HMM.

The GO enrichment of the gene lists was examined, and was overall found to be low on significant cardiac-related genes particularly when compared to the GO enrichment of the stage-specific sets generated by the TRC model (Figure 79).

6.11.3 DEG Analysis

Table 20 shows the enrichment of the DEG lists generated by a previously published study on the heart development dataset [274]. Examining the size of the generated DEG lists shows significantly bigger lists per time point compared to the lists generated by the TRC model. These big lists hold a drawback when it comes to choosing candidate genes to test since a manual, or further computation reduction has to be made before a practical set of candidates can be identified for experimental inspection. On the other hand, examining the GO terms resulting from the enrichment of such lists, no direct cardiac-relevant terms are observed, which might be due to the dilution of the relevant genes in a larger number of non-relevant ones.

Table 20. The top terms of the GO enrichment of the DEG lists generated in a previously published study on the heart development dataset.

Stage	**Number of DEGs**	**GO terms**
Mesoderm Induction	429	• vesicle transport along actin filament • positive regulation of protein localization to plasma membrane • actin filament bundle assembly • regulation of protein tyrosine kinase activity • positive regulation of protein localization to nucleus
Early Cardiac Specification	1241	• positive regulation of rRNA processing • positive regulation of ribosome biogenesis • positive regulation of transcription of nucleolar large rRNA by RNA polymerase I • DNA replication-dependent nucleosome assembly • regulation of rRNA processing
Late Cardiac Specification	36	• L-cystine transport • amino acid transmembrane transport
Early Cardiac Maturation	204	• cellular metabolic process
Late Cardiac Maturation	975	• negative regulation of vesicle fusion • postsynaptic density protein 95 clustering • positive regulation of protein localization to synapse • postsynaptic specialization assembly

7 Conclusion

7.1 Summary

In this thesis, I developed the concept of temporal regulatory cascades (TRCs) in the form of a model that was implemented as an interactive web-service. I then applied the method in order to analyze different temporal gene expression datasets and examined the results from a biological point of view to derive new knowledge and compare it to the existing literature and experimental knowledge.

The TRC method utilizes a collection of template peak patterns (TPPs) and a background general regulatory network. Each TPP that has a peak at one of the time points is then used to attract genes, typically regulators, which are specific for a certain stage. These sets of stage-specific regulators are organized based on their temporal order in a cascade formation. The background regulatory network is based on TF binding site predictions and provides the source for querying the regulatory interactions between stage-specific regulators, which are represented in the form of edges in the TRC.

The TRC method was implemented as the main workflow in a web-service that contains other related and interconnected workflows. The user can visually explore questions related to co-expression and co-regulation in his dataset, refining gene sets of interest and investigating the effects of different regulatory sets to end up with biologically sensible results.

The TRC workflow was applied to different temporal datasets that revolved around cell differentiation experiments. The stage-specific regulators' sets were investigated and evaluated using GO enrichment and the literature knowledge and were found to be highly consistent with the underlying biological stages and events. Previously-known important regulators and temporal regulatory interactions were identified as well as new, potentially significant ones.

The method suffers from sensitivity to the experimental design and choice of time points. However, several solutions were devised in this manuscript to work around this problem. On the other hand, the TRC method computationally out-performs other popular methods and delivers concise information-dense results that can be used and interpreted by biologists.

7.2 Outlooks

While in this manuscript I presented the core TRC model, this model can be expanded in different directions, by adding features and methods to cover more biological scenarios and hypotheses. On the other hand, the webtool is flexible enough to incorporate various features, methods, and relevant links. Hereby I present a set of features that, had the time allowed, would have been implemented to enrich the current method even more. These outlooks can provide the basis for the next version of the method in case it is upgraded and extended by me or any fellow scientist.

One main feature that can be added is a collection of context-dependent libraries that cover different biological contexts such as disease, stress response, drug reaction. These libraries can be used accordingly for different datasets that are relevant to these contexts. This enables a wider set of users to use the model effectively.

Another feature that can be extended in the model is giving the option for a user to construct his own template library and building the cascade. This feature would enable the user to visually construct the template patterns of interest based on his knowledge about the time points and the questions in mind and constructing a TRC around it. The interface, in that case, would include a graph featuring artificial expression patterns of a regulator and its target, where the user can drag the patterns constructing an ideal relative position.

One powerful addition to the method would be linking the evaluation of the stage-specific gene sets to a semantic ontology. The ontology would then evaluate the GO terms and detect those that are biologically relevant to the case. This evaluation would be based on a set of keywords provided by the user about the experimental context and keywords for each particular time point. Such a feature will enable optimizing the model and its parameters to generate the most biologically relevant cascade, where the density of case-relevant stage-specific regulators is maximized. Moreover, it will allow the automatic evaluation of the cascade, which translates into an ability to benchmark the model, calculate its significance statistically, and compare it to other methods.

Chromatin accessibility information such as ATAC-seq, ChIP-seq, and DNase-seq, when available in the experiment, would provide a very good addition to the current model. These data can be integrated in various ways to enrich the model or validate the predictions. Regulatory predictions, for example, can be filtered based on the availability of the promoter of the target gene at the relevant time-point. ChIP-seq data for certain TFs can provide

additional evidence for a regulatory prediction that might be occurring at a particular time-point.

As different technologies of measuring gene expression evolve, methods and models to analyze the output evolve in parallel. We hope that the methods described in this thesis can provide one useful way to approach the puzzle of gene expression and regulation, and merely lay some pieces in the right places within the enormous puzzle of understanding biological systems.

[1] K. Takahashi and S. Yamanaka, “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors,” *Cell*, vol. 126, no. 4, pp. 663–676, 2006.

[2] E. Tzahor and K. D. Poss, “Cardiac regeneration strategies: staying young at heart,” *Science (80-.).*, vol. 356, no. 6342, pp. 1035–1039, 2017.

[3] Y.-K. Kwon and K.-H. Cho, “Analysis of feedback loops and robustness in network evolution based on Boolean models,” *BMC Bioinformatics*, vol. 8, no. 1, p. 430, 2007.

[4] S. Liang, S. Fuhrman, and R. Somogyi, “Reveal, a general reverse engineering algorithm for inference of genetic network architectures,” 1998.

[5] B. Ristevski, “A survey of models for inference of gene regulatory networks,” *Nonlinear Anal Model Control*, vol. 18, no. 4, pp. 444–465, 2013.

[6] D. M. Chickering, D. Heckerman, and C. Meek, “A Bayesian approach to learning Bayesian networks with local structure,” in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, 1997, pp. 80–89.

[7] R. E. Neapolitan and others, *Learning bayesian networks*, vol. 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[8] M. Zou and S. D. Conzen, “A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data,” *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 2004.

[9] L. F. A. Wessels, E. P. V. A. N. SOMEREN, and M. J. T. Reinders, “A comparison of genetic network models,” in *Biocomputing 2001*, World Scientific, 2000, pp. 508–519.

[10] A. A. Margolin *et al.*, “ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” in *BMC bioinformatics*, 2006, vol. 7, no. 1, p. S7.

[11] J. Schäfer and K. Strimmer, “Learning Large-Scale Graphical Gaussian Models from Genomic Data,” in *AIP Conference Proceedings*, 2005, vol. 776, no. 1, pp. 263–276.

[12] M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph, “DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data,” *BMC Syst. Biol.*, 2012.

[13] J. Ding, J. S. Hagood, N. Ambalavanan, N. Kaminski, and Z. Bar-Joseph, “iDREM: Interactive visualization of dynamic regulatory networks,” *PLoS Comput. Biol.*, 2018.

[14] M. Sehgal, I. Gondal, and L. Dooley, “CF-GeNe: Fuzzy Framework for Robust Gene Regulatory Network Inference,” *JCP*, vol. 1, pp. 1–8, 2006.

[15] A. Brazma and T. Schlitt, “Reverse engineering of gene regulatory networks: a finite state linear model,” *Genome Biol.*, vol. 4, no. 6, p. P5, 2003.

[16] Z. Li, S. M. Shaw, M. J. Yedwabnick, and C. Chan, “Using a state-space model with hidden variables to infer transcription factor activities,” *Bioinformatics*, vol. 22, no. 6, pp. 747–754, 2006.

[17] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, “Comparing statistical methods for constructing large scale gene networks,” *PLoS One*, vol. 7, no. 1, p. e29348, 2012.

[18] Y. Chuan Tai and T. P. Speed, “On gene ranking using replicated microarray time course data,” *Biometrics*, vol. 65, no. 1, pp. 40–51, 2009.

[19] A. A. Kalaitzis and N. D. Lawrence, “A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression,” *BMC Bioinformatics*, vol. 12, no. 1, p. 180, 2011.

[20] C. Cheng, X. Ma, X. Yan, F. Sun, and L. M. Li, “MARD: a new method to detect differential gene expression in treatment-control time courses,” *Bioinformatics*, vol. 22, no. 21, pp. 2650–2657, 2006.

[21] X. L. Xu, J. M. Olson, and L. P. Zhao, “A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington’s disease transgenic model,” *Hum. Mol. Genet.*, vol. 11, no. 17, pp. 1977–1985, 2002.

[22] G. K. Smyth, “Limma: linear models for microarray data,” in *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer, 2005, pp. 397–420.

[23] D. Chudova, C. Hart, E. Mjolsness, and P. Smyth, “Gene expression clustering with functional mixture models,” in *Advances in Neural Information Processing Systems*, 2004, pp. 683–690.

[24] T. Park *et al.*, “Statistical tests for identifying differentially expressed genes in time-course microarray experiments,” *Bioinformatics*, vol. 19, no. 6, pp. 694–703, 2003.

[25] O. ElBakry, M. O. Ahmad, and M. N. S. Swamy, “Identification of differentially expressed genes for time-course microarray data based on modified RM ANOVA,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 9, no. 2, pp. 451–466, 2012.

[26] M. J. Nueda *et al.*, “Discovering gene expression patterns in time course microarray experiments by ANOVA--SCA,” *Bioinformatics*, vol. 23, no. 14, pp. 1792–1800, 2007.

[27] M. Yuan and C. Kendziorski, “Hidden Markov models for microarray time course data in multiple biological conditions,” *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1323–1332, 2006.

[28] C.-Y. Li, X. Mao, and L. Wei, “Genes and (common) pathways underlying drug addiction,” *PLoS Comput. Biol.*, vol. 4, no. 1, p. e2, 2008.

[29] Z. Bar-Joseph, G. Gerber, I. Simon, D. K. Gifford, and T. S. Jaakkola, “Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 18, pp. 10146–10151, 2003.

[30] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, “Significance analysis of time course microarray experiments,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 36, pp. 12837–12842, 2005.

[31] F. Hong and H. Li, “Functional hierarchical models for identifying genes with different time-course expression profiles,” *Biometrics*, vol. 62, no. 2, pp. 534–544, 2006.

[32] C. Angelini, L. Cutillo, D. De Canditiis, M. Mutarelli, and M. Pensky, “BATS: a Bayesian user-friendly software for analyzing time series microarray experiments,” *BMC Bioinformatics*, vol. 9, no. 1, p. 415, 2008.

[33] C. Angelini, D. De Canditiis, M. Mutarelli, and M. Pensky, “A Bayesian approach to estimation and testing in time-course microarray experiments,” *Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, 2007.

[34] P. Ma, W. Zhong, and J. S. Liu, “Identifying differentially expressed genes in time course microarray data,” *Stat. Biosci.*, vol. 1, no. 2, p. 144, 2009.

[35] X. Liu and M. C. K. Yang, “Identifying temporally differentially expressed genes through functional principal components analysis,” *Biostatistics*, vol. 10, no. 4, pp. 667–679, 2009.

[36] P. Langfelder and S. Horvath, “WGCNA: an R package for weighted correlation network analysis,” *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[37] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, “Systematic determination of

genetic network architecture," *Nat. Genet.*, vol. 22, no. 3, p. 281, 1999.

[38] P. Tamayo *et al.*, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci.*, vol. 96, no. 6, pp. 2907–2912, 1999.

[39] M. P. S. Brown *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Natl. Acad. Sci.*, vol. 97, no. 1, pp. 262–267, 2000.

[40] R. Sharan and R. Shamir, "CLICK: a clustering algorithm with applications to gene expression analysis," in *Proc Int Conf Intell Syst Mol Biol*, 2000, vol. 8, no. 307, p. 16.

[41] J. Kim and J. H. Kim, "Difference-based clustering of short time-course microarray data with replicates," *BMC Bioinformatics*, vol. 8, no. 1, p. 253, 2007.

[42] T. Hastie *et al.*, "'Gene shaving'as a method for identifying distinct sets of genes with similar expression patterns," *Genome Biol.*, vol. 1, no. 2, pp. research0003--1, 2000.

[43] A. B. Tchagang, K. V Bui, T. McGinnis, and P. V Benos, "Extracting biologically significant patterns from short time series gene expression data," *BMC Bioinformatics*, vol. 10, no. 1, p. 255, 2009.

[44] P. Magni, F. Ferrazzi, L. Sacchi, and R. Bellazzi, "TimeClust: a clustering tool for gene expression time series," *Bioinformatics*, vol. 24, no. 3, pp. 430–432, 2007.

[45] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *J. Comput. Biol.*, vol. 6, no. 3–4, pp. 281–297, 1999.

[46] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[47] R. Šášik, N. Iranfar, T. Hwa, and W. F. Loomis, "Extracting transcriptional events from temporal gene expression patterns during Dictyostelium development," *Bioinformatics*, vol. 18, no. 1, pp. 61–66, 2002.

[48] Z. Bar-Joseph, G. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "A new approach to analyzing gene expression time series data," in *Proceedings of the sixth annual international conference on Computational biology*, 2002, pp. 39–48.

[49] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proc. Natl. Acad. Sci.*, vol. 99, no. 14, pp. 9121–9126, 2002.

[50] L. Wang, M. Ramoni, and P. Sebastiani, "Clustering short gene expression profiles," in *Annual International Conference on Research in Computational Molecular Biology*, 2006, pp. 60–68.

[51] Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics*, vol. 19, no. 4, pp. 474–482, 2003.

[52] T. Scharl, B. Grün, and F. Leisch, "Modelling time course gene expression data with finite mixtures of linear additive models," *Bioinformatics*, vol. 26, no. 3, pp. 370–377, 2010.

[53] I. Costa, A. Schönhuth, A. S.- Bioinformatics, and undefined 2005, "The Graphical Query Language: a tool for analysis of gene expression time-courses," *academic.oup.com*.

[54] A. Schliep, I. G. Costa, C. Steinhoff, and A. Schonhuth, "Analyzing gene expression time-courses," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 2, no. 3, pp. 179–193, 2005.

[55] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talón, "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments," *Bioinformatics*, vol. 22, no. 9, pp. 1096–1102, 2006.

[56] N. A. Heard, C. C. Holmes, and D. A. Stephens, "A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of

curves," *J. Am. Stat. Assoc.*, vol. 101, no. 473, pp. 18–29, 2006.

[57] J. W. Chou, T. Zhou, W. K. Kaufmann, R. S. Paules, and P. R. Bushel, "Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes," *BMC Bioinformatics*, 2007.

[58] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, and D. M. Umbach, "Gene selection and clustering for time-course and dose--response microarray experiments using order-restricted inference," *Bioinformatics*, vol. 19, no. 7, pp. 834–841, 2003.

[59] T. Liu, N. Lin, N. Shi, and B. Zhang, "Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments," *BMC Bioinformatics*, vol. 10, no. 1, p. 146, 2009.

[60] D. Sahoo, D. L. Dill, R. Tibshirani, and S. K. Plevritis, "Extracting binary signals from microarray time-course data," *Nucleic Acids Res.*, vol. 35, no. 11, pp. 3705–3712, 2007.

[61] N. Ramakrishnan, S. Tadepalli, L. T. Watson, R. F. Helm, M. Antoniotti, and B. Mishra, "Reverse engineering dynamic temporal models of biological processes and their relationships," *Proc. Natl. Acad. Sci.*, vol. 107, no. 28, pp. 12511–12516, 2010.

[62] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, no. 1, p. 191, 2006.

[63] T. Springer, K. Ickstadt, and J. Stöckler, "Frame potential minimization for clustering short time series," *Adv. Data Anal. Classif.*, vol. 5, no. 4, pp. 341–355, 2011.

[64] T. R. Hvidsten, A. Lægreid, and J. Komorowski, "Learning rule-based models of biological process from gene expression time profiles using gene ontology," *Bioinformatics*, vol. 19, no. 9, pp. 1116–1123, 2003.

[65] L. Wang, X. Chen, R. D. Wolfinger, J. L. Franklin, R. J. Coffey, and B. Zhang, "A unified mixed effects model for gene set analysis of time course microarray experiments," *Stat. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–18, 2009.

[66] M. Hummel, R. Meister, and U. Mansmann, "GlobalANCOVA: exploration and assessment of gene group effects," *Bioinformatics*, vol. 24, no. 1, pp. 78–85, 2007.

[67] K. Zhang, H. Wang, A. C. Bathke, S. W. Harrar, H.-P. Piepho, and Y. Deng, "Gene set analysis for longitudinal gene expression data," *BMC Bioinformatics*, vol. 12, no. 1, p. 273, 2011.

[68] F. Yao, H.-G. Müller, and J.-L. Wang, "Functional data analysis for sparse longitudinal data," *J. Am. Stat. Assoc.*, vol. 100, no. 470, pp. 577–590, 2005.

[69] M. J. Nueda *et al.*, "Functional assessment of time course microarray data," in *BMC bioinformatics*, 2009, vol. 10, no. 6, p. S9.

[70] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: software for visualization and analysis of biological networks.," *Methods Mol. Biol.*, 2011.

[71] R. Saito *et al.*, "A travel guide to Cytoscape plugins," *Nature Methods*. 2012.

[72] I. H. Goenawan, K. Bryan, and D. J. Lynn, "DyNet: Visualization and analysis of dynamic molecular interaction networks," in *Bioinformatics*, 2016.

[73] M. Modrák and J. Vohradský, "Genexpi: A toolset for identifying regulons and validating gene regulatory networks using time-course expression data," *BMC Bioinformatics*, 2018.

[74] E. Wingender, "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation," *Brief. Bioinform.*, vol. 9, no. 4, pp. 326–332, 2008.

[75] "MATCH: A tool for searching transcription factor binding sites in DNA sequences. , author = Kel, A E and Gössling, E and Reuter, I and Cheremushkin, E and Kel-Margoulis, O V and Wingender, E , journal = Nucleic acids research , journal-iso = Nucleic Acids."

[76] C. Meckbach, R. Tacke, X. Hua, S. Waack, E. Wingender, and M. Gültas, "PC-TraFF: Identification of potentially collaborating transcription factors using pointwise mutual information," *BMC Bioinformatics*, 2015.

[77] M. Haubrock, J. Li, and E. Wingender, "Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks," in *BMC systems biology*, 2012, vol. 6, no. 2, p. S15.

[78] X. Hua *et al.*, "MirTrans: A resource of transcriptional regulation on microRNAs for human cell lines," *Nucleic Acids Res.*, 2018.

[79] J. Göke *et al.*, "Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development," *PLoS Comput. Biol.*, vol. 7, no. 12, p. e1002304, 2011.

[80] S. M. Kiełbasa and M. Vingron, "Transcriptional autoregulatory loops are highly conserved in vertebrate evolution," *PLoS One*, vol. 3, no. 9, p. e3210, 2008.

[81] A. Woolfe *et al.*, "Highly conserved non-coding sequences are associated with vertebrate development," *PLoS Biol.*, 2005.

[82] R. M. Cripps and E. N. Olson, "Control of cardiac development by an evolutionarily conserved transcriptional network," *Dev. Biol.*, vol. 246, no. 1, pp. 14–28, 2002.

[83] J. J. Miller, "Graph database applications and concepts with Neo4j," in *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, 2013.

[84] T. Gene and O. Consortium, "Gene Ontology : tool for the," *Gene Expr.*, 2000.

[85] P. Gaudet, "The Gene Ontology," in *Encyclopedia of Bioinformatics and Computational Biology*, 2019.

[86] P. Shannon *et al.*, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Res.*, 2003.

[87] C. T. Lopes *et al.*, "Cytoscape Web: An interactive web-based network browser," in *Bioinformatics*, 2011.

[88] Q. Liu *et al.*, "Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs," *Circ. Res.*, vol. 121, no. 4, pp. 376–391, 2017.

[89] M. Golumbeanu *et al.*, "Proteo-Transcriptomic Dynamics of Cellular Response to HIV-1 Infection," *Sci. Rep.*, 2019.

[90] M. Haubrock, J. Li, and E. Wingender, "Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks," *BMC Syst. Biol.*, vol. 6, no. SUPPL.2, Dec. 2012.

[91] C. A. Sloan *et al.*, "ENCODE data at the ENCODE portal," *Nucleic Acids Res.*, 2016.

[92] S. G. Landt *et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia," *Genome Research*. 2012.

[93] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinformatics*, 2008.

[94] S. Kitajima, A. Takagi, T. Inoue, and Y. Saga, "MesP1 and MesP2 are essential for the development of

cardiac mesoderm," *Development*, 2000.

[95] A. Bondue *et al.*, "Defining the earliest step of cardiovascular progenitor specification during embryonic stem cell differentiation," *J. Cell Biol.*, 2011.

[96] R. David *et al.*, "Forward programming of pluripotent stem cells towards distinct cardiovascular cell types," *Cardiovasc. Res.*, 2009.

[97] F. Lescroart *et al.*, "Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development," *Nat. Cell Biol.*, 2014.

[98] Y. Saga, S. Miyagawa-Tomita, A. Takagi, S. Kitajima, J. I. Miyazaki, and T. Inoue, "MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube," *Development*, 1999.

[99] Y. Saga, S. Kitajima, and S. Miyagawa-Tomita, "Mesp1 expression is the earliest sign of cardiovascular development," *Trends in Cardiovascular Medicine*. 2000.

[100] Y. Takahashi, Y. Yasuhiko, S. Kitajima, J. Kanno, and Y. Saga, "Appropriate suppression of Notch signaling by Mesp factors is essential for stripe pattern formation leading to segment boundary formation," *Dev. Biol.*, 2007.

[101] M. Morimoto, M. Kiso, N. Sasaki, and Y. Saga, "Cooperative Mesp activity is required for normal somitogenesis along the anterior-posterior axis," *Dev. Biol.*, 2006.

[102] Y. Takahashi, S. Hiraoka, S. Kitajima, T. Inoue, J. Kanno, and Y. Saga, "Erratum: Differential contributors of Mesp1 and Mesp2 to the epithelialization and rostro-caudal patterning of somites (Development vol. 132 (787-796))," *Development*. 2008.

[103] S. Haraguchi, S. Kitajima, A. Takagi, H. Takeda, T. Inoue, and Y. Saga, "Transcriptional regulation of Mesp1 and Mesp2 genes: Differential usage of enhancers during development," *Mech. Dev.*, 2001.

[104] Y. Takahashi, S. Kitajima, T. Inoue, J. Kanno, and Y. Saga, "Differential contributions of Mesp1 and Mesp2 to the epithelialization and rostro-caudal patterning of somites," *Development*, 2005.

[105] G. Chiapparo *et al.*, "Mesp1 controls the speed, polarity, and directionality of cardiovascular progenitor migration," *J. Cell Biol.*, 2016.

[106] S. Tada *et al.*, "Characterization of mesendoderm: A diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture," *Development*, 2005.

[107] B. Soibam *et al.*, "Genome-wide identification of MESP1 targets demonstrates primary regulation over mesendoderm gene activity," *Stem Cells*, 2015.

[108] M. Beland *et al.*, "Cdx1 Autoregulation Is Governed by a Novel Cdx1-LEF1 Transcription Complex," *Mol. Cell. Biol.*, 2004.

[109] S. Grainger, J. G. A. Savory, and D. Lohnes, "Cdx2 regulates patterning of the intestinal epithelium," *Dev. Biol.*, 2010.

[110] D. G. Silberg, G. P. Swain, Eun Ran Suh, and P. G. Traber, "Cdx1 and Cdx2 expression during intestinal development," *Gastroenterology*, 2000.

[111] E. van den Akker *et al.*, "Cdx1 and Cdx2 have overlapping functions in anteroposterior patterning and posterior axis elongation," *Development*. 2002.

[112] C. Lengerke *et al.*, "Interactions between Cdx genes and retinoic acid modulate early cardiogenesis," *Dev. Biol.*, 2011.

[113] G. Seebohm, S. Peischard, and N. Strutz-Seebohm, "On Track for Cardiac Subtype Specific Differentiation of Embryonic Stem Cells," 2018.

[114] G. Weidinger, C. J. Thorpe, K. Wuennenberg-Stapleton, J. Ngai, and R. T. Moon, "The Sp1-related transcription factors sp5 and sp5-like act downstream of Wnt/β-catenin signaling in mesoderm and neuroectoderm patterning," *Curr. Biol.*, 2005.

[115] Y. Liu *et al.*, "Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells," *Proc. Natl. Acad. Sci. U. S. A.*, 2007.

[116] F. Galvagni *et al.*, "Snai1 promotes ESC exit from the pluripotency by direct repression of self-renewal genes," *Stem Cells*, 2015.

[117] F. Galvagni and F. Neri, "Snai1 represses Nanog to promote embryonic stem cell differentiation," *Genomics Data*, 2015.

[118] C. C. Bell *et al.*, "The Evx1/Evx1as gene locus regulates anterior-posterior patterning during gastrulation," *Sci. Rep.*, 2016.

[119] S. J. Rodda, S. J. Kavanagh, J. Rathjen, P. D. Rathjen, and P. Rathjen, "Embryonic stem cell differentiation and the analysis of mammalian development Pluripotent Cells and Early Mouse Development," *Int. J. Dev. Biol*, 2002.

[120] T. J. Cunningham *et al.*, "Id genes are essential for early heart formation," *Genes Dev.*, 2017.

[121] C. C. Huang, G. D. Orvis, K. M. Kwan, and R. R. Behringer, "Lhx1 is required in Müllerian duct epithelium for uterine development," *Dev. Biol.*, 2014.

[122] S. Kaufhold and B. Bonavida, "Central role of Snail1 in the regulation of EMT and resistance in cancer: A target for therapeutic intervention," *Journal of Experimental and Clinical Cancer Research*. 2014.

[123] L. Gan, S. Schwengberg, and B. Denecke, "Transcriptome analysis in cardiomyocyte-specific differentiation of murine embryonic stem cells reveals transcriptional regulation network," *Gene Expr. Patterns*, 2014.

[124] L. Qian *et al.*, "In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes," *Nature*, 2012.

[125] Q. Lin, J. Schwarz, C. Bucana, and E. N. Olson, "Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C," *Science (80-.).*, vol. 276, no. 5317, pp. 1404–1407, 1997.

[126] K. Song *et al.*, "Heart repair by reprogramming non-myocytes with cardiac transcription factors," *Nature*, 2012.

[127] J. P. Muñoz *et al.*, "The transcription factor MEF2C mediates cardiomyocyte hypertrophy induced by IGF-1 signaling," *Biochem. Biophys. Res. Commun.*, 2009.

[128] Q. Lin, J. Schwarz, C. Bucana, and E. N. Olson, "Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C," *Science (80-.).*, 1997.

[129] M. Ieda *et al.*, "Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors," *Cell*, 2010.

[130] M. Higuchi, T. Kato, S. Yoshida, H. Ueharu, N. Nishimura, and Y. Kato, "PRRX1- and PRRX2-positive mesenchymal stem/progenitor cells are involved in vasculogenesis during rat embryonic pituitary development," *Cell Tissue Res.*, 2015.

[131] J. F. Reiter *et al.*, "Gata5 is required for the development of the heart and endoderm in zebrafish," *Genes Dev.*, 1999.

[132] B. Laforest, G. Andelfinger, and M. Nemer, "Loss of Gata5 in mice leads to bicuspid aortic valve," *J. Clin. Invest.*, 2011.

[133] M. K. Singh *et al.*, "Gata4 and Gata5 cooperatively regulate cardiac myocyte proliferation in mice," *J.*

Biol. Chem., 2010.

[134] D. Wei, H. Bao, N. Zhou, G. F. Zheng, X. Y. Liu, and Y. Q. Yang, "GATA5 loss-of-function mutation responsible for the congenital ventriculoseptal defect," *Pediatr. Cardiol.*, 2013.

[135] M. H. Paul, R. P. Harvey, M. Wegner, and E. Sock, "Cardiac outflow tract development relies on the complex function of Sox4 and Sox11 in multiple cell types," *Cell. Mol. Life Sci.*, 2014.

[136] M. W. Schilham *et al.*, "Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4," *Nature*, 1996.

[137] Y. Kojima *et al.*, "Evolutionarily Distinctive Transcriptional and Signaling Programs Drive Human Germ Cell Lineage Specification from Pluripotent Stem Cells," *Cell Stem Cell*, 2017.

[138] L. Cambier, M. Plate, H. M. Sucov, and M. Pashmforoush, "Nkx2-5 regulates cardiac growth through modulation of Wnt signaling by R-spondin3," *Dev.*, 2014.

[139] J. J. Schott *et al.*, "Congenital heart disease caused by mutations in the transcription factor NKX2-5," *Science (80-.).*, 1998.

[140] R. P. Harvey, "NK-2 homeobox genes and heart development," *Developmental Biology*. 1996.

[141] Y. Sun *et al.*, "Islet 1 is expressed in distinct cardiovascular lineages, including pacemaker and coronary vascular cells," *Dev. Biol.*, 2007.

[142] C. L. Cai *et al.*, "Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart," *Dev. Cell*, 2003.

[143] Y. Qyang *et al.*, "The Renewal and Differentiation of Isl1+ Cardiovascular Progenitors Are Controlled by a Wnt/β-Catenin Pathway," *Cell Stem Cell*, 2007.

[144] K. L. Laugwitz *et al.*, "Postnatal isl1+ cardioblasts enter fully differentiated cardiomyocyte lineages," *Nature*, 2005.

[145] L. Caputo *et al.*, "The Isl1/Ldb1 Complex Orchestrates Genome-wide Chromatin Organization to Instruct Differentiation of Multipotent Cardiac Progenitors," *Cell Stem Cell*, 2015.

[146] L. Bu *et al.*, "Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages," *Nature*, 2009.

[147] E. Dodou, M. P. Verzi, J. P. Anderson, S. M. Xu, and B. L. Black, "Mef2c is a direct transcriptional target of ISL1 and GATA factors in the anterior heart field during mouse embryonic development," *Development*, 2004.

[148] N. Heins *et al.*, "Glial cells generate neurons: The role of the transcription factor Pax6," *Nat. Neurosci.*, 2002.

[149] S. N. Sansom *et al.*, "The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis," *PLoS Genet.*, 2009.

[150] R. L. Chow, C. R. Altmann, R. A. Lang, and A. Hemmati-Brivanlou, "Pax6 induces ectopic eyes in a vertebrate," *Development*, 1999.

[151] X. Zhang *et al.*, "Pax6 is a human neuroectoderm cell fate determinant," *Cell Stem Cell*, 2010.

[152] J. Ericson *et al.*, "Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling," *Cell*, 1997.

[153] L. Jones, G. López-Bendito, P. Gruss, A. Stoykova, and Z. Molnŕ, "Pax6 is required for the normal development of the forebrain axonal connections," *Development*, 2002.

[154] O. Shaham, Y. Menuchin, C. Farhy, and R. Ashery-Padan, "Pax6: A multi-level regulator of ocular development," *Progress in Retinal and Eye Research*. 2012.

[155] Y. W. Hsieh and X. J. Yang, "Dynamic Pax6 expression during the neurogenic cell cycle influences proliferation and cell fate choices of retinal progenitors," *Neural Dev.*, 2009.

[156] V. van Heyningen, "PAX6 in sensory development," *Hum. Mol. Genet.*, 2002.

[157] M. Kohwi, N. Osumi, J. L. R. Rubenstein, and A. Alvarez-Buylla, "Pax6 is required for making specific subpopulations of granule and periglomerular neurons in the olfactory bulb," *J. Neurosci.*, 2005.

[158] N. Osumi, H. Shinohara, K. Numayama-Tsuruta, and M. Maekawa, "Concise Review: Pax6 Transcription Factor Contributes to both Embryonic and Adult Neurogenesis as a Multifunctional Regulator," *Stem Cells*, 2008.

[159] Z. Qin *et al.*, "ZNF536, a Novel Zinc Finger Protein Specifically Expressed in the Brain, Negatively Regulates Neuron Differentiation by Repressing Retinoic Acid-Induced Gene Transcription," *Mol. Cell. Biol.*, 2009.

[160] X. Xu, "Analysis of the Target Genes of Transcription Factor ZNF536 in Lung Adenocarcinoma," 2019.

[161] S. B. Thyme *et al.*, "Phenotypic Landscape of Schizophrenia-Associated Genes Defines Candidates and Their Shared Functions," *Cell*, 2019.

[162] Z. Guo, L. Zhang, Z. Wu, Y. Chen, F. Wang, and G. Chen, "In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model," *Cell Stem Cell*, 2014.

[163] Z. Gao *et al.*, "Neurod1 is essential for the survival and maturation of adult-born neurons," *Nat. Neurosci.*, 2009.

[164] A. Pataskar *et al.*, "NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program," *EMBO J.*, 2016.

[165] C. Boutin *et al.*, "NeuroD1 induces terminal neuronal differentiation in olfactory neurogenesis," *Proc. Natl. Acad. Sci. U. S. A.*, 2010.

[166] T. Matsuda *et al.*, "Pioneer Factor NeuroD1 Rearranges Transcriptional and Epigenetic Profiles to Execute Microglia-Neuron Conversion," *Neuron*, 2019.

[167] C. S. L. Lai, D. Gerrelli, A. P. Monaco, S. E. Fisher, and A. J. Copp, "FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder," *Brain*, 2003.

[168] S. E. Fisher and C. Scharff, "FOXP2 as a molecular window into speech and language," *Trends in Genetics*. 2009.

[169] D. Tsui, J. P. Vessey, H. Tomita, D. R. Kaplan, and F. D. Miller, "FoxP2 regulates neurogenesis during embryonic cortical development," *J. Neurosci.*, 2013.

[170] S. C. Vernes *et al.*, "FOXP2 regulates gene networks implicated in neurite outgrowth in the developing brain," *PLoS Genet.*, 2011.

[171] M. J. Phillips *et al.*, "Modeling human retinal development with patient-specific induced pluripotent stem cells reveals multiple roles for visual system homeobox 2," *Stem Cells*, 2014.

[172] M. A. Cwinn *et al.*, "Suppressor of fused is required to maintain the multipotency of neural progenitor cells in the retina," *J. Neurosci.*, 2011.

[173] C. L. Sigulinsky, M. L. German, A. M. Leung, A. M. Clark, S. Yun, and E. M. Levine, "Genetic chimeras reveal the autonomy requirements for Vsx2 in embryonic retinal progenitor cells," *Neural Dev.*, 2015.

[174] C. Zou and E. M. Levine, "Vsx2 Controls Eye Organogenesis and Retinal Progenitor Identity Via Homeodomain and Non-Homeodomain Residues Required for High Affinity DNA Binding," *PLoS Genet.*, 2012.

[175] B. Bin Xie *et al.*, "Differentiation of retinal ganglion cells and photoreceptor precursors from mouse induced pluripotent stem cells carrying an Atoh7/Math5 lineage reporter," *PLoS One*, 2014.

[176] F. Chiodini *et al.*, "A positive feedback loop between ATOH7 and a notch effector regulates cell-cycle progression and neurogenesis in the retina," *Cell Rep.*, 2013.

[177] L. Prasov *et al.*, "ATOH7 mutations cause autosomal recessive persistent hyperplasia of the primary vitreous," *Hum. Mol. Genet.*, 2012.

[178] R. Sinn, R. Peravali, S. Heermann, and J. Wittbrodt, "Differential responsiveness of distinct retinal domains to Atoh7," *Mech. Dev.*, 2014.

[179] L. Prasov and T. Glaser, "Dynamic expression of ganglion cell markers in retinal progenitors during the terminal cell cycle," *Mol. Cell. Neurosci.*, 2012.

[180] A. C. DeCarvalho, S. L. T. Cappendijk, and J. M. Fadool, "Developmental Expression of the POU Domain Transcription Factor Brn-3b (Pou4f2) in the Lateral Line and Visual System of Zebrafish," *Dev. Dyn.*, 2004.

[181] Y. Fu *et al.*, "Feedback induction of a photoreceptor-specific isoform of retinoid-related orphan nuclear receptor β by the rod transcription factor NRL," *J. Biol. Chem.*, 2014.

[182] L. Jia *et al.*, "Retinoid-related orphan nuclear receptor RORβ is an early-acting factor in rod photoreceptor development," *Proc. Natl. Acad. Sci. U. S. A.*, 2009.

[183] C. L. McGrath *et al.*, "Evidence for genetic association of RORB with bipolar disorder," *BMC Psychiatry*, 2009.

[184] E. M. Mandel *et al.*, "The BMP pathway acts to directly regulate Tbx20 in the developing heart," *Development*, 2010.

[185] S. Chakraborty and K. E. Yutzey, "Tbx20 regulation of cardiac cell proliferation and lineage specialization during embryonic and fetal development in vivo," *Dev. Biol.*, 2012.

[186] F. A. Stennard *et al.*, "Murine T-box transcription factor Tbx20 acts as a repressor during heart development, and is essential for adult heart integrity, function and adaptation," *Development*, 2005.

[187] T. Shen *et al.*, "Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function," *J. Clin. Invest.*, 2011.

[188] D. D. Brown *et al.*, "Tbx5 and Tbx20 act synergistically to control vertebrate heart morphogenesis," *Development*, 2005.

[189] T. F. Plageman and K. E. Yutzey, "T-box genes and heart development: Putting the 'T' in hearT," *Developmental Dynamics*. 2005.

[190] C. L. Cai *et al.*, "T-box genes coordinate regional rates of proliferation and regional specification during cardiogenesis," *Development*, 2005.

[191] F. Greulich, C. Rudat, and A. Kispert, "Mechanisms of T-box gene function in the developing heart," *Cardiovascular Research*. 2011.

[192] M. K. Singh *et al.*, "Tbx20 is essential for cardiac chamber differentiation and repression of Tbx2," *Development*, 2005.

[193] X. Cai *et al.*, "Tbx20 acts upstream of Wnt signaling to regulate endocardial cushion formation and valve remodeling during mouse cardiogenesis," *Dev.*, 2013.

[194] T. F. Plageman and K. E. Yutzey, "Differential Expression and Function of Tbx5 and Tbx20 in Cardiac Development," *J. Biol. Chem.*, 2004.

[195] E. P. Kirk *et al.*, "Mutations in crdiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy," *Am. J. Hum. Genet.*, 2007.

[196] M. G. Posch *et al.*, "A gain-of-function TBX20 mutation causes congenital atrial septal defects, patent foramen ovale and cardiac valve defects," *J. Med. Genet.*, 2010.

[197] F. A. Stennard *et al.*, "Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart," *Dev. Biol.*, 2003.

[198] J. K. Takeuchi *et al.*, "Tbx20 dose-dependently regulates transcription factor networks required for mouse heart and motoneuron development," *Development*, 2005.

[199] N. J. Sakabe *et al.*, "Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function," *Hum. Mol. Genet.*, 2012.

[200] R. K. Patient and J. D. McGhee, "The GATA family (vertebrates and invertebrates)," *Current Opinion in Genetics and Development*. 2002.

[201] V. Garg *et al.*, "GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5," *Nature*, 2003.

[202] A. Holtzinger, G. E. Rosenfeld, and T. Evans, "Gata4 directs development of cardiac-inducing endoderm from ES cells," *Dev. Biol.*, 2010.

[203] Y. S. Ang *et al.*, "Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis," *Cell*, 2016.

[204] C. T. Kuo *et al.*, "GATA4 transcription factor is required for ventral morphogenesis and heart tube formation," *Genes Dev.*, 1997.

[205] A. J. Watt, M. A. Battle, J. Li, and S. A. Duncan, "GATA4 is essential for formation of the proepicardium and regulates cardiogenesis," *Proc. Natl. Acad. Sci. U. S. A.*, 2004.

[206] W. T. Pu, T. Ishiwata, A. L. Juraszek, Q. Ma, and S. Izumo, "GATA4 is a dosage-sensitive regulator of cardiac morphogenesis," *Dev. Biol.*, 2004.

[207] J. D. Molkentin, Q. Lin, S. A. Duncan, and E. N. Olson, "Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis," *Genes Dev.*, 1997.

[208] C.-L. Lien, J. McAnally, J. A. Richardson, and E. N. Olson, "Cardiac-specific activity of an Nkx2--5 enhancer requires an evolutionarily conserved Smad binding site," *Dev. Biol.*, vol. 244, no. 2, pp. 257–266, 2002.

[209] D. A. Elliott *et al.*, "NKX2-5 eGFP/w hESCs for isolation of human cardiac progenitors and cardiomyocytes," *Nat. Methods*, 2011.

[210] D. J. Anderson *et al.*, "NKX2-5 regulates human cardiomyogenesis via a HEY2 dependent transcriptional network," *Nat. Commun.*, 2018.

[211] L. Qian *et al.*, "Tinman/Nkx2-5 acts via miR-1 and upstream of Cdc42 to regulate heart function across species," *J. Cell Biol.*, 2011.

[212] J. M. Reecy *et al.*, "Identification of upstream regulatory regions in the heart-expressed homeobox gene Nkx2-5," *Development*, 1999.

[213] J. Brocher, B. Vogel, and R. Hock, "HMGA1 down-regulation is crucial for chromatin composition and a gene expression profile permitting myogenic differentiation," *BMC Cell Biol.*, 2010.

[214] R. Reeves and L. Beckerbauer, "HMGI/Y proteins: Flexible regulators of transcription and chromatin structure," *Biochimica et Biophysica Acta - Gene Structure and Expression*. 2001.

[215] M. Harrer, H. Lührs, M. Bustin, U. Scheer, and R. Hock, "Dynamic interaction of HMGA1a proteins with chromatin," *J. Cell Sci.*, 2004.

[216] R. Reeves, "Structure and function of the HMGI(Y) family of architectural transcription factors," *Environ. Health Perspect.*, 2000.

[217] M. Fedele *et al.*, "Haploinsufficiency of the Hmga1 gene causes cardiac hypertrophy and myelo-lymphoproliferative disorders in mice," *Cancer Research*. 2006.

[218] S. Mendjan *et al.*, "NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency," *Cell Stem Cell*, vol. 15, no. 3, pp. 310–325, 2014.

[219] S. Masui *et al.*, "Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells," *Nat. Cell Biol.*, vol. 9, no. 6, p. 625, 2007.

[220] F. Gao, S. W. Kwon, Y. Zhao, and Y. Jin, "PARP1 poly (ADP-ribosyl) ates Sox2 to control Sox2 protein levels and FGF4 expression during embryonic stem cell differentiation," *J. Biol. Chem.*, vol. 284, no. 33, pp. 22263–22273, 2009.

[221] M. J. Murphy, A. Wilson, and A. Trumpp, "More than just proliferation: Myc function in stem cells," *Trends Cell Biol.*, vol. 15, no. 3, pp. 128–137, 2005.

[222] T. Akagi, S. Kuure, K. Uranishi, H. Koide, F. Costantini, and T. Yokota, "ETS-related transcription factors ETV4 and ETV5 are involved in proliferation and induction of differentiation-associated genes in embryonic stem (ES) cells," *J. Biol. Chem.*, vol. 290, no. 37, pp. 22460–22473, 2015.

[223] C. Harmelink, Y. Peng, P. DeBenedittis, H. Chen, W. Shou, and K. Jiao, "Myocardial Mycn is essential for mouse ventricular wall morphogenesis," *Dev. Biol.*, vol. 373, no. 1, pp. 53–63, 2013.

[224] D. G. McFadden, A. C. Barbosa, J. A. Richardson, M. D. Schneider, D. Srivastava, and E. N. Olson, "The Hand1 and Hand2 transcription factors regulate expansion of the embryonic cardiac ventricles in a gene dosage-dependent manner," *Development*, vol. 132, no. 1, pp. 189–201, 2005.

[225] Y. Morikawa and P. Cserjesi, "Cardiac neural crest expression of Hand2 regulates outflow and second heart field development," *Circ. Res.*, vol. 103, no. 12, pp. 1422–1429, 2008.

[226] B. D. Thattaliyath, C. B. Livi, M. E. Steinhelper, G. M. Toney, and A. B. Firulli, "HAND1 and HAND2 are expressed in the adult-rodent heart and are modulated during cardiac hypertrophy," *Biochem. Biophys. Res. Commun.*, vol. 297, no. 4, pp. 870–875, 2002.

[227] R. M. Barnes *et al.*, "Hand2 loss-of-function in Hand1-expressing cells reveals distinct roles in epicardial and coronary vessel development," *Circ. Res.*, vol. 108, no. 8, pp. 940–949, 2011.

[228] O. Machon, J. Masek, O. Machonova, S. Krauss, and Z. Kozmik, "Meis2 is essential for cranial and cardiac neural crest development," *BMC Dev. Biol.*, vol. 15, no. 1, p. 40, 2015.

[229] G. Mehta *et al.*, "MITF interacts with the SWI/SNF subunit, BRG1, to promote GATA4 expression in cardiac hypertrophy," *J. Mol. Cell. Cardiol.*, vol. 88, pp. 101–110, 2015.

[230] S.-W. Chang *et al.*, "Genetic abnormalities in FOXP1 are associated with congenital heart defects," *Hum. Mutat.*, vol. 34, no. 9, pp. 1226–1230, 2013.

[231] J. J. Liu *et al.*, "miR-218 involvement in cardiomyocyte hypertrophy is likely through targeting REST," *Int. J. Mol. Sci.*, vol. 17, no. 6, Jun. 2016.

[232] A. Carè *et al.*, "MicroRNA-133 controls cardiac hypertrophy," *Nat. Med.*, vol. 13, no. 5, pp. 613–618, May 2007.

[233] Q. R. Lu *et al.*, "Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection," *Cell*, vol. 109, no. 1, pp. 75–86, 2002.

[234] Y. Liu, P. Jiang, and W. Deng, "OLIG gene targeting in human pluripotent stem cells for motor neuron and oligodendrocyte differentiation," *Nat. Protoc.*, vol. 6, no. 5, p. 640, 2011.

[235] Y. Liu and M. S. Rao, "Olig genes are expressed in a heterogeneous population of precursor cells in the developing spinal cord," *Glia*, vol. 45, no. 1, pp. 67–74, 2004.

[236] D. Nozawa, N. Suzuki, M. Kobayashi-osaki, X. Pan, J. D. Engel, and M. Yamamoto, "GATA2-dependent and region-specific regulation of Gata2 transcription in the mouse midbrain," *Genes to Cells*, 2009.

[237] K. Lilleväli, M. Haugas, F. Pituello, and M. Salminen, "Comparative analysis of Gata3 Gata2 expression during chicken inner ear development," *Dev. Dyn.*, 2007.

[238] G. Sheng and C. D. Stern, "Gata2 and Gata3: novel markers for early embryonic polarity and for non-neural ectoderm in the chick embryo," *Mech. Dev.*, 1999.

[239] A. El Wakil, C. Francius, A. Wolff, J. Pleau-Varet, and J. Nardelli, "The GATA2 transcription factor negatively regulates the proliferation of neuronal progenitors," *Development*, 2006.

[240] M. Pieper, K. Ahrens, E. Rink, A. Peter, and G. Schlosser, "Differential distribution of competence for panplacodal and neural crest induction to non-neural and neural ectoderm," *Development*, 2012.

[241] K. Tsarovina *et al.*, "Essential role of Gata transcription factors in sympathetic neuron development," *Development*, 2004.

[242] D. Kerschensteiner *et al.*, "Genetic control of circuit function: Vsx1 and Irx5 transcription factors regulate contrast adaptation in the mouse retina," *J. Neurosci.*, 2008.

[243] M. Vitorino, P. R. Jusuf, D. Maurus, Y. Kimura, S. I. Higashijima, and W. A. Harris, "Vsx2 in the zebrafish retina: Restricted lineages through derepression," *Neural Dev.*, 2009.

[244] C. Francius *et al.*, "Vsx1 transiently defines an early intermediate V2 interneuron precursor compartment in the mouse developing spinal cord," *Front. Mol. Neurosci.*, 2016.

[245] X. Xu, Y. He, L. Sun, S. Ma, and C. Luo, "Maternal Vsx1 plays an essential role in regulating prechordal mesendoderm and forebrain formation in zebrafish," *Dev. Biol.*, 2014.

[246] T. L. Hoffman, A. L. Javier, S. A. Campeau, R. D. Knight, and T. F. Schilling, "Tfap2 transcription factors in zebrafish neural crest development and ectodermal evolution," *J. Exp. Zool. Part B Mol. Dev. Evol.*, 2007.

[247] W. A. Pastor *et al.*, "TFAP2C regulates transcription in human naive pluripotency by opening enhancers," *Nat. Cell Biol.*, 2018.

[248] W. Li and R. A. Cornell, "Redundant activities of Tfap2a and Tfap2c are required for neural crest induction and development of other non-neural ectoderm derivatives in zebrafish embryos," *Dev. Biol.*, 2007.

[249] J. D. Burrill, L. Moran, M. D. Goulding, and H. Saueressig, "PAX2 is expressed in multiple spinal cord interneurons, including a population of EN1+ interneurons that require PAX6 for their development," *Development*, vol. 124, no. 22, pp. 4493–4503, 1997.

[250] C. Soukkarieh, E. Agius, C. Soula, and P. Cochard, "Pax2 regulates neuronal--glial cell fate choice in the embryonic optic nerve," *Dev. Biol.*, vol. 303, no. 2, pp. 800–813, 2007.

[251] J. Terzić, C. Muller, S. Gajović, and M. Saraga-Babić, "Expression of PAX2 gene during human development," *Int. J. Dev. Biol.*, 1998.

[252] J. Stanke, H. E. Moose, H. M. El-Hodiri, and A. J. Fischer, "Comparative study of Pax2 expression in glial

cells in the retina and optic nerve of birds and mammals," *J. Comp. Neurol.*, 2010.

[253] M. Schwarz, G. Alvarez-Bolado, G. Dressler, Pavel Urbánek, M. Busslinger, and P. Gruss, "Pax2/5 and Pax6 subdivide the early neural tube into three domains," *Mech. Dev.*, 1999.

[254] C. Soukkarieh, E. Agius, C. Soula, and P. Cochard, "Pax2 regulates neuronal-glial cell fate choice in the embryonic optic nerve," *Dev. Biol.*, 2007.

[255] P. L. Pfeffer, M. Bouchard, and M. Busslinger, "Pax2 and homeodomain proteins cooperatively regulate a 435 bp enhancer of the mouse Pax5 gene at the midbrain-hindbrain boundary," *Development*, 2000.

[256] Y. Zhao *et al.*, "Control of hippocampal morphogenesis and neuronal differentiation by the LIM homeobox gene Lhx5," *Science (80-.).*, 1999.

[257] H. Z. Sheng *et al.*, "Expression of murine Lhx5 suggests a role in specifying the forebrain," *Dev. Dyn.*, 1997.

[258] J. Inoue, Y. Ueda, T. Bando, T. Mito, S. Noji, and H. Ohuchi, "The expression of LIM-homeobox genes, Lhx1 and Lhx5, in the forebrain is essential for neural retina differentiation," *Dev. Growth Differ.*, 2013.

[259] Y. Zhao *et al.*, "LIM-homeodomain proteins Lhx1 and Lhx5, and their cofactor Ldb1, control Purkinje cell differentiation in the developing cerebellum," *Proc. Natl. Acad. Sci. U. S. A.*, 2007.

[260] S. Bertuzzi *et al.*, "Molecular cloning, structure, and chromosomal localization of the mouse LIM/homeobox gene Lhx5," *Genomics*, 1996.

[261] A. C. Cepeda-Nieto, S. L. Pfaff, and A. Varela-Echavarría, "Homeodomain transcription factors in the development of subsets of hindbrain reticulospinal neurons," *Mol. Cell. Neurosci.*, 2005.

[262] T. Kawaue *et al.*, "Lhx1 in the proximal region of the optic vesicle permits neural retina development in the chicken," *Biol. Open*, 2012.

[263] M. Yossy, "Identification of novel determinants of dopaminergic and isotocinergic neural fates: LHX5 and wnt-antagonists act downstream of the conserved transcriptional regulator FEZF2.," 2014.

[264] A. Fiorino *et al.*, "Retina-derived POU domain factor 1 coordinates expression of genes relevant to renal and neuronal development," *Int. J. Biochem. Cell Biol.*, 2016.

[265] A. Cvekl and X. Zhang, "Signaling and Gene Regulatory Networks in Mammalian Lens Development," *Trends in Genetics*. 2017.

[266] K. Schäfer, P. Neuhaus, J. Kruse, and T. Braun, "The homeobox gene Lbx1 specifies a subpopulation of cardiac neural crest necessary for normal heart development," *Circ. Res.*, 2003.

[267] M. Krüger, K. Schäfer, and T. Braun, "The homeobox containing gene Lbx1 is required for correct dorsal-ventral patterning of the neural tube," *J. Neurochem.*, 2002.

[268] S. Schmitteckert, C. Ziegler, L. Kartes, and A. Rolletschek, "Transcription factor Lbx1 expression in mouse embryonic stem cell-derived phenotypes," *Stem Cells Int.*, 2011.

[269] A. G. Nadadhur *et al.*, "Patterning factors during neural progenitor induction determine regional identity and differentiation potential in vitro," *Stem Cell Res.*, 2018.

[270] L. Ji and K. L. Tan, "Identifying time-lagged gene clusters using gene expression data," *Bioinformatics*, 2005.

[271] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein, "Comparing protein abundance and mRNA expression levels on a genomic scale," *Genome Biology*. 2003.

[272] L. Peshkin *et al.*, "On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development," *Dev. Cell*, 2015.

[273] F. Edfors *et al.*, " Gene-specific correlation of RNA and protein levels in human cells and tissues ," *Mol. Syst. Biol.*, 2016.

[274] S. Zeidler *et al.*, "Computational detection of stage-specific transcription factor clusters during heart development," *Front. Genet.*, vol. 7, no. MAR, Mar. 2016.

PLOS ONE

RESEARCH ARTICLE

Constructing temporal regulatory cascades in the context of development and cell differentiation

Rayan Daou[1], **Tim Beißbarth**[1], **Edgar Wingender**[1], **Mehmet Gültas**[2,3], **Martin Haubrock**[1]*

1 Department of Medical Bioinformatics, University Medical Center Göttingen, Goettingen, Niedersachsen, Germany, **2** Breeding Informatics Group, Department of Animal Science, Georg-August University, Goettingen, Niedersachsen, Germany, **3** Center for Integrated Breeding Research (CiBreed), Georg-August University, Goettingen, Niedersachsen, Germany

* martin.haubrock@bioinf.med.uni-goettingen.de

OPEN ACCESS

Citation: Daou R, Beißbarth T, Wingender E, Gültas M, Haubrock M (2020) Constructing temporal regulatory cascades in the context of development and cell differentiation. PLoS ONE 15(4): e0231326. https://doi.org/10.1371/journal.pone.0231326

Editor: Roberto Mantovani, Università degli Studi di Milano, ITALY

Received: October 9, 2019

Accepted: March 20, 2020

Published: April 10, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0231326

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Abstract

Cell differentiation is a complex process orchestrated by sets of regulators precisely appearing at certain time points, resulting in regulatory cascades that affect the expression of broader sets of genes, ending up in the formation of different tissues and organ parts. The identification of stage-specific master regulators and the mechanism by which they activate each other is a key to understanding and controlling differentiation, particularly in the fields of tissue regeneration and organoid engineering. Here we present a workflow that combines a comprehensive general regulatory network based on binding site predictions with user-provided temporal gene expression data, to generate a a temporally connected series of stage-specific regulatory networks, which we call a temporal regulatory cascade (TRC). A TRC identifies those regulators that are unique for each time point, resulting in a cascade that shows the emergence of these regulators and regulatory interactions across time. The model was implemented in the form of a user-friendly, visual web-tool, that requires no expert knowledge in programming or statistics, making it directly usable for life scientists. In addition to generating TRCs the tool links multiple interactive visual workflows, in which a user can track and investigate further different regulators, target genes, and interactions, directing the tool along the way into biologically sensible results based on the given dataset. We applied the TRC model on two different expression datasets, one based on experiments conducted on human induced pluripotent stem cells (hiPSCs) undergoing differentiation into mature cardiomyocytes and the other based on the differentiation of H1-derived human neuronal precursor cells. The model was successful in identifying previously known and new potential key regulators, in addition to the particular time points with which these regulators are associated, in cardiac and neural development.

Funding: R.D FKZ:81X2300184 DZHK https://dzhk.de The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The work was supported by the DZHK (German Center for Cardiovascular Research; FKZ:81X2300184) where DZHK stands for (Deutsches Zentrum für Herz-Kreislauf-Forschung).

Competing interests: The authors have declared that no competing interests exist.

Introduction

Cell differentiation, the building block of development, is a strong representation of regulatory precision. In stem cell differentiation, a handful of regulators kick off a regulatory mechanism that leads to the activation or repression of other regulators and non-regulatory genes, through consecutive waves, starting processes that are geared towards specification and giving rise to different kinds of cells and tissues [1–5]. The discovery of induced pluripotent stem cells (iPSCs) [6–8], opened the door to a rising number of cell differentiation experiments. Owing to the decreasing prices of RNA-seq, these experiments generated a big and growing number of time series datasets that aim to track a certain process of differentiation by taking snapshots of the gene expression at different time points. These datasets could be further analyzed to obtain a better extensive explanatory model of the regulatory processes and to identify new important regulators that can be manipulated to enhance the process. Deriving as much information as possible from such experiments is a crucial goal in the fields of medical and biological research [9–13], yet there is still a need for computational methods that analyze such unique models in a way tailored to their special properties.

One common challenge to the researchers in these fields is identifying a set of candidate genes that are crucial for the study case, from the thousands of genes in the dataset, that if manipulated can impact the quality and the outcome of the process under study. This candidate set has to be small enough to make the experimental validation of each candidate feasible. One approach is constructing co-expression networks, clustering the genes into modules, usually large ones, then attempting to reduce these modules based on topological feautures [14]. Other approaches, like Short Time-series Expression Miner (STEM), find statistically significant gene patterns and the genes associated with them [15]. Differential gene expression (DEG) analysis is one of the most popular methods to create lists of genes that can be stage-associated. DEG lists provide a good start but often are large in size, and the stage-specific regulators often get diluted in more general genes leading to rather general GO terms when enriched. TFRank [16] is a popular network-based prioritization method, but it doesn't integrate time series expression data. There are other different approaches to prioritize genes and reducing gene lists resulting from previous methods [17], yet none of these specifically take into account the unique properties of cell differentiation.

Another challenge lies in identifying and understanding the important regulatory interactions and programs that trigger and control the expression of different essential genes. One of the most useful and general approaches to address these regulatory programs is via constructing gene regulatory networks (GRNs), typically a directed graph with the genes as nodes and the edges connecting the nodes usually indicating the regulatory interactions. In the past years, many methods and models have been developed to construct GRNs based on either expression data [18], Chip-seq, binding sites analysis, or other data types and models. Some of these models depend solely on one data type to build these networks while others more effectively combine one or more data sources. Despite the general success of some methods which derive GRNs from gene expression data, they have commonly known limitations, such as the inability to deal with time series data in the case of Bayesian Networks (BNs), excessive computational time in the case of Dynamic Baysian Networks (DBNs), and the fact that the number of genes is mostly greater than the number of experimental conditions can cause problems when it comes to methods like Graphical Gaussian Models (GGMs) and BNs [19, 20]. Another different approach is using binding site analysis in the genome to predict the capability of transcription factors (TFs) to regulate the expression of target genes. TFs have the potential to bind to a DNA region via a binding site with a specific pattern of nucleotides that can be recognized by the DNA-binding domain (DBD) associated with each TF. The challenge in this approach lies

mainly in finding the proper library of positional weight matrices (PWMs), the ideal thresholds, and cutoffs and defining the regions of search. The result is an extensive regulatory network that covers a large number of potential regulatory interactions. While these regulatory effects are potentially possible, only a subset of these interactions takes place in a specific context and time. Finding these subsets and refining the global regulatory network according to the biological context under study would result in a more meaningful and case-relevant network.

To tackle these challenges, we constructed a novel workflow and a model of a regulatory network that incorporates the element of time and temporal order, integrates the expression levels of genes, is concise enough to be inspected visually, and identifies candidate regulators efficiently. The method is time and memory efficient, yet it generates a model with a specific architecture to display the primary transcriptional regulators, such as TF genes and miRNAs, and regulatory events unfolding with time. It pre-computes an extensive gene regulatory network that is based on binding site analysis, is independent of the expression data and is used as a background regulatory network. The workflow then uses expression data to identify stage-specific regulators based on their expression pattern. These regulators are finally organized in a cascade architecture that we call a temporal regulatory cascade (TRC). In a TRC, master regulators specific for each stage are organized in ordered vertical columns, and potential regulatory interactions that are based on the background network are displayed as edges between these regulators. To demonstrate this model, we developed an online tool aimed for experimentalists as well as bioinformaticians interested in investigating the regulatory forces that might explain the observed expression of genes in a particular time series dataset. Our novel workflow offers the automatic generation of a TRC from an uploaded time series dataset and visualizes it in an animated interactive manner. In order to facilitate direct interpretation, the results at any stage of the workflow are distilled to an amount that can be handled and analyzed visually, keeping the top significant genes, interactions, and information and discarding those with lower significance and specificity.

In this manuscript, we describe the workflow in detail and report on its application to two time series expression datasets. Both datasets characterize the differentiation of pluripotent stem cells into mature cardiac myocytes and neural progenitors, and the corresponding TRC was generated and analyzed in each case. The main aim was to analyze the specific regulatory activity in each stage, identify and evaluate regulators specific for each time point in the differentiation process, and to test the efficiency of the workflow in re-identifying some well-known case-relevant regulators and regulatory interactions without prior knowledge.

Materials and methods

Background regulatory network

A library of position weight matrices (PWMs) from TRANSFAC® [21] is used in combination with the MATCH™ [22] program to predict transcription factor binding sites (TFBSs) in the conserved promoter regions of the human genome as follows.

Based on 49,344 RefSeq-annotated human transcription units (UCSC track refGene, Jan. 22, 2014), the -1kb upstream region was selected as a proximal promoter. The transcription start site (TSS) indicated in RefSeq was used as the reference point.

On the basis of pre-calculated whole genome alignments provided by the UCSC (46_WAY_MULTIZ_hg19) these promoter definitions were utilized to retrieve the sequence conserved regulatory regions between human (hg19), mouse (mm9), dog (canFam2) and cow (bosTau4). Afterwards, gaps resulting from the multiple genome alignment were removed.

MATCH was used to predict potential TFBSs in the previously identified conserved promoter regions, based on all vertebrate defined matrices using the PWM library from TRANSFAC (release 2013.1, 1446 vertebrate matrices). All matrices with default minFN threshold (minimize false negatives) were used in order to predict potential TFBSs that have at least the quality of an annotated TFBS in TRANSFAC. 1360 out of 1446 TRANSFAC-PWMs had a sequence-conserved TFBS prediction. We ranked all predicted TFBSs associated with each PWM, according to their MATCH score. We chose the best 5% predicted binding sites for each PWM and constructed the background transcriptional regulatory network accordingly. The PWMs are translated to human TF-gene names (HGNC-defined) using the TRANSFAC database. Each TF-gene, identified by its official HGNC-defined gene name, was represented as a node, with a directed edge connecting it with its target gene node. Further information about the construction of the regulatory network can be found in our previous manuscript [23].

The core network included 829 TFs and their 16354 targets, summing up to 749949 interactions. Another expanded network, which includes microRNA binding predictions, was constructed and contained 2239 regulators and 20160 targets. This network was computed once and is independent in the process of its derivation from the expression data, making it usable with every human expression dataset.

While the tool offers the user the option to upload a custom regulatory network to be used for the analysis, we recommend the built-in network just described. The conservation property of these sites makes the prediction ideal for the differentiation context since several pieces of research have shown that conserved regions in the DNA are critical binding sites for development and differentiation [24–27].

Temporal regulatory cascades (TRCs)

The method utilizes the concept of constructing template expression patterns that represent an expression behaviour of interest, then attracting genes that behave similarly to these patterns using correlation. The template patterns we used were stage-specific patterns, peaking at one time point only, and denoted by template peak patterns (TPPs). While different kinds of template patterns can be used, we chose the single-peak TPPs, as a default for its ability to attract stage-specific regulators that are unique to each time point. Regulatory interactions are queried from the background regulatory network and form the edges between the genes in the cascade accordingly.

TRC construction steps.

Step 1: Create a library of TPPs, one TPP for each time point in the dataset. For each time point the corresponding TPP has an expression level of 100 percent at that time point and zero every other time point(Fig 1A).

Step 2: For each TPP, calculate the top *n* correlated regulators to this reference pattern (Fig 1B). These genes are said to be the stage-specific regulators of stage *s* and are displayed in the same column (Fig 2). If a time point has no correlated regulators, no column is created for this stage in the TRC.

Step 3: All regulatory interactions between the regulators of the same stage are mapped, according to their connections in the background regulatory network, and represented in the form of directed edges.

Step 4: All regulatory interactions between the regulators of stage *s* and the next stage are mapped according to their connections in the background regulatory network and

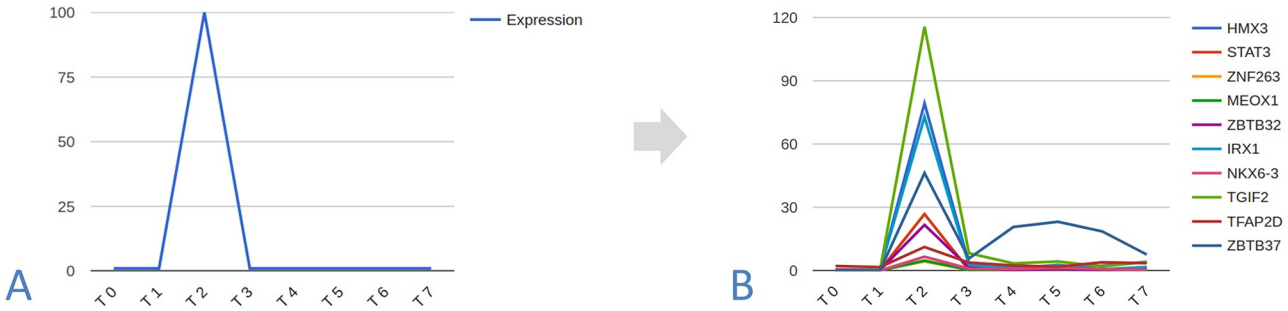

Fig 1. Indetifying stage-specific regulators. (A) The TPP of T2: The template peaking pattern is calculated where the expression is at 100 percent T2 and zero every other timepoint. One TPP for each time point is calculated similarly and the collection of these TPPs form the TPP library. (B) Top 10 regulators that are highly correlated with the previous TPP of T2 and their noticeable T2-specific peaking pattern, these regulators will form nodes in the T2 column in the TRC, the same is done for every TPP in the library.

https://doi.org/10.1371/journal.pone.0231326.g001

represented by directed edges, linking each stage to the next and tying the cascade together (Fig 2).

Parameters. To adjust the temporal regulatory cascade, we use three primary parameters:

minE: A threshold for gene expression levels. A gene that does not have an expression level higher than this threshold in any of the replicates or time points is eliminated and omitted from the calculation that leads to the TRC. This eliminates peaking genes that are lowly expressed even at their peak.

minC: A minimum correlation threshold. Regulatory genes that have a correlation above this threshold to the TPP of a stage are kept as the initial set of regulators associated with that stage.

maxS: The maximum number of genes that can be associated with a specific time point. The initial regulators associated with a time point based on *minC* are sorted by their correlation to the TPP of that stage, and the top *n* (*maxS*) regulators are picked to be in the column associated with the stage. If the initial regulators set has less than *maxS* genes, then the whole set is taken. The max number of nodes in the cascade is *maxS* multiplied by the number of time points.

Implementation

This workflow was implemented in the form of a web service with an interactive visual web interface, which eliminates then the need to install any additional software. The algorithm to generate the TRCs was implemented using Java. In order to display the resulting TRC, a visualizer was implemented using JavaScript, utilizing, and extending the Cytoscape library cy.js. The framework used PHP to manage the files and sessions. The visualizer was embedded in an interactive webpage that includes helpful information such as graphs of the expression levels of the genes in the cascade, tables, and metrics, in addition to direct links to perform GO enrichment and other workflows in the platform. The web tool is a part of a more comprehensive web service that revolves around gene regulation and expression data analysis that is under construction.

Data

While any formatted time series data can be used as input, this model performs the best with RNA-Seq data over other sources of inferior quality and less variability such as microarray

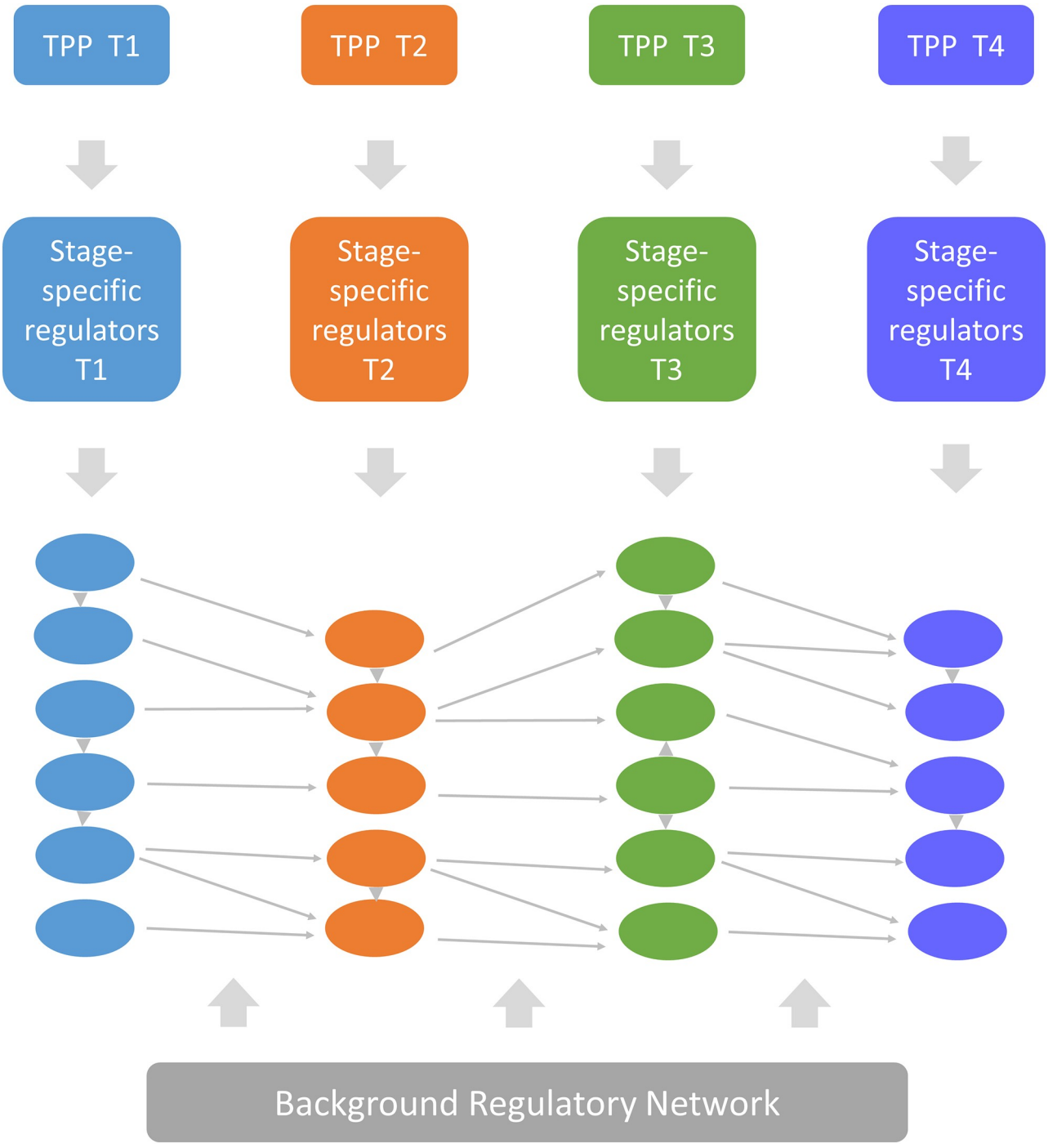

Fig 2. The TRC workflow. Regulators specific for each time point are grouped in the same column with the same color and sorted by their correlation to the TPP of that stage. The edges between the two stages and within the same stage are retrieved and mapped from the regulatory network.

https://doi.org/10.1371/journal.pone.0231326.g002

data. Normalized input data provides a better quality TRC, nevertheless even using the raw counts leads to reasonably significant TRCs. As study cases to demonstrate the TRC model, two time series gene expression datasets were used and denoted Dataset1 and Dataset2.

Dataset1 was assembled using public RNA-Seq data that is captured during the differentiation of H1 derived human neuronal precursor cells (NPCs) across the days 0,1,2,4,5,11, and 18 after induction of neuronal differentiation. Publicly available DEG and GO enrichment analysis on the same dataset was used for comparison. The dataset and the analysis results could be found in the expression Atlas under the accession E-GEOD-56785. The assembled and formatted data can be found in S1 File.

Dataset2 was derived from the normalized expression datasets from the previously published study by Qing Liu et al [28], publicly available in the GEO repository under the accession number GSE85332. We chose one of the four expression datasets available, the RNA-Seq profiling of the differentiation of C20 derived cardiomyocytes at four stages: pluripotent stem cells (day 0), mesoderm (day 2), cardiac mesoderm (day 4), and differentiated cardiomyocytes (day 30). The assembled and formatted data can be found in S2 File.

GO enrichment

To evaluate the relevance of the gene sets in each stage, Gene Ontology (GO) enrichment analysis using the biological processes and a Fisher's Exact test on each column in these cascades was applied using one set at a time as an input. Terms that have a pvalue less than 0.05 after the Bonferroni correction are sorted by their fold enrichment and the top terms were examined. These terms were evaluated based on their consistency with the differentiation stage under observation at that time point.

Results

We applied the TRC workflow to Dataset1 and Dataset2 and generated a cascade for each study case. In addition to the GO enrichment, detailed literature research was performed, investigating the roles of the different regulators predicted by the cascade.

Neural differentiation cascade

Upon the visual inspection of the cascade, we observe a missing time point that is day 2, indicating that this time point does not have any peak strength or any genes that exceed the correlation threshold to the TPP, suggesting that day 2 might be a time point that doesn't underly any unique stage-specific activity (Fig 3).

Examining the GO enrichment of each time point reveals high enrichment of relevant terms in day 1 and day 11. Regulators of day 1 showed enrichment for specific terms such as cell and neuron fate commitment, neuron differentiation, and cell differentiation in the spinal cord. Regulators of day 11 showed high enrichment of even more specific terms such as spinal cord association neuron differentiation, dorsal spinal cord development, cell fate determination, cell differentiation in the spinal cord, hindbrain development. On the other hand, examining the GO enrichment based on the DEG analysis publicly available for the same dataset, differentially expressed genes in day 0 vs. day 1 and day 0 vs. day 11 showed no significant enrichment of specific terms associated with neural development but rather more general terms.

A deeper look into the identity of the regulators in the cascade shows that *OLIG1* and *OLIG3*, which are known for their importance in neural and spinal development [29–31], are active in day 1, suggesting that their importance lies in the earlier part of the differentiation. A microRNA *MIR3659* peaking at day 1 with a high indegree raises the question on the nature of its involvement in neural differentiation, which needs to be further investigated. *PAX2* on day 11, with the highest outdegree, regulates 13 different regulators in the same and next time point which hints that its known essential role in neural development [32–34] is due to its

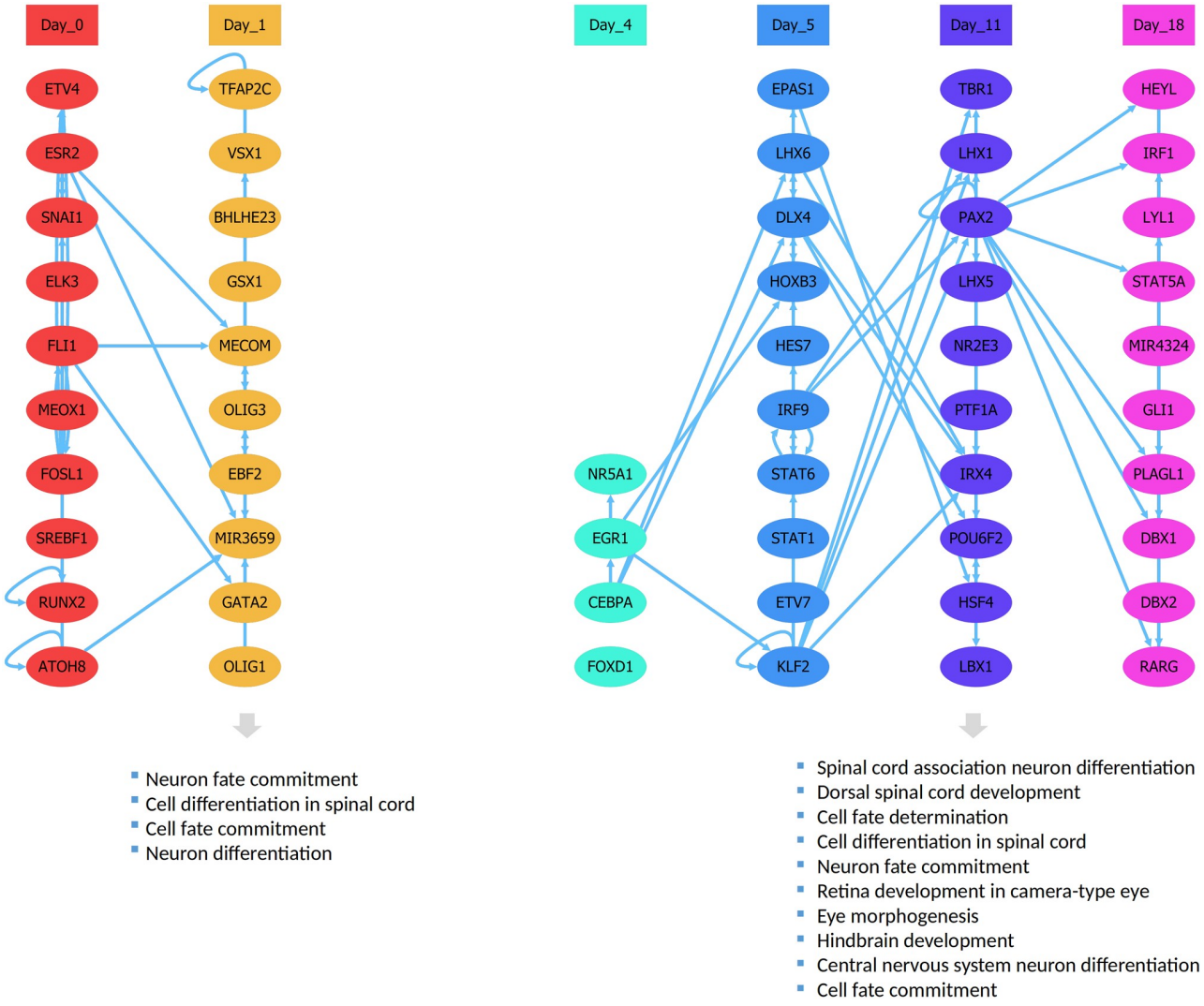

Fig 3. Neural differentiation cascade. The TRC generated for the differentiation of neural proginators based on dataset 1 and the following parameters: *minC = 0.6*, *minE = 4*, and *maxS = 10*.

https://doi.org/10.1371/journal.pone.0231326.g003

regulatory impact on a big set of neural regulators. *KLF2* in day 5 stands out as a potential significant regulator of the day 11 regulatory wave due to its potential ability to regulate a big portion of day 11 regulators. The TRC shows an overall same-stage presence of certain TFs that belong to the same family or subfamily according to the classification experimental conditions TFs in TFClass [35, 36], such as *OLIG1*, *OLIG3* and *BHLHE23* in day 1, *STAT1* and *STAT6* in day 5, the *LHX1* and *LHX5* in day 11, *DBX1* and *DBX2* in day 18. A hypothesis can be made that these TFs are part of the redundancy that leads to the robustness of such regulatory programs, or that these families and subfamilies of TFs collaborate in certain regulatory stages.

Cardiac differentiation cascade

Regulators of the first time point show enrichment of terms related to stem cell maintenance, which is coherent with the biological context since the process of differentiation has not started yet, and the cells are still in the induced stem cell state (Fig 4). These regulators could be

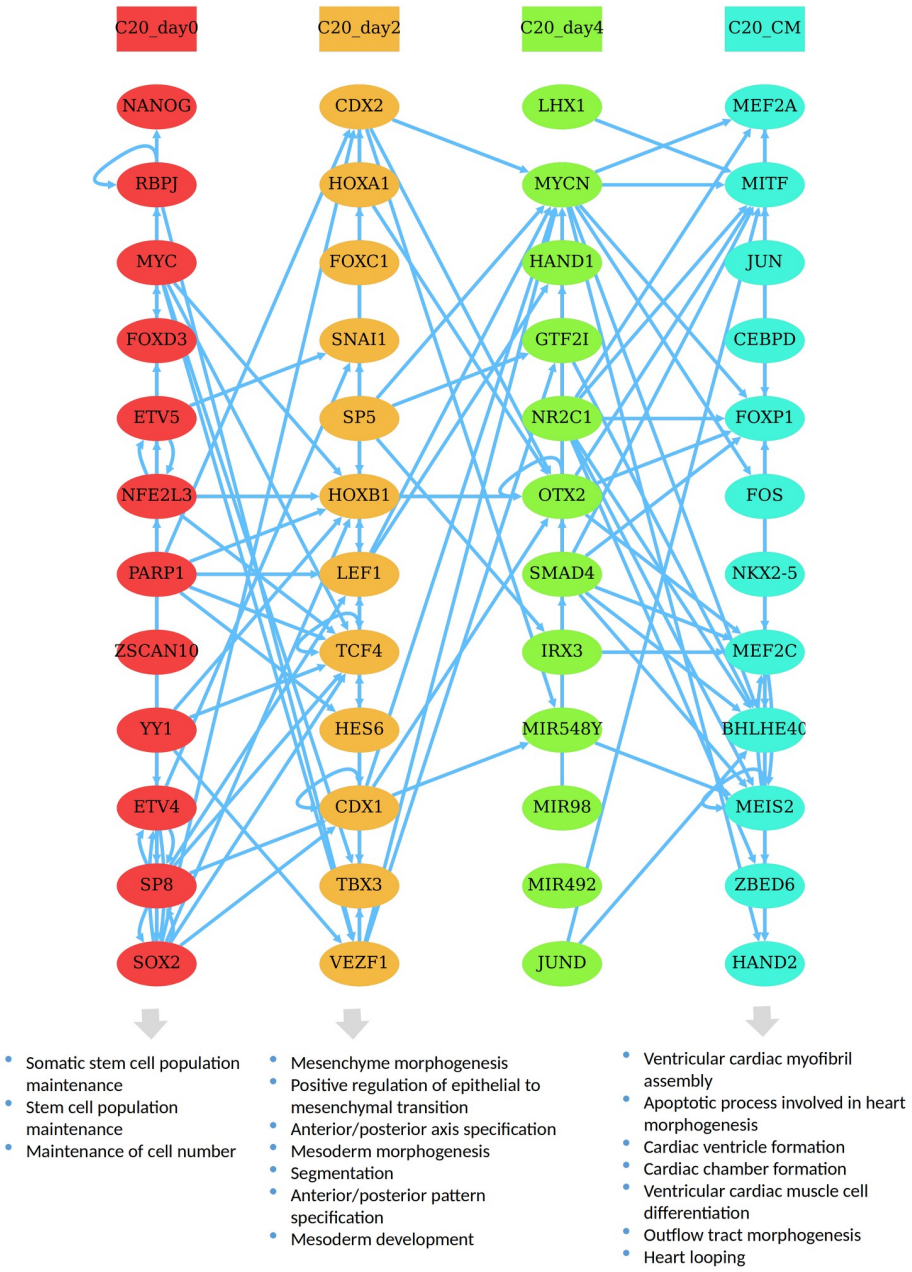

Fig 4. Cardiac differentiation cascade. The TRC generated for the differentiation of cardiomyocytes based on dataset2 and the following parameters: *minC* = *0.6*, *minE* = *30*, and *maxS* = *12*.

https://doi.org/10.1371/journal.pone.0231326.g004

essential for maintaining the pluripotency state and also could be repressing differentiation. Regulators of day 2 show enrichment of terms associated with mesenchymal and mesoderm morphogenesis, which give rise to cardiac cells. Regulators of the last stage the cardiomyocyte (CM) stage show high enrichment of heart-specific terms such as cardiac ventricle and chamber formation, ventricular cardiac muscle differentiation, heart looping, and outflow tract morphogenesis. These terms show a high consistency with the underlying stage of differentiation reported by the experiment.

In the first time point TFs associated with maintanining the pluripotency state like *NANOG* [37], *PARP1* [38], *SOX2* [39], *MYC* [40, 41], *ETV4* and *ETV5* [42] appear. *CDX1* and *CDX2* [43] which are known to modulate early cardiogenesis peak at day 2, alongside some potentially important early cardiac regulators such as *TCF4* and *LEF1*. On day 4, *MYCN* stands out with a high outdegree and indegree confirming its known role in heart development [44] along side with some potential candidate regulators such as *LHX1*, *OTX2*, *NR2C1*, *MIR548Y*. The last stage where the cardiomyocytes have already matured, features core regulators essential for cardiac development such as *MEF2C* [45], *HAND2* [46], *NKX2-5* [47], *MEIS2* [48], *MITF* [49], FOXP1 [50] and some new candidate regulators that could be significant in the cardiac maturation such as *MEF2A* and *BHLHE40*. Like in the previous dataset, a strong same-stage presence of certain TF family members is observed, such as the members of the HOX family CDX1, CDX2, HOXA1 and HOXB1 in day 2.

Discussion

Unlike some of the classic regulatory models such as BNs, the TRC model takes advantage of the sequential order of the time series data to allow more intricate interpretations of regulatory interactions. It takes advantage of the emerging property from the peaking patterns, that is: each node in the cascade is positively correlated in its expression pattern to the other nodes in the same stage (Fig 5A), and correlated via a time-lagged correlation to the nodes in the other stages (Fig 5C). Thus each edge in the cascade is always coupled with a correlation between the expression pattern of the regulator and its target. This coupling can be viewed as a reinforcement of the regulatory interaction predictive quality and gives it an edge over interactions based solely on the binding site analysis or solely derived from gene expression data. From another view, the binding site prediction behind the edge can explain the perceived correlation in the expression patterns between the target and the source. Fig 5 summarizes the five common types of regulatory interactions displayed within the cascade through edge patterns. Fig 5A is an example of a regulatory interaction coupled with high positive correlation indicating that X is potentially one of the activators of Y and contributes to its peaking pattern. Y is inactive where X is inactive and activated when X is activated (stage i), coupled with the fact that X can bind to the promoter of Y, this hypothesis of the regulatory influence of X on Y is strongly enforced. Fig 5A has a one-direction property that supports the causality, whereas cases such as the double edge displayed in Fig 5B cannot decisively assert whether X is an activator of Y or the other way around due to the non-causal nature of correlation and the double potential of these regulators to bind to each other's promoters. Fig 5C is an example of where a regulator in a certain stage potentially needs more time to activate the target thus the target is activated after a time delay and captured in the next stage. This kind of regulatory behavior has been shown and captured using time-lagged correlation models. Another common hypothesis that surrounds co-expressed genes is that they might be coregulated by a master regulator or a set of master regulators. Some of these master regulators can be captured through configurations in the cascade where a regulator emerging in a stage single-handedly has the potential to activate a wide set of correlated regulators, whether in the same stage as in the case of Fig 5D or a set of targets in the next stage via a time-lagged regulation as shown in Fig 5E.

The previous analyses of the two datasets showed clear stage-specific regulatory waves and a GO enrichment that is highly consistent with the biological context of the experiment and, even more specifically, the context in the particular time points of the experiment. The question arises whether these peaking profiles and case-specific GO enrichments are statistically significant, and constitute a characteristic of developmental gene expression datasets in particular, or if they randomly occur in any dataset. While applying the TRC model to a sufficiently

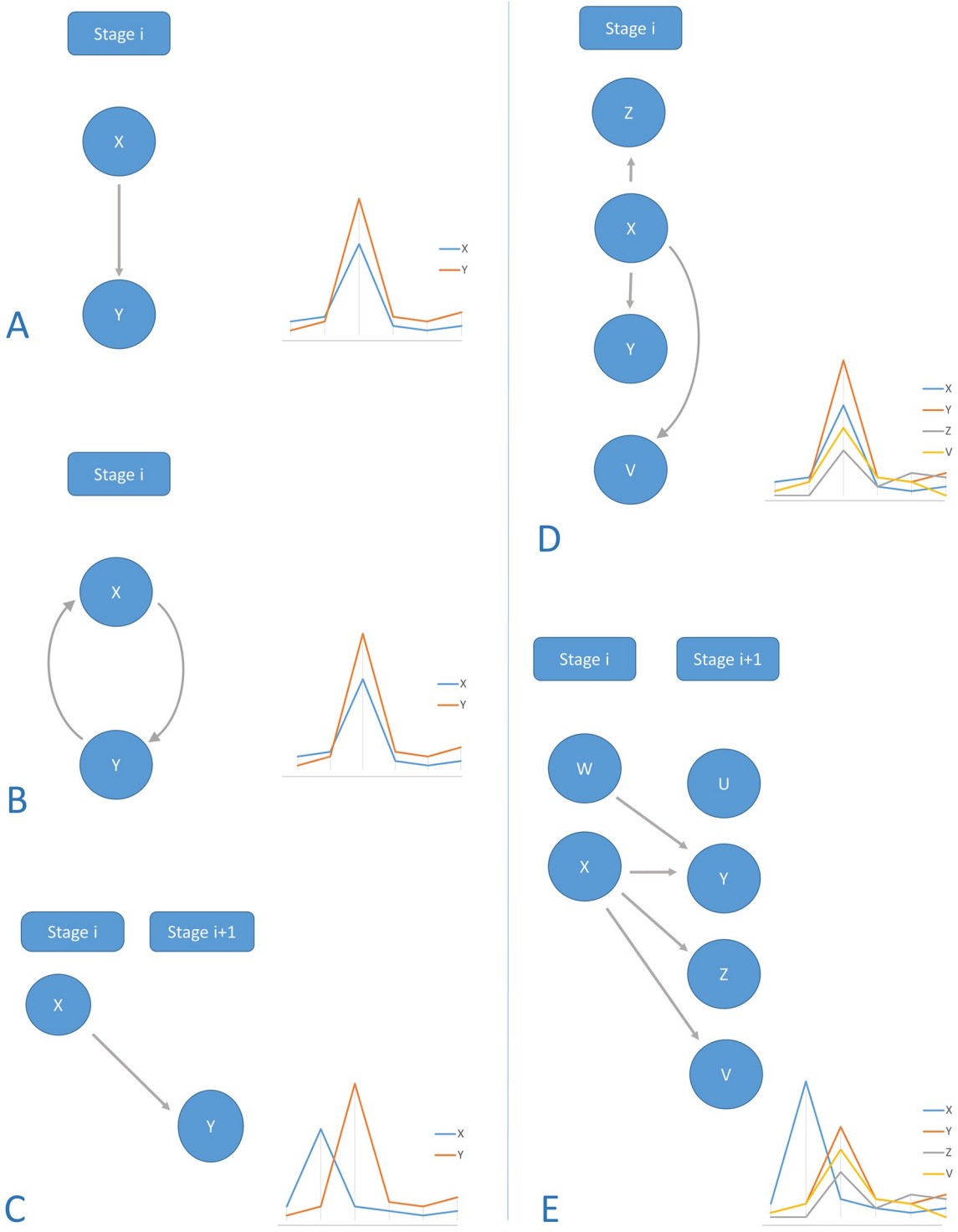

Fig 5. Different cases of regulatory interactions contained in the TRC model. (A) A one-way regulatory prediction within one stage coupled with a high positive correlation. (B) A two-way regulatory prediction within one stage coupled with a high positive correlation. (C) A regulatory interaction from one stage to the next, coupled with a high positive time-lagged correlation. (D) X a potential master regulator of X, Y, and Z coupled with a high positive correlation to each of its targets. (E) X a potential master regulator activating X, Y and Z coupled with a high positive time-lagged correlation to each of its targets.

https://doi.org/10.1371/journal.pone.0231326.g005

large number of random and shuffled datasets and evaluating the resulting TRCs would be optimal to proof the statistical significance of the results, it is merely unfeasible due to the manual process of assessing the resulting TRCs. Alternatively, we applied the model to randomly generated and shuffled gene expression datasets (see the supplementary files) aiming towards a comparative analysis rather than a statistical proof of significance. We examined the resulting TRCs in terms of the GO enrichment of the stages to evaluate their relevance compared to a TRC generated from a real experimental dataset. The first test involved shuffling dataset2 by re-assigning genes to other expression profiles (S3 File), to check whether any set of peaking regulators will show a specific GO enrichment, and none of the stages did lead to any relevant terms. The test was repeated by shuffling the regulator's profiles only, and the enrichment was again insignificant. The previous test showed that the identity of the peaking genes is essential, precise, and specific. Moreover, the workflow was applied to dataset2 without restricting the stage-specific sets to regulators only. Interestingly, the generated cascade was overwhelmed by non-regulatory genes and the GO enrichment showed no significant terms in any of the stages, with the exception of two terms related to cardiac muscle differentiation in the last stage (S1 Fig). This observation supports the choice in the TRC model of limiting the cascade to regulators where less relevant non-regulatory genes do not dilute the small stage-specific gene sets. Next, dataset2 was shuffled by permuting all the values in the expression matrix (S4 File). The result was again a lack of significance in GO the enrichment terms. The last test was applying the TRC workflow to a randomly generated gene expression dataset, using the gene names and the time points from dataset2 combined with randomly generated expression values (S5 File). The GO enrichment showed the absence of any relevant significant terms again.

The default library used in this model is the one-stage peak pattern library, which works optimally with development and differentiation. However this library can be changed, and multiple libraries for different biological contexts such as diseases and immune responses can be developed accordingly, which would require further research or alternatively allowing the user to construct a custom library in the future.

One drawback of this model is the fact that it does not capture every important regulator, particularly those regulators that are expressed in multiple consecutive or non-consecutive time points. However, we argue that the sets of regulators identified by the cascade contain a large percentage of essential stage-specific regulators which is supported by the GO enrichment. On the other hand, the regulatory network might not cover every TF due to missing PWM information or lack of conservation. Another more general drawback is the fact that the model relies on transcript levels which do not translate directly into protein levels, but relative measures [51] [52] can be a potential method for further analysis whenever protein data is not available. Moreover, the candidate regulators can provide a small concise set for a proteomic investigation as a next step in the experiment. The captured regulators can also provide a starting point for further analysis such as target set enrichment analysis, pathway analysis, and investigating the potential collaboration of regulators using tools such as PC-Traff [53]. The TRC model merely lays down, in place, some important starting pieces that can be built on to complete the biological puzzle of developmental regulatory programs.

The unique type of the output of the TRC makes it difficult to accurately compare it to other existing methods, as no other method has the same definition of a regulatory cascade. However, we utilized the context-relevance of the GO enrichment of the gene sets predicted by other methods as a basis for the comparison. We first applied the STEM in order to predict the top 10 significant gene expression patterns in the cardiac differentiation dataset and evaluated the GO enrichment of the genes set associated with each of these profiles. The GO enrichment of these sets showed very general terms not specific to the cardiac differentiation context.

Next, we applied iDREM [54], which we consider the closest method to the TRC in terms of inputs and aims, using the cardiac differentiation dataset and the regulatory network provided by iDREM (human_predicted_1000), to generate a dynamic regulatory network. The resulting model was in the form of a dynamic regulatory map that highlights major bifurcation events, each of which has a list of associated regulatory genes. The GO enrichment of these gene lists showed a mild enrichment of developmental GO terms in some bifurcation points and no enrichment in most of the others. However, proving the validity of a generated network or cascade requires an actual experimental validation of the predicted regulatory interactions in the particular cellular context, which is currently unpractical.

This workflow is built within a broader framework dedicated to studying regulation from different points of view. It blends expression data and a regulatory network and links concepts such as coexpression and coregulation forming a more extensive tool. Users can interactively investigate different hypothesis and track different genes and regulators of interest exploring the regulatory forces governing the time series data, the timing of such forces and the impact of such regulatory interactions on the expression of genes and regulators.

Conclusion

We developed a workflow to analyze and represent regulatory cascades and a web tool based on the corresponding model. It takes time series expression data as an input, generates and visualizes an interactive cascade that identifies relevant and stage-specific regulators associated with each time point and the interactions between these regulators. The workflow was applied to multiple datasets that revolved around cell differentiation and was successful in identifying previously-known TFs relevant to the time points and the cell types, in addition to some new candidate regulators, as well as pinpointing the time points were unique regulation activities are emerging. A demo of the web tool is available under TF-investigator.sybig.de/TRC.

Supporting information

S1 File. NPC differentiation (Dataset1). The formatted data expression file based on human H1-derived NPC differentiation differentiation. This format is ready for upload via the webtool.
(CSV)

S2 File. Cardiac differentiation (Dataset2). The formatted data expression file based on C20 derived cardiomyocyte differentiation. This format is ready for upload via the webtool.
(CSV)

S3 File. Shuffled profile assignment of dataset2. A version of dataset2 where gene profiles are randomly re-assigned. This format is ready for upload via the webtool.
(CSV)

S4 File. Shuffled dataset2 by permuting the matrix. A version of dataset2 where cells in the expression matrix are permuted across columns and rows. This format is ready for upload via the webtool.
(CSV)

S5 File. Random expression values with dataset2 time points and gene names. Random expression values with time points and gene names taken from dataset2. This format is ready for upload via the webtool.
(CSV)

S1 Fig. TRC based on dataset2 where regulatory and non-regulatory genes are included. Stage-specific gene sets are not restricted to regulators in this example. This allows the TRC to include peaking non regulatory genes as well.
(TIF)

Acknowledgments

We would like to thank Sebastian Zeidler and Hryhorii Chereda for their insights and helpful discussions.

Author Contributions

Conceptualization: Rayan Daou, Edgar Wingender, Mehmet Gültas, Martin Haubrock.

Methodology: Rayan Daou.

Software: Rayan Daou.

Supervision: Tim Beißbarth, Edgar Wingender, Martin Haubrock.

Visualization: Rayan Daou.

Writing – original draft: Rayan Daou.

Writing – review & editing: Tim Beißbarth, Edgar Wingender, Mehmet Gültas, Martin Haubrock.

References

1. Bolouri H, Davidson EH. Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. Proceedings of the National Academy of Sciences. 2003; 100 (16):9371–9376. https://doi.org/10.1073/pnas.1533293100
2. Zimmer B, Kuegler P, Baudis B, Genewsky A, Tanavde V, Koh W, et al. Coordinated waves of gene expression during neuronal differentiation of embryonic stem cells as basis for novel approaches to developmental neurotoxicity testing. Cell death and differentiation. 2011; 18(3):383. https://doi.org/10.1038/cdd.2010.109 PMID: 20865013
3. Davidson EH, McClay DR, Hood L. Regulatory gene networks and the properties of the developmental process. Proceedings of the National Academy of Sciences. 2003; 100(4):1475–1480. https://doi.org/10.1073/pnas.0437746100
4. Zuber ME, Gestri G, Viczian AS, Barsacchi G, Harris WA. Specification of the vertebrate eye by a network of eye field transcription factors. Development. 2003; 130(21):5155–5167. https://doi.org/10.1242/dev.00723 PMID: 12944429
5. Mootoosamy RC, Dietrich S. Distinct regulatory cascades for head and trunk myogenesis. Development. 2002; 129(3):573–583. PMID: 11830559
6. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell. 2006; 126(4):663–676. https://doi.org/10.1016/j.cell.2006.07.024 PMID: 16904174
7. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. cell. 2007; 131(5):861–872. https://doi.org/10.1016/j.cell.2007.11.019 PMID: 18035408
8. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced pluripotent stem cell lines derived from human somatic cells. science. 2007; 318(5858):1917–1920. https://doi.org/10.1126/science.1151526 PMID: 18029452
9. Csete M. Translational prospects for human induced pluripotent stem cells. Regenerative medicine. 2010; 5(4):509–519. https://doi.org/10.2217/rme.10.39 PMID: 20632855
10. Lindvall O, Kokaia Z. Stem cells in human neurodegenerative disorders—time for clinical translation? The Journal of clinical investigation. 2010; 120(1):29–40. https://doi.org/10.1172/JCI40543 PMID: 20051634

11. Lindvall O, Kokaia Z, Martinez-Serrano A. Stem cell therapy for human neurodegenerative disorders–how to make it work. Nature medicine. 2004; 10(7s):S42. https://doi.org/10.1038/nm1064 PMID: 15272269

12. Inoue H, Nagata N, Kurokawa H, Yamanaka S. iPS cells: a game changer for future medicine. The EMBO journal. 2014; 33(5):409–417. https://doi.org/10.1002/embj.201387098 PMID: 24500035

13. Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature biotechnology. 2009; 27(3):275. https://doi.org/10.1038/nbt.1529 PMID: 19252484

14. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9(1):559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008

15. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics. 2006; 7(1):191. https://doi.org/10.1186/1471-2105-7-191 PMID: 16597342

16. Gonçalves JP, Francisco AP, Mira NP, Teixeira MC, Sá-Correia I, Oliveira AL, et al. TFRank: network-based prioritization of regulatory associations underlying transcriptional responses. Bioinformatics. 2011; 27(22):3149–3157. https://doi.org/10.1093/bioinformatics/btr546 PMID: 21965816

17. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y. A guide to web tools to prioritize candidate genes. Briefings in bioinformatics. 2010; 12(1):22–32. https://doi.org/10.1093/bib/bbq007 PMID: 21278374

18. Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D. How to infer gene networks from expression profiles. Molecular systems biology. 2007; 3(1):78. https://doi.org/10.1038/msb4100120 PMID: 17299415

19. Ristevski B. A survey of models for inference of gene regulatory networks. Nonlinear Anal Model Control. 2013; 18(4):444–65. https://doi.org/10.15388/NA.18.4.13972

20. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. PloS one. 2012; 7(1):e29348. https://doi.org/10.1371/journal.pone.0029348 PMID: 22272232

21. Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Briefings in bioinformatics. 2008; 9(4):326–332. https://doi.org/10.1093/bib/bbn016 PMID: 18436575

22. MATCH: A tool for searching transcription factor binding sites in DNA sequences., author = Kel, A E and Gössling, E and Reuter, I and Cheremushkin, E and Kel-Margoulis, O V and Wingender, E, journal = Nucleic acids research, journal-iso = Nucleic Acids Res., volume = 31, number = 13, year = 2003, month = Jul, pages = 3576-9, pmid = 12824369.

23. Haubrock M, Li J, Wingender E. Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks. In: BMC systems biology. vol. 6. BioMed Central; 2012. p. S15. https://doi.org/10.1186/1752-0509-6-S2-S15

24. Göke J, Jung M, Behrens S, Chavez L, O'Keeffe S, Timmermann B, et al. Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. PLoS computational biology. 2011; 7(12):e1002304. https://doi.org/10.1371/journal.pcbi.1002304 PMID: 22215994

25. Kiełbasa SM, Vingron M. Transcriptional autoregulatory loops are highly conserved in vertebrate evolution. PLoS One. 2008; 3(9):e3210. https://doi.org/10.1371/journal.pone.0003210 PMID: 18791639

26. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS biology. 2004; 3(1):e7. https://doi.org/10.1371/journal.pbio.0030007 PMID: 15630479

27. Cripps RM, Olson EN. Control of cardiac development by an evolutionarily conserved transcriptional network. Developmental biology. 2002; 246(1):14–28. https://doi.org/10.1006/dbio.2002.0666 PMID: 12027431

28. Liu Q, Jiang C, Xu J, Zhao MT, Bortle KV, Cheng X, et al. Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. Circulation research. 2017; 121(4):376–391. https://doi.org/10.1161/CIRCRESAHA.116.310456 PMID: 28663367

29. Lu QR, Sun T, Zhu Z, Ma N, Garcia M, Stiles CD, et al. Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection. Cell. 2002; 109(1):75–86. https://doi.org/10.1016/s0092-8674(02)00678-5 PMID: 11955448

30. Liu Y, Jiang P, Deng W. OLIG gene targeting in human pluripotent stem cells for motor neuron and oligodendrocyte differentiation. Nature protocols. 2011; 6(5):640. https://doi.org/10.1038/nprot.2011.310 PMID: 21527921

31. Liu Y, Rao MS. Olig genes are expressed in a heterogeneous population of precursor cells in the developing spinal cord. Glia. 2004; 45(1):67–74. https://doi.org/10.1002/glia.10303 PMID: 14648547

32. Mansouri A, Hallonet M, Gruss P. Pax genes and their roles in cell differentiation and development. Current opinion in cell biology. 1996; 8(6):851–857. https://doi.org/10.1016/s0955-0674(96)80087-1 PMID: 8939674

33. Burrill JD, Moran L, Goulding MD, Saueressig H. PAX2 is expressed in multiple spinal cord interneurons, including a population of EN1+ interneurons that require PAX6 for their development. Development. 1997; 124(22):4493–4503. PMID: 9409667

34. Soukkarieh C, Agius E, Soula C, Cochard P. Pax2 regulates neuronal–glial cell fate choice in the embryonic optic nerve. Developmental biology. 2007; 303(2):800–813. https://doi.org/10.1016/j.ydbio.2006.11.016 PMID: 17173889

35. Wingender E, Schoeps T, Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. Nucleic acids research. 2012; 41(D1):D165–D170. https://doi.org/10.1093/nar/gks1123 PMID: 23180794

36. Wingender E, Schoeps T, Haubrock M, Dönitz J. TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic acids research. 2014; 43(D1):D97–D102. https://doi.org/10.1093/nar/gku1064 PMID: 25361979

37. Mendjan S, Mascetti VL, Ortmann D, Ortiz M, Karjosukarso DW, Ng Y, et al. NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency. Cell stem cell. 2014; 15(3):310–325. https://doi.org/10.1016/j.stem.2014.06.006 PMID: 25042702

38. Gao F, Kwon SW, Zhao Y, Jin Y. PARP1 poly (ADP-ribosyl) ates Sox2 to control Sox2 protein levels and FGF4 expression during embryonic stem cell differentiation. Journal of Biological Chemistry. 2009; 284(33):22263–22273. https://doi.org/10.1074/jbc.M109.033118 PMID: 19531481

39. Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nature cell biology. 2007; 9 (6):625. https://doi.org/10.1038/ncb1589 PMID: 17515932

40. Murphy MJ, Wilson A, Trumpp A. More than just proliferation: Myc function in stem cells. Trends in cell biology. 2005; 15(3):128–137. https://doi.org/10.1016/j.tcb.2005.01.008 PMID: 15752976

41. Smith KN, Singh AM, Dalton S. Myc represses primitive endoderm differentiation in pluripotent stem cells. Cell stem cell. 2010; 7(3):343–354. https://doi.org/10.1016/j.stem.2010.06.023 PMID: 20804970

42. Akagi T, Kuure S, Uranishi K, Koide H, Costantini F, Yokota T. ETS-related transcription factors ETV4 and ETV5 are involved in proliferation and induction of differentiation-associated genes in embryonic stem (ES) cells. Journal of Biological Chemistry. 2015; 290(37):22460–22473. https://doi.org/10.1074/jbc.M115.675595 PMID: 26224636

43. Lengerke C, Wingert R, Beeretz M, Grauer M, Schmidt AG, Konantz M, et al. Interactions between Cdx genes and retinoic acid modulate early cardiogenesis. Developmental biology. 2011; 354(1):134–142. https://doi.org/10.1016/j.ydbio.2011.03.027 PMID: 21466798

44. Harmelink C, Peng Y, DeBenedittis P, Chen H, Shou W, Jiao K. Myocardial Mycn is essential for mouse ventricular wall morphogenesis. Developmental biology. 2013; 373(1):53–63. https://doi.org/10.1016/j.ydbio.2012.10.005 PMID: 23063798

45. Lin Q, Schwarz J, Bucana C, Olson EN. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. Science. 1997; 276(5317):1404–1407. https://doi.org/10.1126/science.276.5317.1404 PMID: 9162005

46. Morikawa Y, Cserjesi P. Cardiac neural crest expression of Hand2 regulates outflow and second heart field development. Circulation research. 2008; 103(12):1422–1429. https://doi.org/10.1161/CIRCRESAHA.108.180083 PMID: 19008477

47. Hiroi Y, Kudoh S, Monzen K, Ikeda Y, Yazaki Y, Nagai R, et al. Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. Nature genetics. 2001; 28(3):276. https://doi.org/10.1038/90123 PMID: 11431700

48. Machon O, Masek J, Machonova O, Krauss S, Kozmik Z. Meis2 is essential for cranial and cardiac neural crest development. BMC developmental biology. 2015; 15(1):40. https://doi.org/10.1186/s12861-015-0093-6 PMID: 26545946

49. Tshori S, Gilon D, Beeri R, Nechushtan H, Kaluzhny D, Pikarsky E, et al. Transcription factor MITF regulates cardiac growth and hypertrophy. The Journal of clinical investigation. 2006; 116(10):2673–2681. https://doi.org/10.1172/JCI27643 PMID: 16998588

50. Wang B, Weidenfeld J, Lu MM, Maika S, Kuziel WA, Morrisey EE, et al. Foxp1 regulates cardiac outflow tract, endocardial cushion morphogenesis and myocyte proliferation and maturation. Development. 2004; 131(18):4477–4487. https://doi.org/10.1242/dev.01287 PMID: 15342473

51. Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Molecular systems biology. 2016; 12(10):883. https://doi.org/10.15252/msb.20167144 PMID: 27951527

52. Wang X, Teng L, Zhang H, Zhou Q, Su X, Cui X, et al. Bi-clustering interpretation and prediction of correlation between gene expression and protein abundance. bioRxiv. 2018; p. 270397.

53. Meckbach C, Tacke R, Hua X, Waack S, Wingender E, Gültas M PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. BMC bioinformatics. 2015; 16 (1):400 https://doi.org/10.1186/s12859-015-0827-2 PMID: 26627005

54. Ding J, Hagood JS, Ambalavanan N, Kaminski N, Bar-Joseph Z: Interactive visualization of dynamic regulatory networks. PLoS computational biology. 2018 Mar 14; 14(3):e1006019. https://doi.org/10.1371/journal.pcbi.1006019 PMID: 29538379