Modelling Gene Expression in Terms of DNA Sequence

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

ĐORĐE RELIĆ

Basel, 2023
Originaldokument gespeichert auf dem Dokumentenserver
der Universität Basel edoc.unibas.ch

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Erik van Nimwegen (First Supervisor)

Prof. Dr. Alexander F. Schier (Second Supervisor)

Associate Prof. Dr. Nadine Vastenhouw (External Expert)

Basel, den 14. Dezember 2021

Prof. Dr. Marcel Mayor

Dekan

Abstract

Understanding the gene regulatory networks that control gene expression remains one of the most of important questions in molecular biology. Much of gene expression is controlled through transcription initiation, whose regulation is ultimately encoded in the constellations of small sequence motifs in the DNA that are bound by transcription factors (TFs) in a sequence-specific manner. In this thesis, we addressed the task of understanding gene regulation on two levels. Firstly, we present a computational pipeline for inferring a set of gene regulatory elements in a given organism which includes identifying genes that encode DNA-binding domains (DBDs), mapping them to known binding motifs by leveraging similarity in DBDs between species, annotating promoter regions genome-wide, aligning promoters with orthologous regions from related genomes, and predicting genome-wide transcription factor binding sites (TFBSs). We demonstrated the use of our pipeline by applying it to zebrafish. Furthermore, we integrated these results into our previously developed Integrated System for Motif Activity Response Analysis (ISMARA) which models gene expression data in terms of predicted regulatory sites. Using ISMARA, we predicted known and novel key regulatory TFs in zebrafish using a number of RNA-seq datasets. Secondly, we zoom in at the scale of one single TF regulating a set of constitutive promoters in *Escherichia coli*. We analyzed an artificially evolved set of synthetic promoter sequences which are selected for expression constitutive promoters regulated by σ^{70} transcription factor. We looked closely into promoter sequences and TF binding dynamics and investigated the predictive power of TF binding affinity on gene expression.

Acknowledgments

"Because it's there." - George Mallory

I have not climbed Everest, nor am I a mountaineer for the matter, but I like this quote from Mallory because it symbolizes pursuing that inner drive that is sometimes just too hard to explain. Looking back over the past four years, it was the drive of curiosity that led me to the decision to pursue a PhD degree at Biozentrum and it was one of the best decisions of my life. However, starting something is much easier than finishing and I would not be here without the support from many people who made this journey so special.

First and foremost, I would like to thank my supervisors Erik van Nimwegen and Alex Schier. Thank you for giving me the opportunity to learn from you about the beauty of science spanning from applied mathematics to molecular biology and all the ways they interplay. Thank you for your support, guidance, and patience throughout the last four years. Finally, thank you for fueling my scientific curiosity from the first time we met and for the opportunity to contribute to your scientific endeavors.

I would like to thank Nadine Vastenhouw for reviewing my work and accepting to be part of my PhD committee. I would also like to thank Mikhail Pachkov for his great help on my projects. Many thanks go to Yvonne Steger, Sarah Güthe, and Angie Klarer for their priceless administrative and personal support throughout the last years.

Furthermore, I would like to thank all past and present Biozentrists I had the pleasure to meet. Biozentrum is a place filled with interesting people and it was a pleasure to spend the last four years here. Thanks to the present and past members of van Nimwegen lab for being such good colleagues and friends, and for making this whole journey so enjoyable inside and outside of the lab. Special thanks go to Christina, Enea, and Athos for sharing the highs and the lows over these years. It wouldn't have been the same without you.

Especially, I would like to thank my girlfriend Lara. For all the understanding and support, particularly in the last stretch of this journey. Thank you for enriching this whole experience and my life by adding a wonderful dimension to it.

Last but not least, I want to thank my parents, Slavka and Živko, and my sister Isidora. Thank you for your continuous and unconditional trust, encouragement, patience, and love. Thank you for always being there for me whenever I needed you. I dedicate this thesis to you.

Table of Contents

List of Figures

List of Tables

1
Introduction

"Progress in science depends on new techniques, new discoveries and new ideas, probably in that order."

Sydney Brenner

In a world as diverse as ours, curious individuals who could afford the time always wondered about understanding what makes the world around us. What makes an organism? What are the underlying reasons which lead to such staggering differences between organisms? How do different organisms interact and why? How do they use resources around them, such as water, minerals, sunlight, or even whole other organisms, to power life, development, and change over time? In biology, there is an infinity of questions that spark curiosity but, by continuously learning more about the underlying processes in living organisms, one can also make a real-world impact.

The world we live in is made of an amazingly large number of different organisms. Differing in many ways, while at the same time sharing many similarities. In 1859, Charles Darwin published "The origin of species"[42] which set foundations for evolutionary biology. Even though controversial at the time, today it is widely accepted that we are all coming from the same ancestor. It implied that there has to be some kind of hereditary transformable information passed on from generation to generation. In 1866, Gregor Mendel showed how certain traits, such as color and size of the fruit, are passed down to generations in pea plants [109]. In 1902, Sir Archibald Edward Garrod showed that traits related to diseases are passed down to human generations [55]. Not so long after the work of Mendel, in 1869 Friedrich Miescher identified "nu-

clein" which is what we call today the DNA[41]. And in 1944, Oswald Avery showed that it is exactly this molecule that is changing over generations[11]. Back then, as today, the scientific world was divided on what are the correct answers to the questions we are asking. These debates fueled more research innovations. Discoveries were following discoveries, often happening in parallel, and, scientist due credit, often had to wait before being properly acknowledged. In 1952 we had the first X-ray photograph of DNA made by Rosalind Franklin, which was followed by the discovery of the double-helix structure of DNA in 1953 published by James Watson and Francis Crick [160]. With the discovery of the physical shape of DNA, in 1951 Barbara McClintock published her work on the interaction of two genetic loci in maize, and its role in seed color formation [107]. Ten years later, in 1961, François Jakob and Jaques Monod, publish their important work on gene regulation in bacteria which explained the mechanism of differential gene regulation depending on the type of nutrient bacteria are consuming [74]. Around the same time, in 1961, we learned from the work of Marshall Nirenberg how information encoded in DNA is translated to proteins [116]. And by 1970, scientists could decipher which combinations of nucleotides make proteins. Few years after, from the work by Wu and Padmanabhan et al. we learned how uncover the composition of a short piece of DNA sequence [75, 120, 126] and in 1977, through the work of Frederick Sanger, we got our first fully sequenced genome of a phage [131]. This discovery led to a slow but steady increase in genome sequencing and in 2003 we had the whole human genome sequenced [38].

With this brief part of the history of genetic regulation, we went from discovering that hereditary information is passed on through generations, to what exactly it is and how it looks like, how it determines certain traits we exhibit, and finally, to how this information is stored and extracted for appropriate function. The burning question of how traits are formed, what influences their expression, and where it happens has been more elucidated with every new discovery made. What we know now is that genes are specific sequences of nucleic acids that are initially transcribed into mRNA molecules and subsequently translated into proteins which give cells their function. We also know that there are special proteins called transcription factors (TFs). They are present in all living organisms and their general role is conserved - they bind DNA and regulate gene expression.

Having the idea of gene regulation settled as a concept, questions that beg for answers are: how exactly do TFs regulate gene expression? Can we identify all TFs in a given organism, predict their DNA binding sites, and describe the gene regulatory network they form? To which extent can we control gene expression? And, ultimately, can we perturb a gene regulatory network and (re)program cells?

1.1 Gene expression

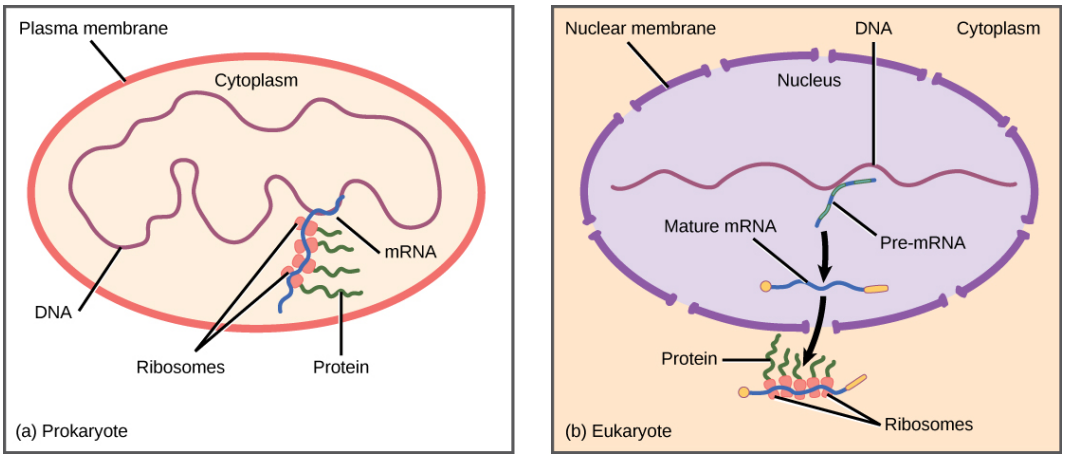


Figure 1.1: **Simplified sketch of gene expression in prokaryotes and eukaryotes.**

Despite large differences in complexity between bacteria, plants, and animals, the core gene regulation concept is mostly conserved. Essentially, the main difference comes from the organization of the cell. Bacteria, which are prokaryotes, do not have a nucleus and DNA is floating freely in the cytoplasm. On the other hand, animals, plants, and fungi, which are eukaryotes, have a nucleus that contains the DNA material[12]. In Figure 1.1[1] we see a sketch showcasing the difference between prokaryotic and eukaryotic gene regulation. Since prokaryotes do not have a nucleus separating DNA and ribosomes from each other, transcription and translation can happen simultaneously. Therefore, there is virtually no post-transcriptional control. Hence, the majority of gene expression regulation in bacteria happens on the gene transcription level. On the other hand, in eukaryotes, gene transcription takes place in the nucleus,

[1] Figure provided by OpenStax under Creative Commons Attribution License 4.0. Access for free at https://openstax.org/books/biology/pages/1-introduction

where pre-mRNA is synthesized and then matured by several RNA processing steps such as capping, splicing, and polyadenylation. Then, mature mRNAs are transported out of the nucleus to the cytoplasm where ribosomes start the process of mRNA translation[5].

For transcription initiation to happen, RNA polymerase has to bind DNA in the promoter region. The double-stranded DNA is opened by "melting" the hydrogen bonds between complementary DNA nucleotides and the RNAP holoenzyme "slides" down the template strand, in 3' to 5' direction, and synthesizes the complementary RNA matching the 5' to 3' direction on the coding strand of the DNA. Once RNAP reaches the "terminator" sequence, transcription stops, mRNA is released from the template DNA, RNAP holoenzyme detaches from DNA, and hydrogen bonds between complementary nucleotides on the DNA are restored [5].

1.2 Measuring gene expression

To understand the underlying complexity of a cell, we need to measure its expression state. Ideally, we would always like to measure gene expression in protein levels since, only when folded, proteins can perform a designated function in the cell. However, such a task is quite hard if we want to capture the complete gene expression state of the cell. On the other hand, we can measure gene transcript content and, assuming the central dogma of molecular biology[40], and capture the true gene expression state of a cell.

Over time, several techniques have been developed to allow for more precise measurement of gene expression such as Northern blotting [6], serial analysis of gene expression (SAGE) [104], DNA microarray [133], RNA-seq [158], measuring transcriptional activity or protein level with fluorescent reporters [146] to name a few. In this thesis, investigation of gene regulation using is addressed with computational analysis data coming from two types of experiments: Fluorescence-Activated Cell Sorting (FACS)[19] - where we measure the expression of fluorescence intensity of translated proteins, and RNA sequencing [158] - where we measure the amount of mature mRNA available for translation by ribosomes.

1.2.1 Fluorescence-Activated Cell Sorting (FACS)

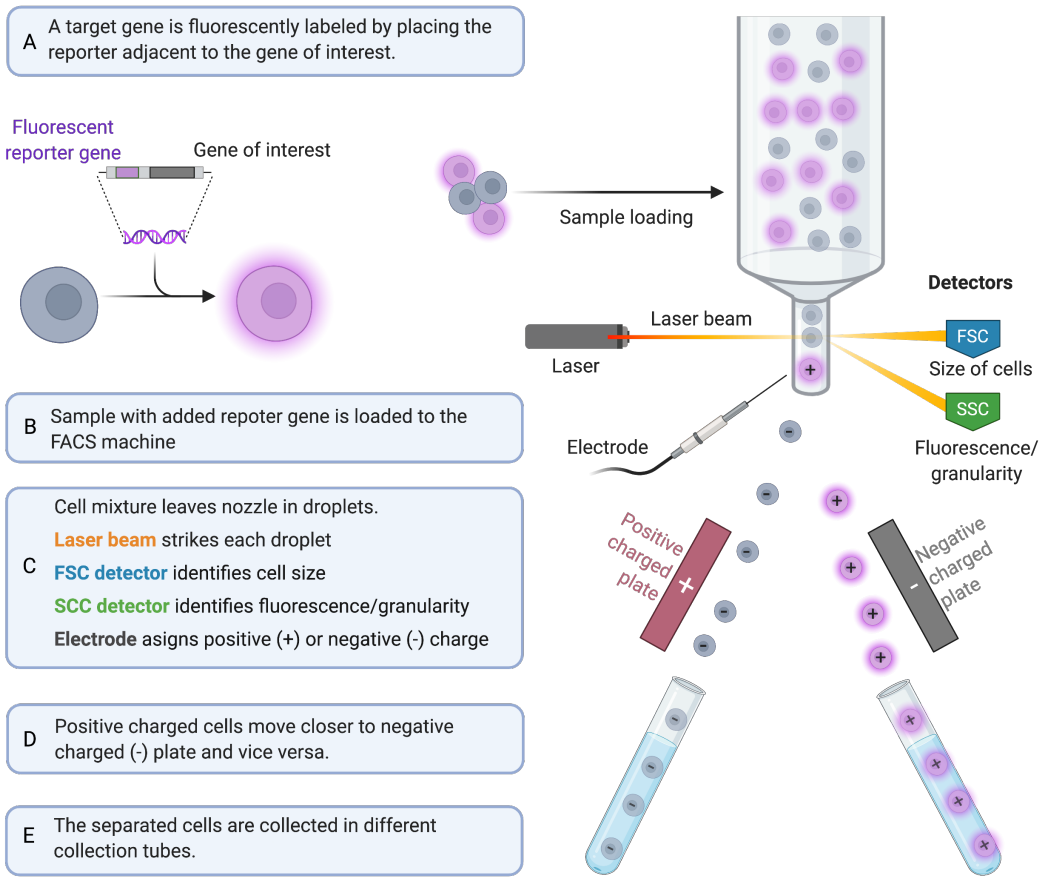


Figure 1.2: **Fluorescence Activated Cell Sorting (FACS).** (A) To label a gene of interest, we place a fluorescent gene reporter adjacent to the gene of interest. The reporter gene is transcribed together with the gene of interest and, once translated and folded, it emits fluorescent light under UV light. (B) Once all cells are prepared, they are loaded into the FACS machine where they go through a nozzle in droplets. (C) When in the nozzle, cells are exposed to a laser beam which allows for measuring of cell size (FSC) by detecting the scattering of light, and the fluorescence intensity (SSC) which is a readout from light emitted from the cell. Based on the readouts, electrodes assign a positive or negative charge. (D) Positively and negatively charged cells are attracted to negative and positive detectors, respectively. (E) Separated cells fall into different collection containers. From these containers, they can be used for further analysis such as sequencing for examples.

The first fluorescent protein, green fluorescent protein (GFP), was discovered in jellyfish and it was quickly shown that it can be incorporated into genomes of different organisms[19, 146]. This enabled the measurement of gene expression on protein level as well as transcriptional activity on promoter level and facilitated the development of different methods for high throughput gene expression measurements on protein level. One such method is Fluorescence-Activated Cell Sorting (FACS).

Figure 1.2 showcases the flow of FACS where we measure protein levels of gene of interest. To measure gene expression on protein level, a so-called reporter gene is inserted adjacent to the gene of interest. This reporter gene encodes for a fluorescent gene, such as GFP. Thus, the reporter gene is expressed together with the gene of choice and can be detected under UV light upon folding. A higher concentration of detected reporter gene implies a higher expression of the gene of choice.

Upon sample injection into the flow cytometer, the cells are focused into a single file with help of a liquid stream (sheath fluid). Droplets are generated having each droplet containing only one cell. A laser beam exciting the fluorescent proteins of every single cell enables the measurement of fluorescence intensity. Based on the intensity, an electrode then charges cells either positively or negatively. After that, cells go through a detector that sorts them based on the assigned charge. Another piece of information obtained from FACS is the scatter of diffracted light coming from the laser hitting the cell. This additional piece of information about cell size allows the implementation of further selection criteria.

FACS experiments are mainly used for selecting cells that express a gene of interest. As we show in Chapter 3, from FACS sorting, we can design experiments that implement artificial selection based on expression levels of the gene of choice.

1.2.2 RNA sequencing (RNA-seq)

Unlike FACS, RNA sequencing (RNA-seq) allows for quantitative transcriptional profiling by measuring the mRNA content of all expressed genes in a sample at a given moment[156].

Figure 1.3[2] showcases the flow of RNA sequencing. The flow can be divided into two parts, library preparation, and library sequencing. For library preparation, the cells of interest are dissociated and only the RNA content of the cells is extracted and purified. These RNAs are fragmented into smaller pieces that are compatible with short-read sequencers. They are then reverse transcribed by reverse transcriptases to generate a single-stranded complemen-

[2] Adapted from [156]

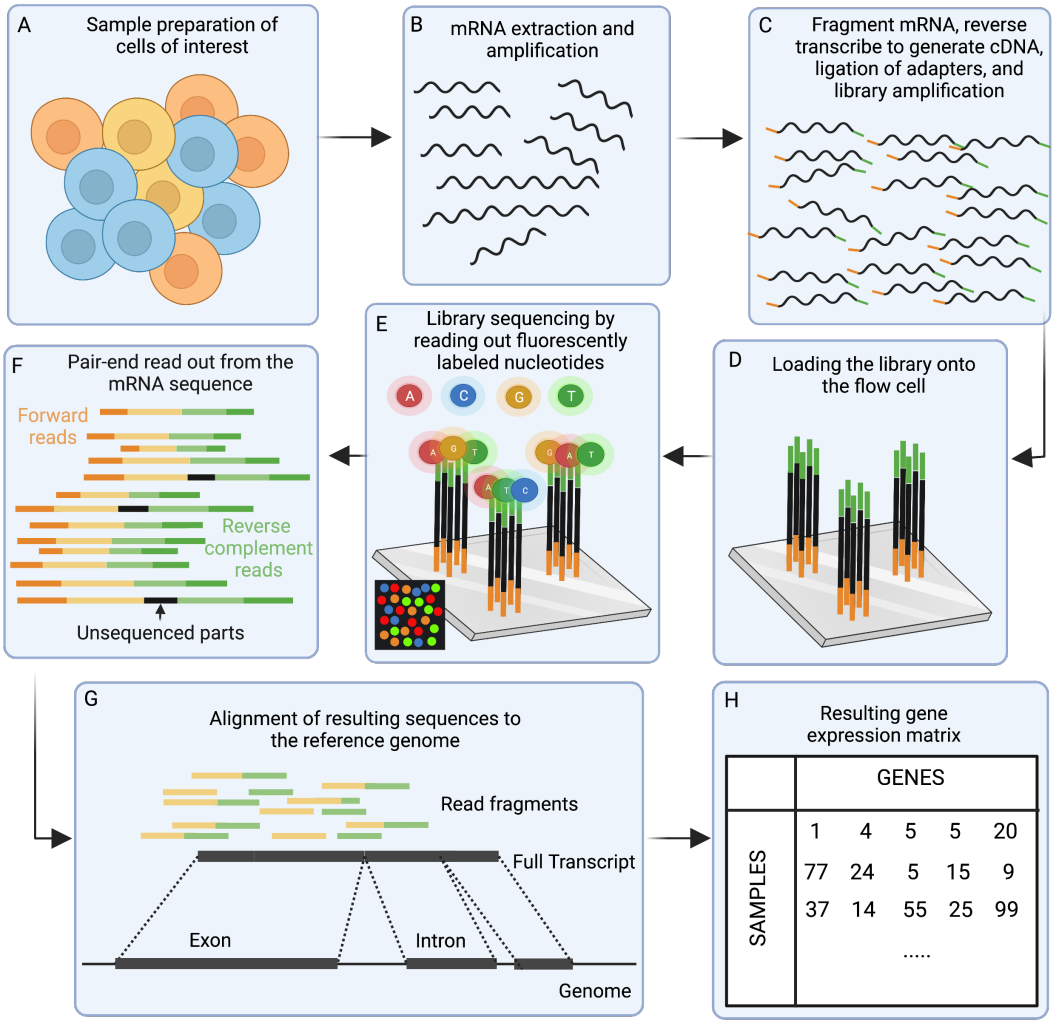


Figure 1.3: **Overview of simplified RNA-sequencing protocol**. (A-C) cDNA library generation, (D-F) library sequencing (Illumina sequence by synthesis), (G-H) Data analysis.

tary DNA (cDNA), followed by a DNA polymerase-mediated complementation to the double-stranded cDNA. Next, adapters are ligated to each cDNA sequence, to enable labeling and amplification of the fragments, as well as sequencing of multiple samples simultaneously. A common technology for cDNA sequencing is Illumina short-read sequencing. For that purpose, the prepared library of single-stranded cDNAs is loaded onto a flow cell where the cDNAs attach to complementary sequences of their adapters. The cDNAs are amplified and so-called "DNA clusters" of each cDNA are formed. Each cluster is then sequenced in several cycles by generating complementary strands. Each cycle consists of adding primers and fluorescently labeled nucleotides, reading the label of the incorporated nucleotide of the complementary strand (the la-

bel ensures incorporation of only one nucleotide at the time), washing excess nucleotides away, and cleaving the label of the incorporated nucleotide to allow the incorporation of the next nucleotide in the next cycle. By reading the labels we can determine which bases were incorporated and ultimately elucidate the sequence of the cDNAs. The generation of complementary DNA is done for both, the forward and the reverse strand. Each cluster (containing one type of cDNA) will generate around 1000 copies of its DNA. The fragments of the DNAs can then be pieced together with help of overlapping areas. This allows for the sequencing of the whole RNA even though it was fragmented beforehand. The sequenced reads can then be aligned to the reference genome ultimately resulting in an expression table that summarizes gene transcript counts across the sample[156].

Using this technique, we can read gene expression across a large number of cells. Moreover, with recent advances in experimental design, we can even measure the RNA content in a single cell. This kind of data allows us to ask all sorts of different questions and the scientific community is hard at work finding ways to analyze such kinds of datasets [70].

1.3 Transcription factors (TFs)

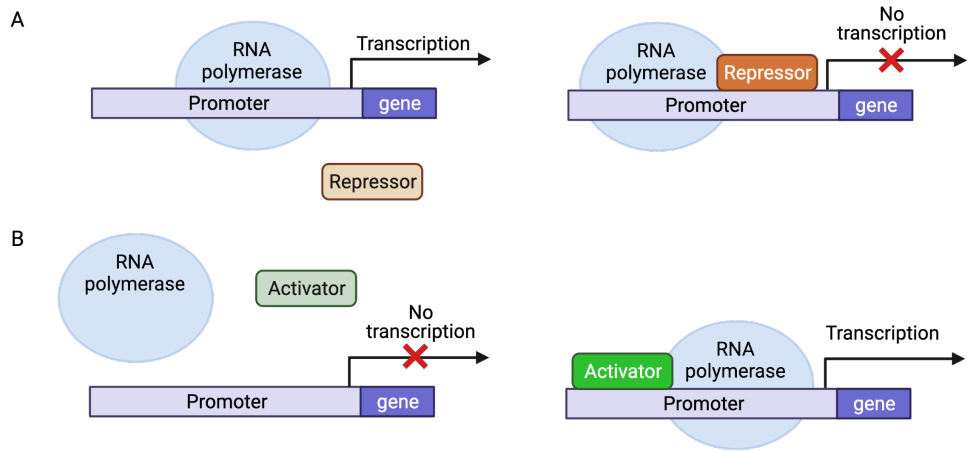


Figure 1.4: **Transcription factors (TFs)**. Transcription factors, in the simplest sense, control transcription in one of two ways. (A) TFs block expression by preventing the RNA polymerase (RNAP) from starting transcription. (B) TFs promote expression by helping the RNAP bind DNA and initiate transcription.

Transcription factors (TFs) are special proteins that control transcription ini-

tiation by directly or indirectly binding DNA [69]. They are mostly either activators or repressors (Figure 1.4), although, in some cases depending on the expression state of the cell, they can be both [114]. An activator promotes expression by helping the RNA polymerase (RNAP) bind DNA and initiate transcription. Sometimes, they act alone [73, 100] and sometimes they act in groups with co-factors [28, 141, 145]. Repressors, on the other hand, block transcription by placing themselves on the DNA and like that, block the passage to the RNAP transcription complex. Notably, TFs can also regulate expression without recruiting or blocking the RNAP transcription complex by changing the chromatin structure. They can either open the chromatin and make it accessible to other TFs [169] or close it and repress gene expression by making the genes inaccessible for transcription[59].

1.3.1 DNA-binding domains (DBDs)

Transcription factors, like all other proteins, are large molecules consisting of one or more protein domains. If we would cut the protein sequence in pieces, the smallest pieces of amino acid sequence that fold into well-defined structures by themselves are called a protein domains. Proteins typically consists of multiple non-overlapping protein domains. Protein domains are classified into classes and one such class of protein domains is called the DNA-binding domain (DBD). Since TFs bind DNA, they are essentially proteins that contain one or more DBDs. Moreover, DBDs encoded in TFs allow for recognition and binding to specific DNA sequences.

1.3.2 Pfam

An exhaustive database of all protein domains is Pfam [111]. At Pfam, protein domains are grouped into protein domain families, based on domain sequence similarity. Protein domain families are manually curated and a similarity threshold is set for each family by hand. The Pfam database is being continuously updated and in the latest release, there is a total of 19,179 domain families. Hundreds of those are DNA-binding specific domains that can be found in prokaryotes or eukaryotes. Pfam is closely integrated with UniProt [39], a database that collects all discovered protein sequences, and in the latest release, there is a total of 47 million sequences. Out of all sequences in the

UniProt database, 74.5% of them contain at least one Pfam domain, and 48.8% of all residues in collected proteins, belong to a Pfam protein domain family. That is, for three-quarters of all proteins, we have some statistical models of domains in those proteins and some information about what their role is. On the other hand, if we look at all residues (parts of protein sequences), for more than half we do not have any characterization [111]. However, in recent years it was shown that many proteins are intrinsically disordered and do not have stable structures [155] (as a feature not as a bug) and it is not so sure we will reduce the number of protein residues which do not have statistical models for their structure.

1.3.3 Hiden Markov Models (HMMs)

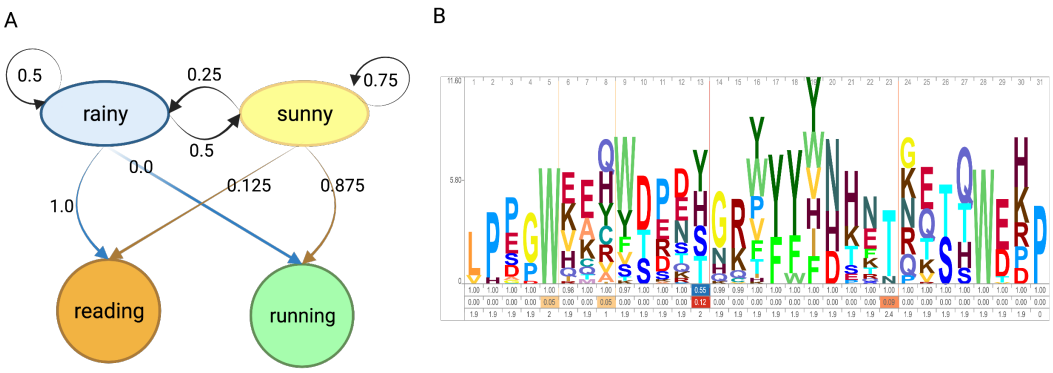


Figure 1.5: **Hidden Markov Model (HMM).** (A) A HMM toy example. (B) Logo representation coming from the HMM of WW domain.

For each domain family, at Pfam, we can retrieve a protein domain model, in the form of a Hiden Markov Model (HMM). The HMM is a statistical generative model which describes a non-observed state in regards to the set of previously observed states [48]. Probably the easiest way to explain introduce the concept of HMMs is through an example.

Let us picture the following scenario. Marvin and Trillian grew up together. At some point, both of them went to college, but in different parts of the country. Nonetheless, they keep in touch by talking every day on the phone. Trillian knows that the weather in Marvin's town is pretty stable during the day, but also, by the laws of nature, tomorrow's weather always depends on today's weather. Trillian also knows that Marvin has two favorite leisure activities which are running and reading his favorite book[71]. He would do only

one of those each day and tell Trillian about it, after the usual weather-related small-talk. One day, Marvin, knowing Trillian collected this data, challenged her to guess the weather that day knowing he was reading in the afternoon.

Now, since Trillian was noting data over time (just for the fun of it), she can estimate transition probabilities between different weathers and Marvin's activities across different days. From her notes, she generated a graph with transition probabilities (Figure 1.5 A). To take Marvin on his challenge, she formalized her notes into two matrices. Transition matrix:

$$\mathbf{T} = \begin{array}{c} \begin{matrix} \text{rainy} & \text{sunny} \end{matrix} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix} \end{array} \begin{array}{l} \text{rainy} \\ \text{sunny} \end{array} \tag{1.1}$$

and emission matrix E:

$$\mathbf{E} = \begin{array}{c} \begin{matrix} \text{rainy} & \text{sunny} \end{matrix} \\ \begin{bmatrix} 0.0 & 0.875 \\ 1.0 & 0.125 \end{bmatrix} \end{array} \begin{array}{l} \text{running} \\ \text{reading} \end{array} \tag{1.2}$$

What Trillian needs to do now is to find which weather (the hidden variable) has the highest probability given that she knows Marvin was reading (observed variable). To calculate which weather condition is more probable, she needs to compute two conditional probabilities. The probability that it was raining and Marvin was reading given it was raining: $(P(\text{rainy}) * P(\text{rainy}|\text{reading}))$ and the probability that it was sunny and Marvin was reading given it was sunny: $(P(\text{sunny}) * P(\text{sunny}|\text{reading}))$. Since she does not know what is the probability of a given weather, she needs to compute the stationary probability of the Markov Chain defined by the transitions between the weather states. To do that, she needs to calculate the normalized left Eigenvector of the transition matrix T by solving the following system of equations:

$$\begin{aligned} &\begin{bmatrix} P(\text{rainy}) & P(\text{sunny}) \end{bmatrix} T = \begin{bmatrix} P(\text{rainy}) & P(\text{sunny}) \end{bmatrix} \\ &P(\text{rainy}) + P(\text{sunny}) = 1 \end{aligned} \tag{1.3}$$

This results in the stationary probability of $P_{\text{rainy}} = 0.66$ and $P_{\text{sunny}} = 0.33$. Now, Trillian concludes that it was most likely rainy that day as

$$\begin{aligned} P(\text{rainy}) * P(\text{rainy}|\text{reading}) &= 0.66 * 0.875 = 0.5575 \\ P(\text{sunny}) * P(\text{sunny}|\text{reading}) &= 0.33 * 0.125 = 0.04125 \end{aligned} \tag{1.4}$$

This approach also allows Trillian to infer a sequence of length n of most likely consecutive weather forecasts knowing only Marvin's n consecutive activities on those days by finding a sequence of weather that maximize the conditional probability of given activity events:

$$\text{argmax}_{W=w_1,w_2,...w_n} P(W = w_1, w_2, ..., w_n | A = a_1, a_2, ..., a_n) \tag{1.5}$$

where $W = w_1, w_2, ...w_n$ are the weather forecasts and $A = a_1, a_2, ..., a_n$ are observed activities.

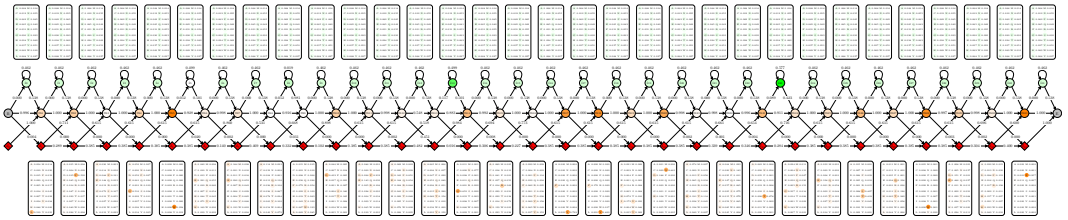


Figure 1.6: **HMM of WW domain.** Visualization of full Hidden Markov Model (HMM) of WW protein domain.

Following the same concepts, we can generate an HMM for a protein domain. Where we define the stationary distribution and transition probabilities from the collected set of similar protein domain sequences. This allows for checking how likely is that a given protein sequence encodes a given domain.

One of the shortest protein domains is the WW protein domain[65]. It is only 31 amino acids long but its HMM is already quite complex (Figure 1.6). Even for such a small protein domain, it is already quite useless to look at the graphical representation of HMM. However, another way to get a feeling of what amino acids are predominant in a given protein domain is to generate a domain HMM logo (Figure 1.5B [3]). Like that, at each position, we can visualize the significance of each amino acid. In the WW domain example, on some positions, there is only one amino acid that was ever found in the curated

[3] created with [163]

set of WW domain sequences (i.e. positions 5 and 28 have always contained Tryptophan (W)), while for some other positions, our domain of interest can have several, almost interchangeable, amino acids (i.e. position 8 can almost equally be occupied by Glutamine (Q), Histidine (H), Tyrosine (Y), Cysteine (C), Arginine (R), Valine (V) or Alanine (A)).

1.3.4 Position Weight Matrix (PWM)

Similar to how we compute the logo representation of protein domains, we can also compute the logo representation of the sequence a given TF binds. With this kind of representation, what we essentially want to do is to define the TF sequence binding specificity. In the context of protein domains, we used HMMs since they cover the transition probability. However, for searching a nucleotide sequence-specific case, we can set up the model such that we care about each nucleotide independently. Like that, we relax the conditions in representing the sequence specificity and consequently, allow for more degrees of freedom when we are predicting the binding of a TF.

For this purpose, we use a position weight matrix (PWM; also known as position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM). PWMs were introduced in 1982 [58, 142] and prior to their introduction, the best way to represent TF binding specificity was the consensus sequences. For the remainder of this thesis, whenever we talk about PWMs or sequence specificity, if not stated otherwise, we are talking about a nucleotide sequence and specificity within that sequence.

A consensus sequence is a simplified representation of the sequence specificity defined by a set of related sequences. For example, sequences known to be facilitating binding of the same TF are collected and aligned[34, 36]. With a defined threshold, position-specific nucleotides are then discriminated based on the frequency at which a nucleotide n is observed at position i. On the other hand, a PWM represents the probability matrix of sequence specificity. Instead of discriminating between nucleotides, it offers a probability for each nucleotide n to be at position i. For example, if we have the following set of nucleotide sequences:

```
AAGGTAAAC
TCCGTAAGA
```

```
CAGCTTGGA
ACAGTCAGT
TAGGTCATT
TAGGGACTG
ATGGTAACT
CAGGTATAC
TGCGTGAGT
AAGGTAAGT
```

the consensus sequence with a threshold of 0.6 is:

```
NAGGTAANN
```

where 1st, 9th, and 10th nucleotides are undetermined since no nucleotides appear on those positions with sufficient frequency. On the other hand, if we use the same set of sequences to generate a PWM, we first start by generating a position frequency matrix (PFM):

$$PFM = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 4 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 2 \\ 2 & 2 & 2 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 1 & 7 & 9 & 1 & 1 & 1 & 5 & 1 \\ 4 & 1 & 0 & 0 & 9 & 1 & 1 & 2 & 5 \end{bmatrix} \tag{1.6}$$

From PFM, we generate a position weight matrix (PWM) by first calculating the position probability matrix (PPM) by dividing all counts by the total number of observations and then converting that matrix into a matrix of log-likelihoods for each letter and position, taking into account the background model. That is, for each position we would like to see how likely it is to have a nucleotide i at position j considering the distribution of nucleotides in the background model. To do that, firstly for each position in the PFM, we compute $M_{i,j} = \frac{M_{i,j}}{\sum_i M_{i,j}}$ and obtain the PPM. Then, we calculate $M_{i,j} = \log_2 \frac{M_{i,j}}{b_i}$ where b_i is the background model probability to see nucleotide i in the dataset. The simplest background model is uniform and is 0.25 for each nucleotide. We add a pseudocount value (usually 0.0001) in order to avoid $-\infty$ values in the PWM which would result if the frequency of nucleotide i was never seen at position j.

$$PWM = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.68 & 1.26 & -1.32 & -11.29 & -11.29 & ... \\ -0.32 & -0.32 & -0.32 & -1.32 & -11.29 & ... \\ -11.29 & -1.32 & 1.49 & 1.85 & -1.32 & ... \\ 0.68 & -1.32 & -11.29 & -11.29 & 1.85 & ... \end{bmatrix}$$

$$\begin{matrix} ... & 1.26 & 1.49 & -0.32 & -0.32 \\ ... & -0.32 & -1.32 & -1.32 & -0.32 \\ ... & -1.32 & -1.32 & 1.0 & -1.32 \\ ... & -1.32 & -1.32 & -0.32 & 1.0 \end{matrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix} \tag{1.7}$$

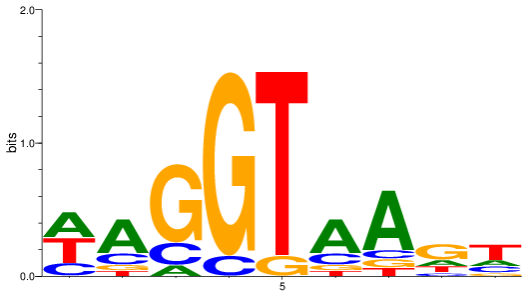


Figure 1.7: **PWM motif**. Example of a motif logo for a generated Position Weight Matrix Equation (1.7).

Visually, we present PWM as a sequence logo (Equation (1.7)), similar to what we had for protein domain logos. In the visual representation of the PWM logo, on the x-axis, we have the *number of bits* for each letter at each position. Since we are using $\log_2$ base, the maximum number of bits is 2. This value essentially represents the *information content* for each position and each nucleotide. That is, if we would have a probability of 1 for nucleotide n at position i, all other nucleotides would have to have a probability 0 of appearing at position i. That would make the log-likelihood of seeing n at position i: $M_{n,i} = \log_2 \frac{1}{0.25} = 2$, which says that, based on the data we have, n is always on position i.

PWMs are important for the prediction of transcription factor binding sites (TFBSs). Given a PWM describing the sequence specificity for a TF of choice, enables the use of computational methods and prediction of potential binding sites on a given set of genomic sequences. Throughout this thesis, we

heavily rely on the use of PWMs.

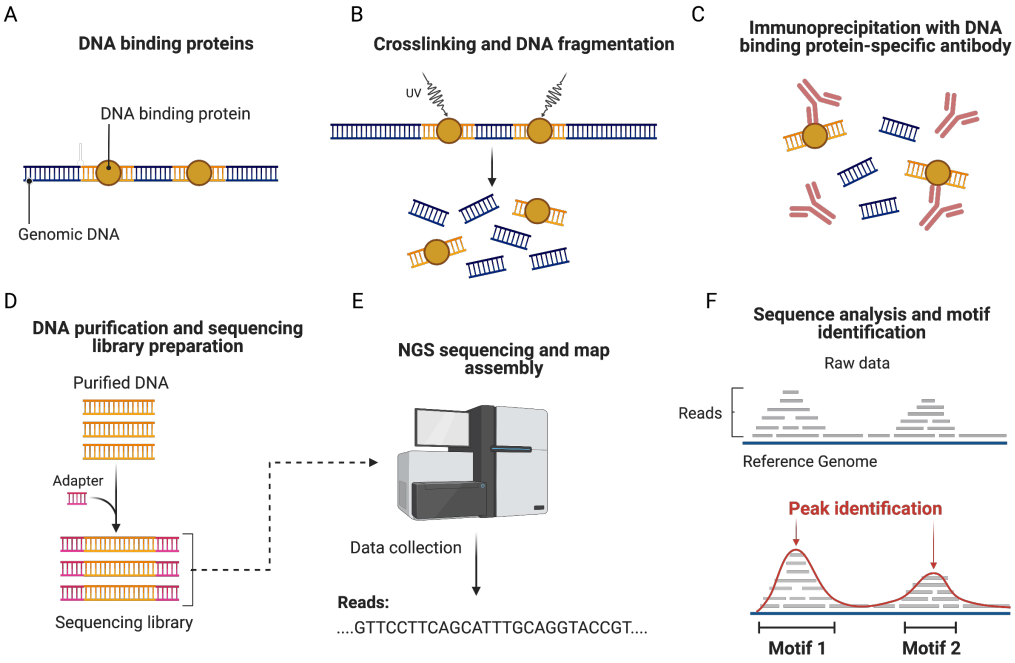


Figure 1.8: **ChIP-seq protocol**. (A) Sample is prepared *in vitro* or *in vivo*. The transcription factor (TF) of choice is mixed with DNA sequences which TF of interest binds. (B) TF and DNA are cross-linked. The next step is fragmenting the DNA so that we can gather only the pieces of DNA that are bound by our TF. (C) DNA-bound TF is immunoprecipitated with an antibody that binds only our TF of interest, in such a way that it does not interfere with TF-DNA interaction. (D) The solution is washed so that only molecules containing DNA-bound TF are extracted. (E) Next, DNA is purified and using next-generation sequencing (NGS), sequences bound by a TF are extracted. (F) Sequenced DNA fragments are aligned to the reference DNA and using computational methods, peaks of stacked DNAs are analyzed which results in the generation of motifs which TF of interest binds.

PWMs are generated using data coming from experiments such as SELEX [154], ChIP-seq [77], protein binding microarray (PBM) [15], or more recently CUT&RUN [83]. PWMs can also be generated computationally by looking at enriched motifs in a given set of DNA sequences with tools like AlignACE [129], MEME [13], or MotEvo [8].

Ideally, we would always have a PWM inferred from experimental data. One of the most popular experiments for this task is ChIP-seq [77]. It gained in popularity with the improvement of sequencing techniques (NGS) and most of the motifs that will be covered in the remainder of the thesis come from that type of experiment. In Figure 1.8 a diagram of the ChIP-seq experiment protocol is shown.

ChIP-sequencing is used to analyze protein DNA interactions by identi-

fying the binding sites of the proteins (TFs)[151]. Upon sample preparation, the DNA and the TFs are cross-linked and the DNA gets sheared in order to fragment the chromatin. Next, a bead-attached antibody against the TF of interest is used in order to immunoprecipitate and therefore isolate the target TF bound to the DNA fragments. DNA recovery and purification are achieved by unlinking the TF from the DNA and extracting the DNA. The DNA fragments are then sequenced and the results are summarized as a set of sequences that can be aligned to the reference genome. Using computational methods [16], a PWM can be generated which describes the TF binding sequence specificity.

1.3.5 Predicting transcription factor binding sites (TFBSs)

Having acquired a PWM for the TF of interest, the next important step is to predict transcription factor binding sites (TFBSs). Correct prediction of TFBS is a really important challenge of computational biology [18, 60] as it would allow us to uncover which TFs regulate expression of which genes [14], or predict the interaction and competition of TFs in gene regulation [25, 79]. This would allow for refined hypothesis generation and more streamlined target selection for future experiments.

Ever since we started acquiring high-throughput sequencing data, the problem of predicting TFBSs was intensely studied [62, 63, 72, 84, 93, 99, 113, 125, 128, 136, 139, 140, 159]. Each of the cited tools addresses the prediction of TFBSs in a slightly different way, and many of the ones that base their modeling on PWMs can be described with a general Bayesian probability framework [157]. A combination of different concepts into one single tool was presented with MotEvo [8]. For a given PWM and set of sequences, MotEvo predicts all TFBSs and for each site, it provides the posterior probability of a TFBS. MotEvo also incorporates the conservation information in regulatory sequences, which has been shown to hold important predictive power when it comes to TFBS predictions[117].

1.4 Controlling gene expression

There are several ways we can perturb gene regulation. We can knock-out, knock-in, knock-down (silence), or overexpress a gene. The first two ap-

proaches introduce irreversible changes to the genome where the expression of a gene is completely stopped (knock-out) or introduced (knock-in). If a gene is knocked-down (silenced), its expression is impaired while on the other hand, when a gene is overexpressed, its expression is increased relative to wild-type (WT) expression. Organisms in which gene perturbation takes place are called mutants.

Mutants can be generated in several ways. For example, gene expression control can be regulated on transcriptional level (by introducing mutations in promoter regions, which create or destroy a TF binding site [76, 130]), on translational level [144] (by synthesizing a RNA complement to the targeted RNA and modify processing of pre-RNA [122], interfering with mRNA transport into the cytoplasm [9], or by preventing mRNA translation [92]) or by directly targeting a gene (either removing it from DNA or replacing with another gene [51, 76]).

1.5 Computational tools for processing gene expression data

Performing an experiment with the aim of perturbing gene expression can be quite a challenging task. Firstly, it is far from easy to hypothesize about genes and their role solely based on experimental data. And secondly, since genes can take on different roles in different tissues or times of development, it can be especially hard to assess their role in a specific situation. However, with the advent and refinement of experimental setups for measuring gene expression, we are generating more and more data that holds valuable information which can help us generate hypotheses about gene regulatory networks (GRNs) using computational methods.

Development of such computational methods started with the results coming from microarray experiments and was quickly extended to processing bulk RNA-seq experiments [14, 30]. These tools were limited to gene expression data coming from a handful of samples. For example, it was possible to compare differential gene expression between two different organs, or from two samples treated with different drugs, but it was not possible to uncover specific cell types contained in a large tissue. With the advent of experimental

setups for measuring RNA content of chromatin accessibility in single cells, we did not only get a chance to look more closely into different cell types within one tissue [50, 171], between different species [23, 135], but we were also able to look at time-series datasets which allow us to study organism development [23, 52, 124]. With the rapid increase of generated data, many computational tools are being published to address the question of understanding GRNs in different systems on the single-cell level[4, 23, 31, 44, 52, 78, 91, 143].

In this thesis, the focus lies on the use of two computational tools for analyzing RNA-seq data. The following paragraphs are a brief explanation of the ideas behind these tools as their output is used in the analysis presented in Chapter 2.

1.5.1 Modelling gene expression in terms of genome-wide regulatory sites (ISMARA)

Integrated System for Motif Activity Response Analysis (ISMARA)[14] is a computational pipeline for modeling gene expression data in terms of genome-wide regulatory sites. For expression matrix E_{ps} and sitecount matrix N_{pm}, ISMARA infers transcription factor motif activities by fitting a linear model:

$$E_{ps} = \sum_{m} N_{pm} * A_{ms} + noise \tag{1.8}$$

The expression matrix, E_{ps}, represents the mRNA readout for each of the genes in a system of choice at a given time. For each sample s we have the total amount of mRNA expressed by promoter p. The sitecount matrix, N_{pm}, is a pre-computed resource that gives us the information of which motifs are targeting which promoters. That is, for each motif m we have a probability of it binding promoter p.

Linear model (Equation (1.8)) fits motif activities A^*_{ms} and also outputs motif activity error bars δA^*_{ms}. Having fitted A^*_{ms}, we can assess how well ISMARA captures the expressional changes between different samples by looking at the fraction of variance (FoV) explained by the fitted model.

For each motif target, ISMARA computes the contribution of a motif in modeling gene expression across the whole dataset. This is done by calculating

the difference in error of the predicted expression *with* (using the original N_{pm} sitecount matrix) and *without* the predicted motif site (modified sitecount matrix N'_{pm}):

$$S_m = \log\left[\frac{\int P(E|N,A)dA}{\int P(E|N',A)dA}\right] \tag{1.9}$$

This score allows for the sorting of motif targets based on inferred significance. Furthermore, ISMARA sorts motifs based on computed z-score z_m

$$z_m = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{A^*_{ms}}{\delta A_{ms}}\right)^2} \tag{1.10}$$

This allows for sorting motifs based on how much they vary across all analyzed samples.

ISMARA also offers additional analysis of motif targets such as first-level regulatory network, GO enrichment analysis [1, 10] of top motif targets as well as the analysis of top targets by String database[147].

ISMARA is publicly available at ismara.unibas.ch and is ready to analyze RNA-seq and microarray datasets, directly coming from an experiment or even publicly available databases.

1.5.2 Single-cell reconstruction of developmental trajectories (URD)

URD is a computational tool for reconstructing developmental trajectories from time series scRNA-seq data [52, 137]. The first step in studying organism or tissue development is the acquisition of expression data from multiple time points. Such datasets allow for computational modeling of cell differentiation and cell fate determination processes. Briefly, the URD algorithm can be described with the following steps:

1. For all cells in the dataset, K-Nearest-Neighbour (KNN) graph is constructed based on the transcript data (mRNA counts for each cell). Distance between cells is defined by edge values assigned during KNN. These distances are later used for defining transition probabilities between cells

(i.e. how likely is that a cell would transition between two expression states).

2. The user identifies the root of the tree as a starting point of development (from the earliest time-point of expression data) and the tree leaves as the endpoint of developmental course (from the last time-point of expression data). Usually, there is only one root, assuming that all cells in the earliest time point are in the same expression state but there can be more roots defined. Leaf points are identified through cell clustering and manual curation of resulting clusters based on marker gene expression.

3. Each cell is assigned a pseudotime that reflects its developmental state, relative to the other cells in the dataset. Pseudotime is calculated by multiple simulations of diffusion processes starting from cells in the root state. The average number of diffusive transitions to get from the root to a cell represents its pseudotime value. This value correlates well with real-time (time points at which cells were harvested) but does not match it since cells usually go through asynchronous development and can have the same transcriptional states at different time points.

4. Once every cell is assigned a pseudotime value, a developmental trajectory is constructed bottom-up. That is, starting from leaves, cells "perform" random walks to other cells, but only in direction of equal or earlier pseudotime points, so that cyclical structures in the tree are avoided. This results in cells "coming together" in branching points of the tree in the next step.

5. The final tree is reconstructed by summarizing inferred trajectories and merging similar parts into larger branches.

6. Finally, the tree is visualized with a force-directed layout which relies on the number of visits each cell had from other cells in their random walk.

URD generates developmental trajectories *de novo* and is agnostic to prior knowledge. It is freely available at https://github.com/farrellja/URD.

1.6 Thesis outline

In this thesis, the problem of modeling gene expression by looking at the specificity of TF binding and its influence on gene regulatory networks is addressed. It is presented in two major parts: 1. enumerating gene regulatory elements for a given organism and modeling gene expression in terms of genome-wide regulatory sites and 2. modeling changes in promoter gene expression in terms of single TF binding affinity.

In Chapter 2 we present an automated pipeline for inferring the necessary set of ingredients to model gene expression in terms of genome-wide TFBS. In brief, the pipeline consists of 1. inferring a list of TFs for an organism of choice, 2. mapping previously experimentally determined and manually curated set of motifs to inferred TFs, 3. generating a promoter set from transcript data and, 4. predicting of genome-wide TFBS while considering conservation information between related genomes. We demonstrated the use of the presented pipeline by inferring a gene regulatory resource in zebrafish (*Danio rerio*). With this pipeline, we enhanced a previously developed tool in the lab called ISMARA[14] for processing zebrafish data. Using ISMARA for zebrafish, we analyzed several bulk RNA-seq and one single-cell RNA-seq. Agnostic to previous knowledge, we predicted known and novel regulators across zebrafish tissues.

In Chapter 3, we zoom in and investigate if a promoter sequence encodes sufficient information for the prediction of changes in gene expression in terms of TF binding affinity. We turn to housekeeping gene regulation by σ^{70} TF in *Escherichia coli*. Using a set of synthetic constitutive promoters expressing in two regimes (medium and high expression), regulated only by the σ^{70}, we model gene expression in terms of σ^{70} binding affinity. We showed that using the concepts from thermodynamics to model gene expression, we cannot explain the change between medium and high expressors solely based on σ^{70} binding affinity and derived features.

Chapter 2 is presented as a standalone publication and Chapter 3 is presented as part of the project work that is still in progress.

2

A pipeline for genome-wide annotation of transcription factors, their sequence specificities, and binding site

Đorđe Relić[4,5], Mikhail Pachkov[4,5], Alexander F. Schier[4,6] and Erik van Nimwegen[4,5]

Abstract

For some organisms, such as human and mouse, hundreds of transcription factors (TFs) have been extensively studied by experimentally inferring their sequence-specific binding motifs, predicting transcription factor binding sites (TBFSs) genome-wide, and computationally modeling transcriptomic and epigenomic data in terms of these regulatory sites. In contrast, for other organisms, sophisticated computational methods modeling gene expression have been restricted to a handful of experimentally investigated TFs. In this study, we present a broadly applicable computational pipeline that generates a gene regulatory resource for any given organism of choice and apply this pipeline to

[4] Biozentrum, University of Basel, Basel CH
[5] Swiss Institute of Bioinformatics, Basel, CH
[6] Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA

zebrafish (*Danio rerio*). The pipeline consists of 1. identification of genes containing DNA binding domains (DBDs), 2. inferring TF binding motifs by leveraging DNA-binding domain (DBD) similarity to TFs from a database of known sequence specificities, 3. annotation of promoters genome-wide from transcript sets, 4. aligning promoter regions with orthologous regions from related genomes and, 5. prediction TFBSs across promoters using the MotEvo algorithm. By leveraging the similarity of zebrafish to human and mouse, we use our pipeline to infer gene regulatory interactions in zebrafish. To model gene expression data in terms of predicted regulatory sites, we integrate these results into our previously developed Integrated System for Motif Activity Response Analysis (ISMARA) and predict novel TF-promoter interactions in zebrafish. The presented pipeline will help prediction of gene regulatory networks in other organisms and make other gene regulation modeling tools more broadly applicable.

2.1 Introduction

Understanding the gene regulatory networks that control gene expression is a central question in molecular biology. Much of gene expression is controlled through transcription initiation whose regulation is ultimately encoded in the constellations of small sequence motifs in the DNA that are bound by transcription factors (TFs) in a sequence-specific manner. Thus, a key step toward understanding gene regulation within a given organism is to comprehensively annotate TFs, identify the sequence-specificities of each TF, and map transcription factor binding sites (TFBSs) in the genome.

Currently, there are only a handful of organisms for which such resources are available for the majority of the TFs, including yeast, mouse, and human. There are several reasons for the lack of such resources in other organisms, but two reasons stand out. First is the popularity of an organism in studying a specific research question. For example, disease model organisms (such as mouse) are more studied in terms of TF-specific gene regulation than embryonic developmental model organisms (such as zebrafish) where a complete cell gene expression state is of greater interest. Second, established experimental techniques do not have the same efficacy in different organisms. One example is ChIP-seq[77], where it is crucial to have the right antibody for the TF of

interest, and finding one in the list of currently available anti-bodies requires laborious effort which often can result in unsatisfactory results[121].

In the lack of experimentally validated resources, we resorted to computational methods to infer a set of TFs with binding specificities. Here, we present a computational pipeline for inferring a gene regulatory network resource. First, assuming that TFs with similar DNA-binding domains (DBDs) bind similar motifs, we infer a set of TFs in an organism of choice and map them to the previously experimentally validated set of motifs. Second, we annotate promoters genome-wide from transcript sets. Third, we align the inferred promoter regions with orthologous regions from related genomes and predict transcription factor binding sites (TFBSs) using the MotEvo[8] algorithm. We demonstrate the use of our pipeline to infer gene regulatory interaction resource in zebrafish. To map a set of motifs to zebrafish TFs, we leveraged its similarity to human and mouse in terms of TF encoding DBDs. Our pipeline yields a set of 994 zebrafish TFs mapped to 552 unique motifs with TFBS prediction on 36259 zebrafish promoter regions aligned to common carp, goldfish, and grass carp.

To use the generated resource and predict gene regulation in zebrafish, we employed a tool earlier developed in the lab called Integrated System for Motif Activity Response Analysis (ISMARA)[14]. For a given experimental transcriptomic data set and set of genome-wide TFBS predictions, ISMARA models gene expression in terms of genome-wide motif activity. Furthermore, it infers the first level gene regulatory network, top targets of each motif and, target GO category enrichment[1, 10] as well known target connections in String database[147].

We analyzed several publicly available zebrafish bulk RNA-seq datasets coming from heart, liver, brain, muscle, blood, gill, testis, oocyte, and ovary tissues. Agnostic to previously published studies, we predicted known down-regulation of rest targets in non-neuronal tissue, up-regulation of hnf-family targets in liver tissue, and up-regulation of rfx-family targets in brain and testis tissues. Furthermore, we predicted previously unknown up-regulation of grhl1 targets in gill and blood tissues and down-regulation of zbtb14 targets in non-neuronal tissue. We also analyzed one single-cell RNA-seq dataset coming from the 12 hpf stage of zebrafish embryonic development clustered to 31 distinctive cell types. We predicted previously known up-regulation of

myod1 targets in adaxial cells. Furthermore, we predicted the differential role of two gata motifs in epidermis tissue. We find that gata2a targets are up-regulated, as it was expected based on the available literature, while gata3 targets are down-regulated which presents a novel hypothesis about the role of gata3 in epidermis tissue.

2.2 Results

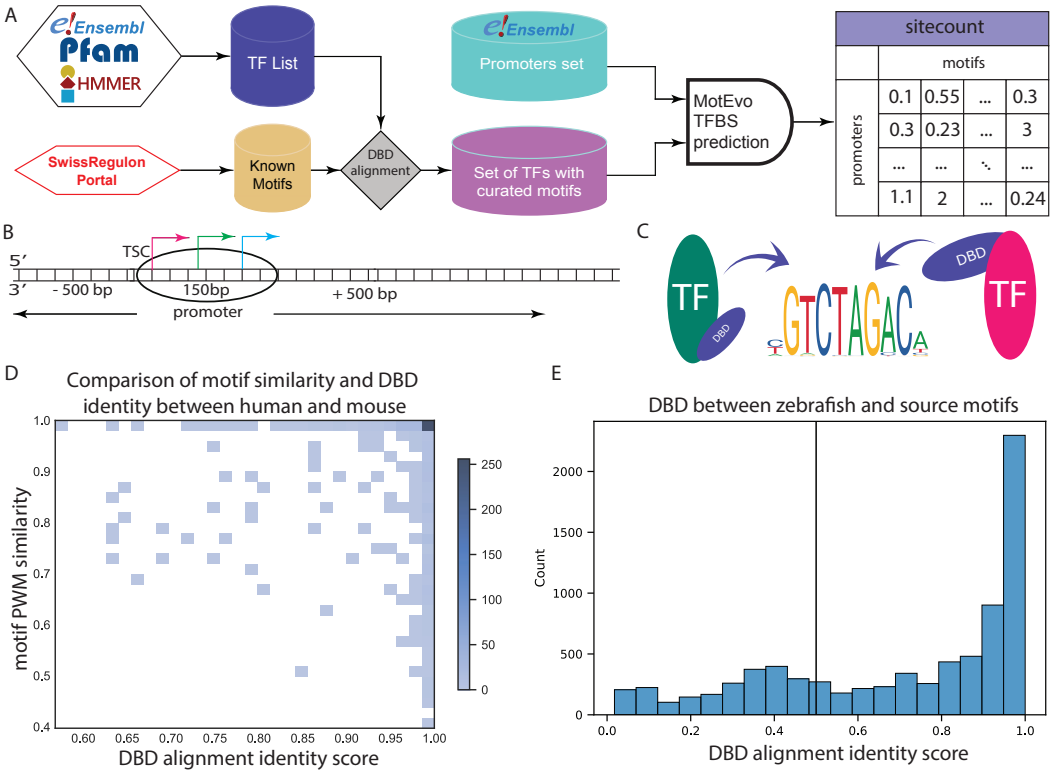


Figure 2.1: **Computational pipeline for inferring a list of transcription factors (TFs) with binding motifs.** (A) Diagram of the pipeline. (B) Graphical representation of the algorithm for inferring a set of promoters. (C) Transcription factors that encode similar DNA-binding domains (DBDs) have the same sequence specificities (motifs). (D) Demonstration of the hypothesis in (C), where we show that similarity in DBD closely matches similarity in motifs. (E) The majority of TFs in zebrafish have identical DBDs to those found in human or mouse. In this study, we consider correct mappings only if DBDs match for more than 0.5 in DBD sequence similarity.

2.2.1 Transcription factors encoding similar DNA-binding domains bind similar motifs

Transcription factors (TFs) are proteins that bind DNA in a sequence-specific manner and regulate gene expression. Thus, proteins that encode for one or more DNA-binding domains are putative TFs. The first step in inferring a list of TFs in an organism of choice is to collect all known DBDs. Pfam[111] aggregates manually curated protein domain families and provides corresponding Hidden Markov Models (HMMs) of each protein domain family. We semi-manually curated a list of over 19 000 Pfam protein domains and, identified a total of 136 protein domains that are present in higher eukaryotes and are known to be DNA-binding (Sup. Table S1; Materials and Methods: Inferring transcription factors). Using `hmmsearch` from software suite HMMER[53], we scanned principal reference protein sequences from zebrafish, human and mouse and inferred 3105, 2790, and 2322 putative TFs respectively (Materials and Methods: Mapping motifs)

We hypothesized that TFs which encode for similar DBDs bind similar motifs. To put this hypothesis to test, we retrieved an experimentally validated set of TFs from SwissRegulon[119] and compared TFs matching based on DBD identity with motif similarity between human and mouse (for details see Materials and Methods: Mapping motifs). Our analysis finds that this hypothesis holds as it is shown in Figure 2.1D.

By matching TFs from zebrafish to TFs in human and mouse, we find that for majority of TF pairs based on DBD identity actually encode for identical DBDs (Figure 2.1E; Materials and Methods: Mapping motifs). We set a threshold of 0.5 in order to ensure that motifs assigned to zebrafish TFs match at least 50% in terms of DBD identity. Finally, we produced a list of 994 zebrafish TFs mapped to 552 unique motifs. This resource is accompanied by a web application where for each predicted motif in zebrafish, we present the full list of TFs which could bind that motif, sorted by the DBD identity score to the TF from which the motif was inferred (http://brlauuu.pythonanywhere.com/ username:zebrafish_tfs, password:zftf2021).

2.2.2 Promoter regions are highly conserved between zebrafish, common carp, goldfish and grass carp

Promoters are DNA sequences located around the gene transcription start site (TSS). Earlier we developed an algorithm for generating a set of promoters which we applied to human and mouse[119]. In this work, we applied the same approach to infer a set of zebrafish promoters. Briefly, the algorithm consists of the following steps: 1. collection of transcription start site (TSS) annotation from Ensembl, 2. using single-linkage clustering, generation a set of transcription start clusters (TSCs) gathering TSSs within blocks of 150 base-pairs and 3. defining a promoter region as +/- 500 base-pairs around a TSC (Figure 2.1B). From 44802 transcripts coming from 25102 genes we retrieved from Ensembl, we infer 36259 promoters. In comparison, the same approach yielded 37700 promoters in human and 30115 promoters in mouse. The majority of genes are mapped to one promoter and in rare cases, they are mapped to up to 20 different promoters (Figure A.1B).

Conservation in promoter regions stores valuable information in terms of predicting transcription factor binding sites[117]. To make use of this evolutionary feature, we aligned the zebrafish genome (danRer11 [67]) to the genomes of 3 other fish: Grass Carp (*Ctenopharyngodon idella*; cteIde1 [115]), Common Carp (*Cyprinus carpio*; cypCar1 [57]) and Goldfish (*Carassius auratus*; casAur01) [35]). We find that each pair of aligned fish genomes has at least 55% conservation of the whole genome while more than 85% of promoter regions is conserved between all 4 fish (Figure A.2A-B). Taking the fraction of conservation between the fish, we generated a phylogenetic tree depicting the conservation distance between the fish (Figure 2.2B). This phylogenetic tree, together with aligned promoter regions, improves our transcription factor binding site (TFBS) prediction.

2.2.3 Genome-wide transcription factor binding site (TFBS) prediction

Having defined aligned promoter regions and inferred a set of motifs mapped to TFs in zebrafish, we predicted genome-wide transcription factor binding sites (TFBSs) using MotEvo[8]. MotEvo is a Bayesian probabilistic method for the prediction of TFBSs from multiple alignments of phylogenetically re-

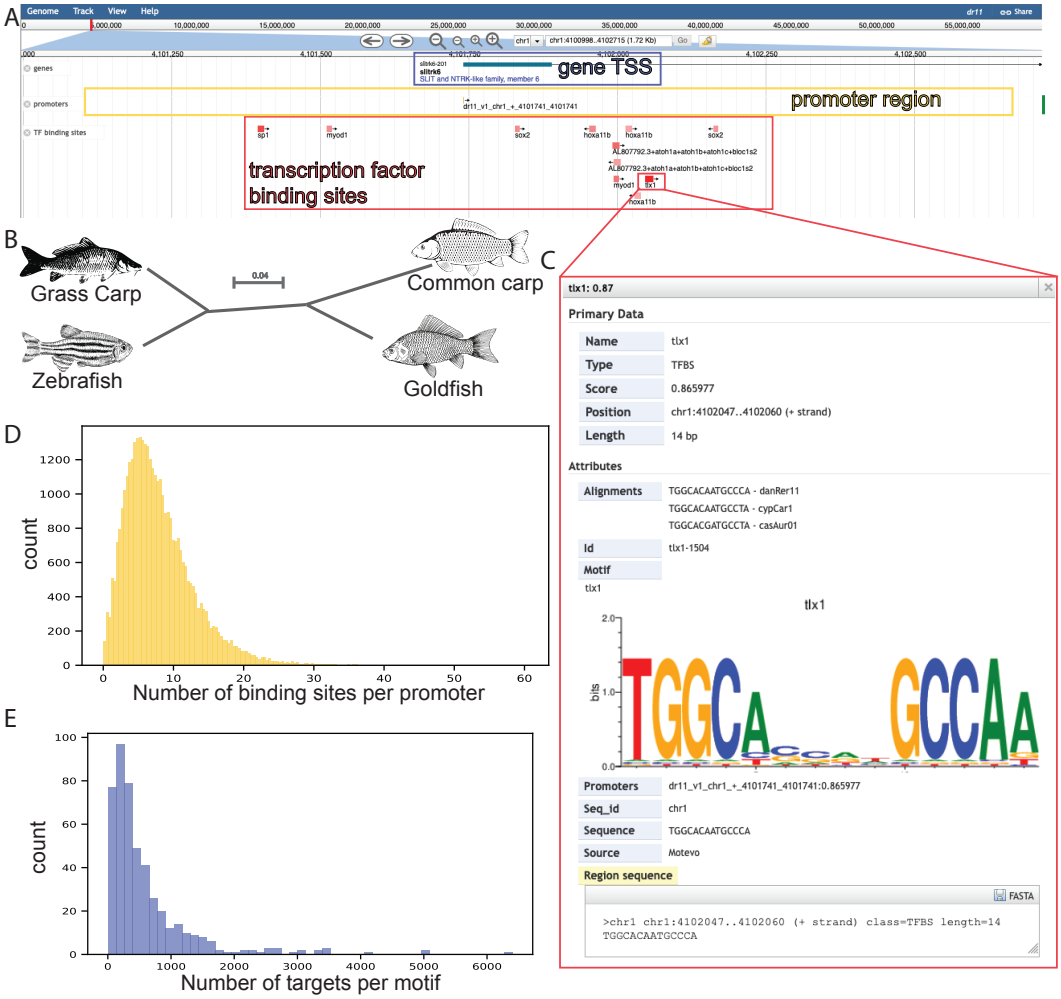


Figure 2.2: **Multiple-genome alignment ensures accounting for conservation information between species when predicting transcription factor binding sites** (A) Subset of TFBS predicted on the promoter of SLIT and NTRK-like family - member 6 (slitrk6) gene. Together with promoter region, in orange, transcription start cluster (TSC) in blue, all TFBSs are annotated in red with the significance of predicted binding (white to red color scale). (B) Using distance calculated from whole-genome alignment between each two fish, we compute a phylogenetic tree for zebrafish, grass carp, common carp, and goldfish. (C) Example of binding information of motif group tlx1 on slitrk6 promoter together with conserved sequences in all other fish. (D) Distribution of predicted binding sites per promoter for all 36259 promoters in zebrafish. (E) Distribution of predicted targets for all 552 motifs in zebrafish.

lated DNA sequences. It incorporates several features into its model including competition of multiple TFs, forming of TF clusters which form cis-regulatory modules, and evolutionary modeling of conservation of TFBSs across species. Running MotEvo on 36259 aligned zebrafish promoter regions with 552 motifs results in average of 805 promoter targets per motif and 10 binding sites per promoter (Figure 2.2D-E; Materials and Methods: Predicting TFBSs).

Each of the predicted TFBSs can be examined through the interactive portal (Figure 2.2A and C). Furthermore, TFBS annotation is freely available at https://swissregulon.unibas.ch/sr/downloads.

2.2.4 Modelling gene expression in terms of genome-wide transcription factor binding site regulatory sites

ISMARA[14] is an established integrated computational pipeline for modeling gene expression in terms of genome-wide regulatory sites. It fits a linear model $E_{ps} = \sum_m N_{pm} * A_{ms} + noise$ where the input parameters are E_{ps}, experimentally acquired expression data where expression of promoter p is quantified in sample s and N_{pm}, sitecount matrix which recapitulates predicted probability of motif m binding promoter p (depicted as final result in Figure 2.1A diagram). The fitted parameter is the activity matrix A where A_{ms} represents an estimation of the expression fold change of targets of motif m in sample s relative to mean target expression across all samples. ISMARA lists motifs sorted by the most significant change in motif activity across all samples. Furthermore, motif targets are sorted by the predicted binding site contribution to the modeling of target expression. That is, for each motif m and each target promoter p, ISMARA computes the target score as the change in fraction of variance (FoV) explained *with* and *without* predicted binding site of motif m in promoter p (for more details see [14]). For targets predicted to be regulated by motif m, ISMARA calculates several statistics such as GO annotation enrichment[1, 10] in several categories and known target connections in the String database[147]. Furthermore, ISMARA infers the first level gene regulatory network around motif m which contains motifs that are regulating or are regulated by motif m.

Using ISMARA, we analyzed three bulk RNA-seq datasets and one single-cell RNA-seq dataset. Bulk RNA-seq datasets gather tissue samples coming from time course and several replicas of heart, brain, liver, gill, and muscle [88], single replicas of brain, heart, liver, muscle, and blood[82], and several replicas of oocyte, ovary, and testis[64]. Single-cell RNA-seq dataset comes from 12 hpf stage of zebrafish embryonic development[52]. Transcriptomics data coming from 2131 single cells were clustered and manually curated into 31 distinctive clusters. We generated pseudobulk data from curated clusters (see Materials and Methods: Pseudobulk ISMARA) and treated each of the

clusters as a separate tissue sample.

Agnostic to previously published knowledge, ISMARA predicted known and novel regulators in zebrafish tissues.

2.2.4.1 Bulk RNA-seq

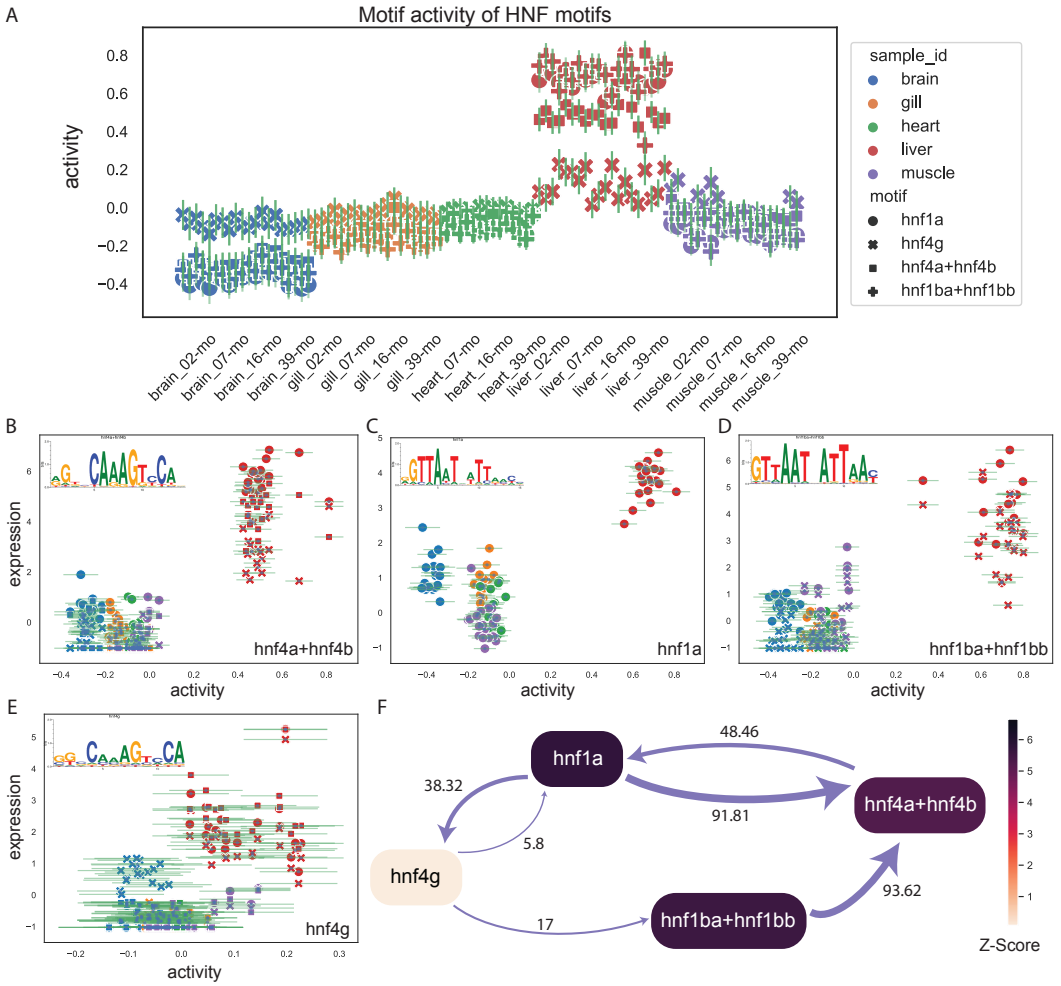


Figure 2.3: **HNF TF family is an activator of transcription in liver tissue.** (A) Motif activity of hnf1a, hnf4a+hnf4b, hnf4g and, hnf1ba+hnf1bb is up in liver tissue in comparison with other tissues. Each symbol is a different motif, error bars on motif activity are shown in green and different tissues are in different colors. (B-E) Expression of all HNF TFs is positively correlated which implies that they are activating their targets. Each symbol, represented in different colors (noting different tissues corresponding to the legend in (A)) shows promoter expression plotted against motif activity. (F) The regulatory network between the HNF TFs. The color of the cells represents the inferred z-score, while numbers and thickness of arrows represent the target scores (For more details see [14]).

Known regulators. ISMARA analysis yielded several expected key regu-

lators. For example, we predicted down-regulation of rest targets in non-neuronal tissue (rest is known to be repressing genes involved in neuronal development [134]) as well as up-regulation of RFX-family TFs in brain (ciliogenesis [37, 96]) and in testis (spermiogenesis [167]). In both gill and blood samples, we found up-regulation of IRF-family targets, which is confirmed in the literature [2, 3]. Furthermore, we noted up-regulation of genes targeted by the HNF-family of TFs in the liver which is expected [94]. In Figure 2.3 we show findings for four HNF motifs: hnf1a, hnf4g, hnf4a+hnf4b and hnf1ba+hnf1bb. All 4 motif activity profiles are well correlated with the expression of the corresponding TFs which implies that hnf TFs are activating their targets in liver tissue (Figure 2.3A-E). Furthermore, we also predicted the interaction between these HNF factors (Figure 2.3F). While the regulatory interaction between hnf1a/b and hnf4a/b is known from earlier studies in mice[94], positive regulation between hnf1ba/b to hnf4a/b is potentially new interesting insight.

Potential novel regulators. Our analysis yielded novel regulators which are potentially interesting for further exploration. We find implications of an activator role of grainy head-like transcription factor 1 (grhl1) in gill and blood tissue (Figure A.3). To the best of our knowledge, we have not found studies exploring the role of grhl1 in gill and blood samples and this might be an interesting insight for further investigation. Furthermore, we note the repressive role of zbtb14 TF in non-neuronal tissue, similar to the one of rest. Zbtb14 was not extensively studied in zebrafish, however, it was earlier shown that Xenopus paralog of zbtb14 suppresses genes involved in neuronal development[149].

2.2.4.2 scRNA-seq

Known regulators. We identified correlated motif activity of myod1 motif and expression of the myod1 promoter across all samples (Figure A.4B). Expression of myod1 is highest in adaxial cells cluster where its targets are up-regulated (Figure A.4A). Moreover, when we looked at the zebrafish embryo development tree generated in [52], we found that myod1 is exclusively expressed in the branch leading to the formation of adaxial cells (Figure A.4C) This finding is confirmed by the literature since adaxial cells are known precursors to specific muscle fiber types and myod1 is a known marker for muscle formation[45].

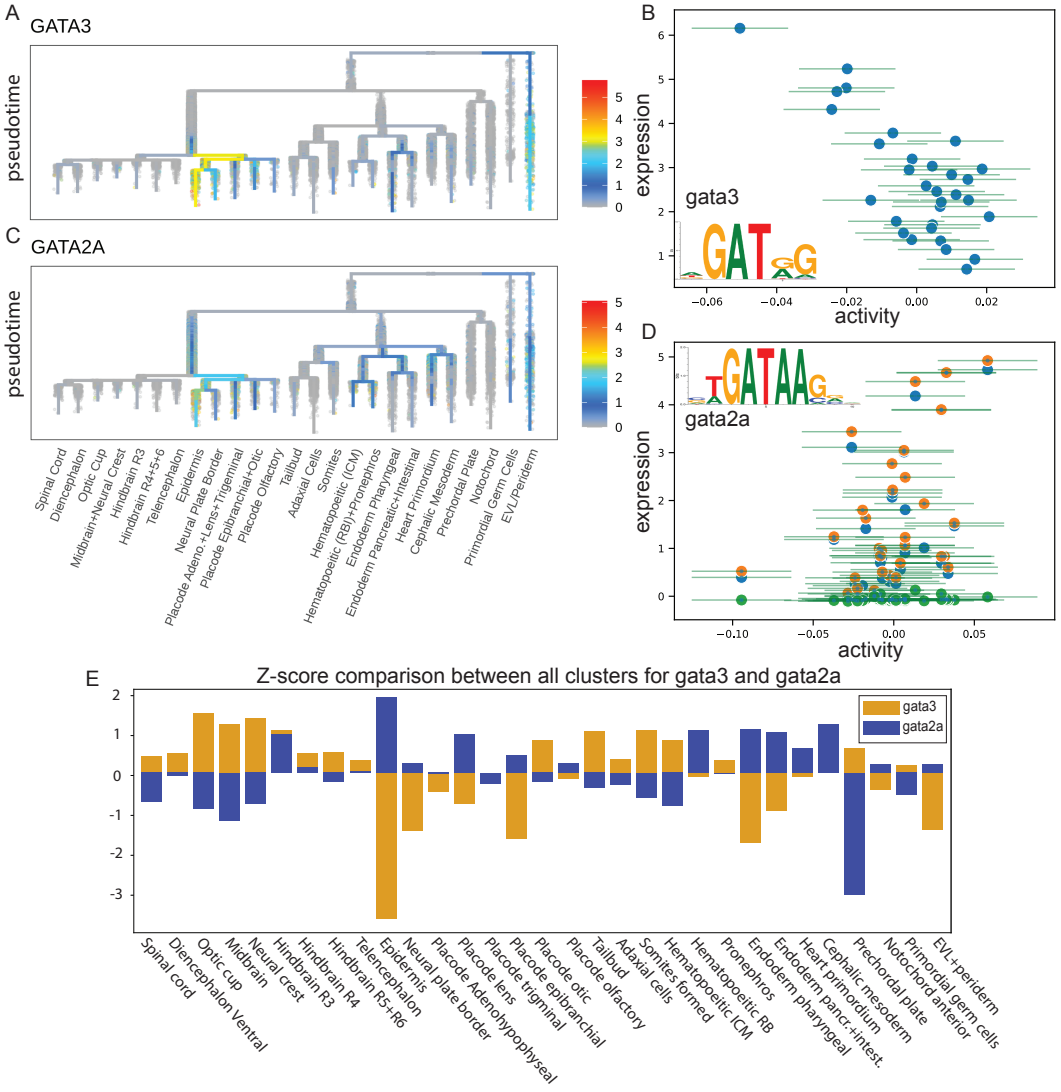


Figure 2.4: **Differential role of gata3 and gata2a in epidermis cell type.** (A) Gata3 is exclusively expressed in the part of the zebrafish development tree leading to the epidermis which is a predecessor cell type to muscle tissue. (B) Gata3 is repressing its targets in the epidermis cluster as its expression is anti-correlated with its activity. (C) Gata2a is a known epidermis marker [52]. (D) Gata2a is an activator TF, with the highest activity in the epidermis cluster. (E) Gata3 and gata2a are significantly enriched in epidermis tissue compared to all other tissues with gata3 being the most down-regulated and gata2a is the most up-regulated motif in the epidermis.

Potential novel regulators. Our analysis yielded an interesting finding capturing the differential role of two GATA motifs. We found that gata2a targets are up-regulated in the epidermis while gata3 targets are down-regulation in the same tissue. Moreover, our analysis showed that regulation mediated by these two gata motifs is most enriched in epidermis tissue (Figure 2.4E). Looking at the correlation between expression and motif activity, our analysis

implied that gata3 has a repressive role while gata2a is exhibiting an activator role (Figure 2.4 panels B and D). Expression of both gata3 and gata2a is almost exclusively leading to the development of epidermis (Figure 2.3 panels A and C). Gata2a has been previously identified as the marker for epidermis [52], unlike gata3. However, we found reported in the literature that gata3 has a role in skin cell differentiation in mice[81], but, to the best of our knowledge, we have not found any data for the role of gata3 in zebrafish embryonic development. By looking at the GO annotation enrichment of both gata2a and gata3 targets, our analysis implied that gata3's repressive role is focused on slowing down cell proliferation (positive regulation of cytoplasmatic translation, maturation of LSU-rRNA, positive regulation of mRNA splicing) while the gata2a target activation promotes cell differentiation (bleb assembly, slow muscle cell migration, regulation of blood vessel endothelial cell proliferation involved in sprouting angiogenesis).

2.3 Discussion

Gene regulation is one of the most important questions of molecular biology and annotation of regulatory elements is the first crucial step in modeling gene expression on a genome-wide scale. To model gene expression and infer gene regulatory networks in a given system, one needs to annotate all regulatory sequences in the DNA, enumerate TFs binding those sequences, and predict most likely genome-wide TFBSs. There have been numerous studies addressing parts of this problem by identifying a set of promoters [47], identifying DNA-binding domains[7, 68, 161, 162] and predicting TFBSs [68, 123, 164, 172].

In zebrafish, the most notable resources covering part of the task we were set to address here include the Eukaryotic Promoter Database [47] which hosts annotated sets of promoters in eukaryotes and CIS-BP [162] which hosts sets of TFs with computationally inferred motifs. Firstly, we improved the annotation of both annotated TFs with motifs and inferred promoter regions by using the latest release of zebrafish genome (danRer11) while both EPB and CIS-BP resources are based on earlier genome versions, danRer7 and danRer10 respectively. Secondly, we generated a set of 36259 promoters which, in comparison to the 10728 promoters hosted at EPB, extends the set of annotated zebrafish promoters for more than two-thirds. This allows for wider genome coverage in

terms of predicting TFBSs. Thirdly, after comparing annotation of TFs with mapped motifs between CIS-BP and one generated in this study in terms of modeling gene expression using ISMARA, we found that we present a larger set of motif groups (308 compared to 475) which in turn increase the predictive power of ISMARA by 33% in terms of the fraction of variance explained by analyzing data from [88] (Figure A.5; Materials and Methods: Comparing different annotations of TFs with mapped motifs).

There are, of course, limitations to our approach that we aim to address in future work. First, our approach to infer a set of TFs potentially yields a fraction of false positives. This issue is symptomatic with all previous computational methods that tackled this problem. Here, we implicitly mediate this issue by considering only TFs with a significant fraction of conserved DBDs which have experimentally confirmed motifs. Taking this approach, we essentially take only about one-third of all putative TFs inferred by identifying DBDs across zebrafish proteome and discarding many potential false positives. Second, one could argue that looking at DBD similarity to map TFs with known motifs could be replaced by just looking at homologous genes between species annotated at Ensembl. However, with that approach, we found that out of 994 inferred TFs with assigned motifs, we would keep only 663, which accounts for the loss of 30% of regulators. Furthermore, in Figure A.1C, we show that the majority of our TF matching between species is scoring as good if not higher on the full protein sequence alignment score. We also found that for 321 TFs, which do not have an annotated homologous in the Ensembl database, there is a significant similarity in DBD sequence ($> 50\%$) even though the full protein sequences do not match (Figure A.1D). Lastly, we are aware of the limitations introduced by predicting TFBSs only in the proximal promoter regions. Enhancers, distal regulatory sequences, are shown to be playing crucial roles in gene regulation, especially in development[97, 110, 168]. However, accurate annotation of all enhancers, across tissues and different conditions, as well as mapping them correctly to genes they regulate is still an open question.

In conclusion, in this study, we present an automated pipeline to infer a set of regulatory elements in a given organism. This pipeline represents a clear, step-by-step set of instructions on how anyone can annotate a set of TFs with binding specificities and predict TFBS for a given organism by leveraging knowledge we have for other, known, and better studied, organisms. We

demonstrate the usage of this pipeline by inferring a set of regulatory elements in zebrafish by leveraging its similarity with human and mouse. Finally, we employ the generated resources and prepare ISMARA, a tool previously developed in the group, and infer known and novel key regulators found in several RNA-seq and scRNA-seq datasets. We anticipate that the presented pipeline will help the generation of gene regulatory networks in other organisms and make other gene regulation modeling tools more broadly applicable.

2.4 Materials and Methods

2.4.1 Inferring transcription factors

We identified all DNA-binding domains from Pfam[111] by scanning the name and description fields for "DNA binding" or "transcription". Furthermore, we also included domains known to have DNA binding activity presented in [161]. Then, for each of the selected DBDs, we extracted the respective Hidden Markov Model (HMMs).

From Ensembl[67], we retrieved all known protein sequences in zebrafish, human, and mouse. Considering that many genes are assigned different transcripts (due to post-transcriptional control mechanisms) which ultimately translate to different amino acid sequences, we made use of the Appris database [127] and selected only for principal reference proteins. This accounts for 33665, 32368, and 29349 protein sequences coming from 25102, 22166, and 23031 annotated genes in zebrafish, human, and mouse respectively.

Next, we ran `hmmersearch` with default settings on principal reference proteome. We only take into consideration confident hits. This analysis yielded a total of 3105, 2790, and 2322 putative TFs in zebrafish, human, and mouse respectively. Details about confident and non-confident hits can be found in HMMER manual[49].

2.4.2 Mapping motifs

DBDs identified in protein sequences are syntenically concatenated. Concatenated DBDs from zebrafish (query species) are aligned against concatenated DBDs of those in human and mouse (source species) using BLAT[85].

We ran BLAT with default settings and compute "DBD identity score" as $DBD_s = n_m/l_q$, where n_m is number of matches and l_q is the length of the query (concatenated DBD sequence). We set the threshold of 0.5 for the DBD identity score, which essentially says that only TFs which match for more than 50% in DBD domains are mapped (Figure 2.1E). Next, we filter for TFs with the highest mapping score. There are a significant fraction of TF mappings where DBDs between zebrafish aligns equally well with both human and mouse TFs (sometimes even to multiple ones). For all "tied" matches we apply the following tie-resolution strategy:

1. Needle[101] align full protein sequences and choose the highest-scoring one (using needle score).
2. If motifs mapped to human/mouse TFs are identical, solve the tie by removing redundancy and choosing human motif.
3. If motifs are different and the tie was between human and mouse TF, choose human TF.
4. Using the sitecount matrix (see Predicting TFBSs), choose the motif that has 2 out of 3 higher values in the following metrics: mean binding probability, binding probability variance, and the total number of sites across zebrafish promoters.

SwissRegulon database [119] hosts a curated set of experimentally validated 500 motifs mapped to 684 TFs in human and 503 motifs mapped to 680 TFs in mouse. Assuming that, if two TFs encode similar DBDs, they are likely to bind similar binding sites in the genome (Figure 2.1C).

Using the SwissRegulon resource and matched TFs between zebrafish (query species) on one side and human and mouse (source species) on the other, our analysis resulted in 552 unique motifs mapped to 994 TFs in zebrafish. The list of inferred TFs in zebrafish with mapped motifs can be found in Sup. Table S2.

2.4.3 Multiple species alignment

We acquired whole genomes for 4 fish [35, 57, 67, 115]. Firstly, we used `last`[87] to perform whole-genome alignment between all pairs of fish, then

calculated fraction of conserved nucleotides in the promoter region, and from that information calculated the phylogeny tree between these fish (Figure 2.2B; Figure A.2B). The phylogeny tree is defined using the calculated the fraction of base-pair similarity between the fish f_{ij} with formula $f_{ij} = \frac{1}{4} + \frac{3}{4}\exp(-4\frac{d_{ij}}{3})$, where d_{ij} the distance between i and j. Next, we used `t-coffee`[118] to generate multiple species alignment from all pair alignments. Using the set of promoters we identified for zebrafish, out of the multiple-genome alignment, we only select the parts that are coinciding with the zebrafish promoters. More than 85% of promoter regions are conserved between all 4 fish (Figure A.2B).

2.4.4 Predicting TFBSs

Earlier in the lab, we developed MotEvo[8], an integrated suite of Bayesian probabilistic methods for the prediction of TFBSs from multiple alignments of phylogenetically related DNA sequences. The key parameter we had to set is called `bgprior`. Essentially, `bgprior` is the sequence background prior and it should be thought of in the following way: 1-`bgprior` is the prior probability of finding a site in a given sequence. For example, for a `bgprior` of 0.995, we have a 0.005 prior probability of finding a site which results in expecting 5 sites on a 1000 bp long promoter. Since, to the best of our knowledge, no studies showed TFBS frequency on promoters, we leveraged knowledge generated in [14] and fit what we know from human and mouse data. That is, for each motif, we find the `bgprior` value such that the mean posterior probability over all promoters in zebrafish is matching the mean posterior probability in the organism from which the motif is originating. We set lower and upper bound for `bgprior` to 0.6 and 0.999999999. We apply a binary search for `bgprior` until the relative difference between ρ_z and ρ_x is smaller than 0.05, where ρ_z is the mean posterior probability in zebrafish and ρ_x is the mean posterior probability of the organism of the origin of the motif m. In Figure A.2C-D we see the final mean posteriors between zebrafish on the y-axis and mouse and human, respectively, on the x-axis. In Figure A.2E-F, we see the total number of sites predicted, compared between zebrafish (on the y-axis) and mouse and human, respectively (on the x-axis).

With all TFBS predictions generated, we sum the posterior probability for all sites per promoter and apply a binding threshold of 0.5 which essentially means that we are considering a motif binding a given promoter only if the

total posterior probability of that motif is higher than 0.5. With this, we generate a sitecount matrix that allows us to run ISMARA. Furthermore, all TFBS predictions are publicly available at the SwissRegulon database (http://swissregulon.unibas.ch/).

2.4.5 Grouping motifs

In order to reduce the redundancy of motifs in ISMARA analysis, we grouped motifs with the following algorithm

1. for each pair of motifs, we calculate the correlation between the sitecount columns,

2. using single-linkage clustering and threshold of 0.6 we define motif clusters and combine their sitecount columns such that $p_{m_g} = \frac{1}{\sqrt{n}} \sum_i p_{m_i}$ where n is the number of motifs in the group, p_{m_i} is the predicted probability of motif m_i is binding promoter p binding and p_{m_g} is the final probability of motif group m_g binds promoter p.

In the final annotation, we use "_" to mark motif groups, while "+" marks identical TFs biding the same motif.

2.4.6 Pseudobulk ISMARA

Embryos sequenced at the 12 hpf stage were clustered and manually curated into 31 clusters (a total of 2131 single cells). We took the raw UMI counts from the 31 clusters, grouped the counts per annotated clusters, removed Poisson sampling noise[21] and obtained log-transcription quotients which we converted to log transcript per million (log TPM). No pseudo-counts were added.

2.4.7 Comparing motifs

Detailed description of comparing PWMs is given in [16]. Briefly, PWMs are compared "nucleotide-by-nucleotide". That is, for each position in the PWM, we compute linear convolution between two arrays of 4 elements (a number for each of the ACGT nucleotides) which tells us about the independence between the same positions in a given PWM.

2.4.8 Comparing different annotations of TFs with mapped motifs

A set of 2351 TF-motif pairs available at CIS-BP was filtered so that only TFs which a) are in the danRer11 annotation and b) have an available motif mapped (some motifs come from TRANSFAC [164] database which requires commercial license). We found that a significant fraction of motifs in CIS-BP are identical even though they are labeled with different identifiers (Figure A.5B; Section 2.4.7).

Next, using MotEvo we predicted TFBSs on a previously prepared set of promoters and generated the sitecount matrix. The smaller number of motifs resulted in the lower number of predicted targets per motif (Figure A.5D; average 502 to 805 targets per motif predicted using results from our pipeline) and the number of sites per promoter (Figure A.5E; an average of 4 sites in comparison to 10 sites per promoter).

After preparing CIS-BP TF-motif annotation for ISMARA (Section 2.4.5), we ended up with 308 motif groups (compared to 475 motif groups from our analysis). Finally, we ran ISMARA on [88] and compared fraction of variance (FoV) explained (for more details about FoV see [14]). We found that, across all samples, ISMARA predicted higher FoV when using sitecount matrix generated from motifs resulting from our pipeline (Figure A.5C). Total FoV increase is 33% (from 3.6% to 5.4% FoV across the whole dataset).

3

Modelling constitutive promoter expression in *Escherichia coli*

The previous Chapter 2 addresses the question of modeling gene expression in terms of genome-wide regulatory sites. In this chapter, we look more closely into a specific transcription factor and investigate a possibility to model its expression in terms of binding site affinity. Work in this chapter is a joint effort of the following authors:

Đorđe Relić[7,8], Thomas Julou[7,8], Maciej Bak[7,8], Diana Blank[7,8], Tobias Zehnder[7,8,9], Louise Wolf[7,8,10], Dany Chauvin[7,8], Erik van Nimwegen[7,8]

3.1 Introduction

Regulatory sequences differ from coding regions in the its ability to attract transcription factors (TFs). As shown in Chapter 2, DNA-binding domains recognize specific small DNA sequences which mediate transcription and, consequently gene expression. For many TFs the specific DNA sequences they bind (motifs) are well known as well as how presence or absence of such motifs influences gene expression. It is also known that mutations in bind-

[7] Biozentrum, University of Basel, Basel CH
[8] Swiss Institute of Bioinformatics, Basel, CH
[9] Current address: Max Planck Institute for Molecular Genetics, Berlin, DE
[10] Current address: Roche Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, CH

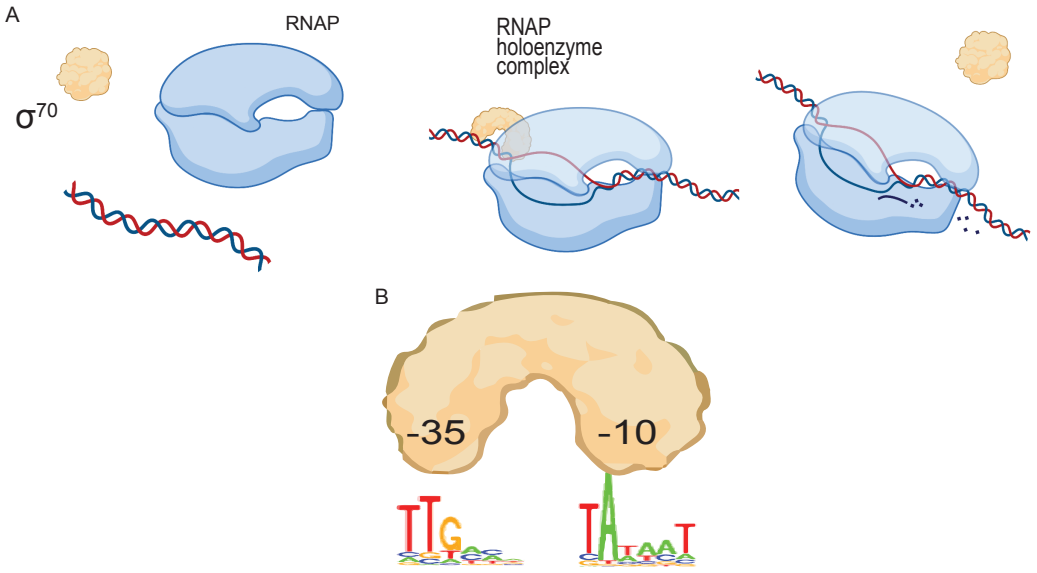


Figure 3.1: σ^{70} **transcription regulation.** (A) σ^{70} binds RNAP and forms the RNAP holoenzyme transcription complex which subsequently binds DNA and initiates transcription. After elongation, σ^{70} is freed from the RNAP holoenzyme complex and can start another transcription initiation process. (B) σ^{70} binds the promoter region in sequence-specific manner with two of its feet -35 and -10 foot.

ing sites can increase or decrease gene expression yield relative to the basal expression[22, 152]. However, it is largely unknown how to predict gene expression levels regulated by a given TF for an arbitrary promoter sequence.

Due to lack of post-transcriptional control (Figure 1.1), bacteria represent a perfect model for studying how differences in TF binding affinity predicts differences in gene expression. Bacterial promoter sequences encode targets for a specific sub-unit of RNAP, the σ factor [26]. The most studied sigma factor is the σ^{70} factor and it was previously shown that the binding affinity of σ^{70} is a predictor of change in transcription levels [90]. Together with RNA polymerase, it forms the RNAP holoenzyme complex[27], and facilitates transcription initiation by binding a specific motif (Figure 3.1). Upon the fixation of RNAP holoenzyme to the promoter sequence, DNA is melted and the open complex is formed. It is the role of σ^{70} to facilitate the bound state during the formation of the open complex. Once the open complex is formed, RNAP escapes the promoter region and initiates start of mRNA synthesis (elongation). After the start of elongation, σ^{70} detaches from the RNAP and is ready to continue facilitating gene transcription initiation by forming new RNAP holoenzyme complexes [153] (Figure 3.1).

σ^{70} binds promoters with its "two feet": the -35 foot with consensus sequence 5'-TTGACA-3' and the -10 foot with consensus sequence 5'-TATAAT-3' with a spacer in between(Figure 3.1B). Even though the spacer sequence has not been characterized in terms of nucleotide composition, it was shown that mutations in the spacer region can influence gene expression [32, 66]. It is also important to note that almost none of the native *E. coli* promoters encode exactly for the described consensus sequence which suggests a certain level of versatility of σ^{70} to bind many different promoters [106].

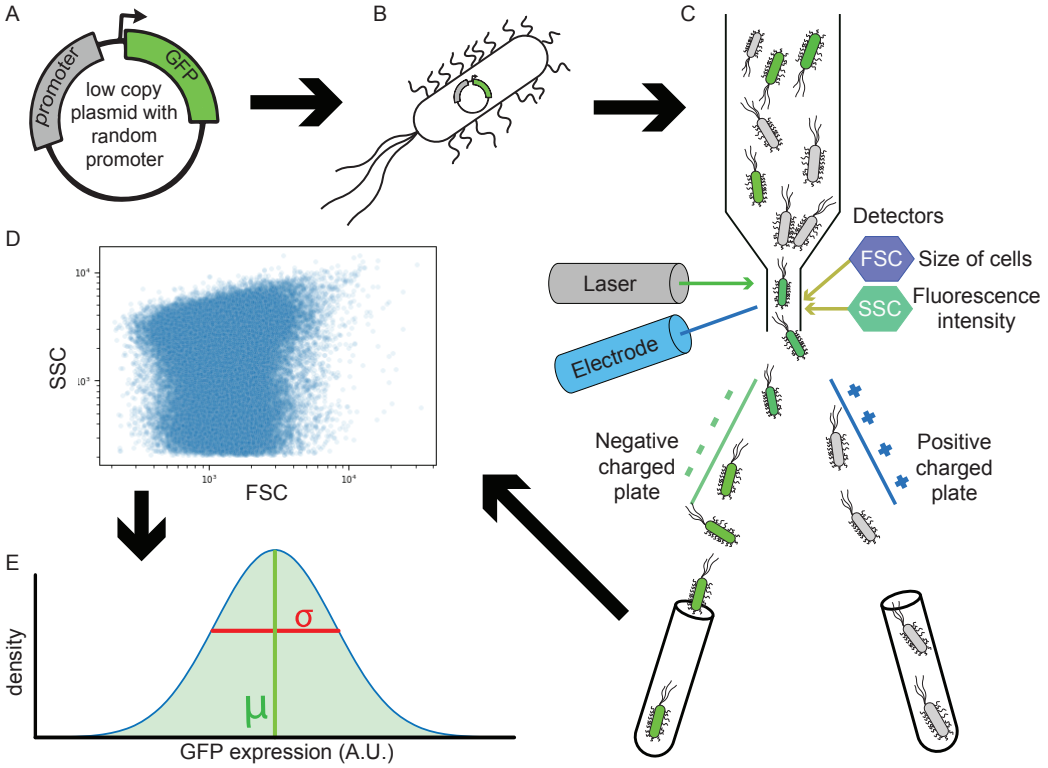


Figure 3.2: **Experimental protocol of gene expression selection in bacteria.** (A) Plasmid vector is designed with a promoter sequence in front of a GFP gene. (B) Plasmid vector is loaded into bacteria. (C) Using FACS, bacteria are sorted based on gene expression of the loaded promoter. (D) FACS measurement results in two values: cell size estimated by laser beam diffraction (FSC) and fluorescence intensity coming from the expression of GFP (SSC). (E) For the population of cells that went through the FACS measurement, we can estimate the mean and variance of gene expression of the whole population.

Studying the relationship between TF binding affinity and gene expression is especially challenging *in vivo* due to the nature of promoters that encode many TF binding sites. TFs compete for similar or overlapping binding sites on native promoters [102] and different promoters can have up to 10 000 fold

difference in expression based on the difference in their composition [108]. However, bacterial systems allow for design of experiments which facilitate evolution of completely random promoters (Figure 3.3A-B) through artificial selection based on gene expression [98, 166] (Figure 3.2).

In this work, we used the data generated in [166], furthermore extended the artificial selection of constitutive promoters, and investigated the hypothesis of modeling gene expression solely based on the σ^{70} binding affinity.

3.2 Results

3.2.1 Characterization of the initial dataset

Wolf et al. [166] conducted an evolutionary experiment that includes 5 rounds of selection, PCR mutagenesis and flow cytometry selection (Figure 3.3A). Promoters are selected for two expression regimes: medium and high expression. Medium expressors are selected based on the reference expression of wild type promoter *gyrB* which has mean expression at 50th percentile of all *E. coli* promoters [170] and high expressors are selected to express in range of wild type promoter *rpmB* which has to mean expression at 97.5th percentile of *E. coli* promoters [138]. In the first round of selection, the expression of one million promoters is measured using Fluorescence-Activated Cell Sorting (FACS) and in each subsequent round, only promoters expressing in a given range are kept for further analysis (Figure 3.3B). The resulting set of promoters does not have any binding sites other than ones targeted by the σ^{70} as shown in Figure B.8 and [165]. We built on this data set and generated expression data of higher resolution (see Appendix B.1).

3.2.2 σ^{70} binding affinity

Thermodynamic models have been proposed earlier as suitable for modeling gene expression regulation in terms of TF binding affinity. Appeal for applying thermodynamics models lies in taking the information of instantaneous gene production and modeling it with a probability framework that looks at changes of RNAP binding rates to the promoter rather than changes in concentration of synthesized protein[17]. That is, all actors in a gene regulatory network are

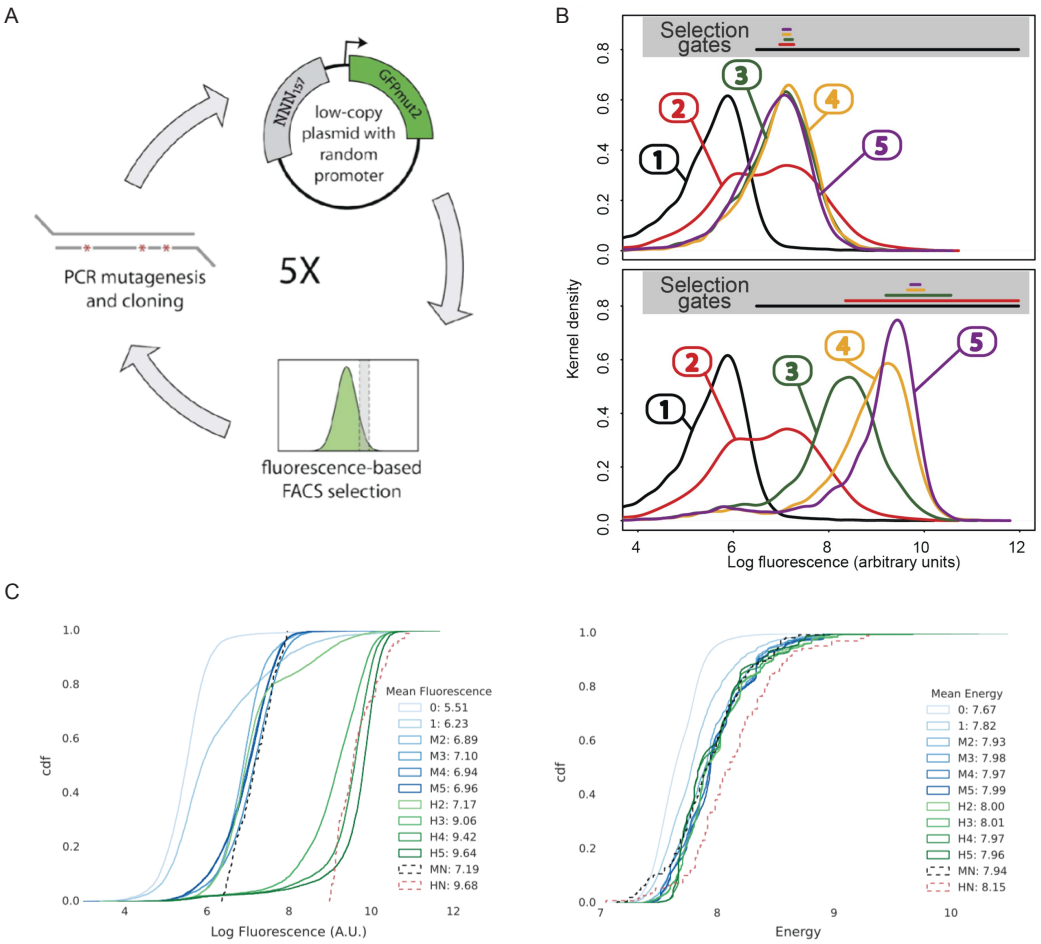


Figure 3.3: **Evolutionary experiment of synthetic promoters.** (A) One million synthetic *E. coli* promoters were loaded into cell culture and selected for expression based on GFP fluorescence. Next, plasmids were extracted and mutated using PCR mutagenesis protocol and then again loaded into bacteria and selected. This process was repeated 5 times. (B) Bacteria were put through an evolutionary experiment in two regimes: medium expression (top) and high expression (bottom). For each of the selection rounds, a given range of expression values was imposed by gating the fluorescence readout. (C) σ^{70} binding energy correlates with expression levels only up to medium expression levels. In the regime of high expression, the sigma factor binding energy does not explain the difference in expression levels.

assigned probabilities of being at the right time at the right place and changes in the probability of binding should drive the change in expression output.

To calculate binding affinity (or binding energy) of σ^{70} we start from a given PWM (in this study we use the PWM initially generated in [89] and then adapted in [22]) and, for each piece of promoter sequence in the length of the PWM, we compute a window score E_w with:

$$E_w = \sum_{i,j} M_{i,j} \tag{3.1}$$

where $M_{i,j}$ corresponds to the log-likelihood of of nucleotide j at position i.

The probability of RNAP binding window w is given by e^{E_w}. Thus, to calculate the binding probability of RNAP to bind the whole promoter sequence, we sum the probabilities of it binding at each position of the promoter.

$$E_{seq} = \log \sum_{w} e^{E_w} \tag{3.2}$$

Furthermore, we implemented an additional level of freedom for σ^{70} in terms of binding differently sized spacers.

$$E_{seq} = E'_{-10} + E_{spacer} + E'_{-35} = \log e^{E_{-10}+e^{E_0}} + E_{spacer} + \log e^{E_{-35}+e^{E_0}} \tag{3.3}$$

where, E_{-10} and E_{-35} are the energies of -10 and -35 foot binding, E_{spacer} is the energy of the spacer and E_0 is the constant which allows -10 and -35 foot to not bind. Allowed spacers span in the length from 15 to 19 base pairs in between the two feet, and, for a given site, we consider all possible spacers.

3.2.3 Promoter features

Having the estimation of mean μ and variance σ^2 for each promoter (Figure 3.4A) and the proposed sequence binding energy model there are several sequence features we can look at. The main question of this exercise is to investigate whether we can use the computed binding affinity of a transcription factor to predict gene expression?

We find that, for a given random promoter sequence, cumulative sequence energy is not sufficient to discriminate between high and medium expressors (Figure 3.4B).

Using the defined model, there are several more features we looked into. Assuming that, instead of cumulative sequence energy, it is the binding site with the highest affinity that determines the transcription levels. In regards

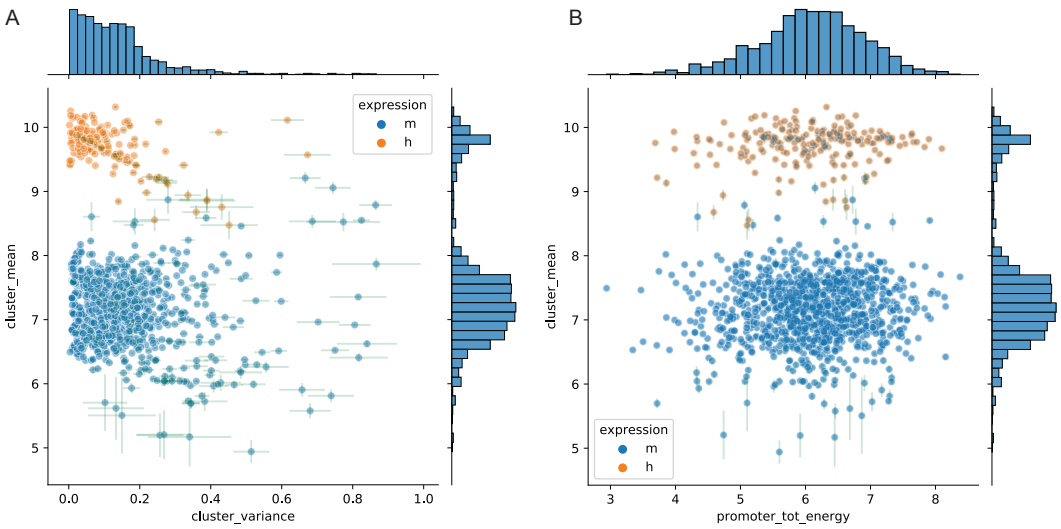


Figure 3.4: σ^{70} **binding affinity does not explain changes in gene expression.** (A) For each promoter (represented as a dot) we estimate the mean and variance of expression together with error bars (in green). Two expression regimes are visible: medium expression regime (in blue) and high expression regime (in orange). (B) The total binding energy of σ^{70} across the whole promoter sequence cannot explain the changes in expression levels between medium and high expressors.

to that hypothesis, we find that the best binding site (exhibiting the highest binding energy) is as well not sufficient for explaining transcriptional changes (Figure B.10A). Relative to the best binding site, we also looked at the distance from the coding region (Figure B.10D). Since starting transcription far from the coding region can mean wasting energy and transcribing irrelevant DNA content, we investigated whether higher expressors exhibit a smaller distance to the coding region. We find no predictive power from the position of the best binding site.

The promoter sequence expression output is known to increase with higher content of AT nucleotides and, specifically, lack of C nucleotides. This is most likely due to the lower required temperature for breaking of weaker AT hydrogen bonds (and easier opening of the DNA), compared to two-fold stronger hydrogen bonds between GC nucleotides[112]. We looked at the nucleotide composition downstream from the strongest binding site (Figure B.10C) and found no clear indication of the difference between the medium and high expressors. Notably, there are a significant fraction of both medium and high expressors that have the best binding site exactly at the promoter end.

In terms of modeling gene expression using the thermodynamics approach,

we also asked the question about the stability of the system. That is, the energy entropy of the sequence can be an indicator of the sequence features which would allow for specific binding preferences. Lower entropy indicates a less disordered system with few strong sites and the rest of the weak sites while higher entropy would indicate a higher disordered system, including sites of similar affinity. Strikingly, there is virtually no difference in entropy of the sequence in terms of binding affinity between medium and high expressors (Figure B.10B).

We also hypothesized that longer promoters might be indicating higher expression due to a higher chance of having a site, however, the length of the promoters is also not a valid predictor (Figure B.10F).

Lastly, we investigated the spacer, initiator, and discriminator sequences of the best binding sites. Spacer sequence, placed between the two feet of σ^{70}, even though not being specific enough to be characterized in known PWMs, is shown to be having an effect on gene expression. Moreover, it is the specific part of the spacer, next to the -10 foot, called discriminator sequence, which is shown to have an effect on gene transcription [32, 66]. However, our study did not yield any discrimination between medium expressors in terms of spacer length or content. Initiator sequence represents the piece of DNA that is transcribed at the start of the elongation[106]. With the start of elongation, it takes RNAP holoenzyme to transcribe 9-11 nucleotides in order to escape the promoter. For that to happen, the open complex needs to be stable enough and it was shown that in the opposite case, abortive transcription happens. That is, in the case of unstable open complex, RNAP transcribes a short mRNA which is released, DNA is fused back and RNAP starts *de novo* forming of open complex [56]. To investigate the bias of the initiator sequence, we stratified our promoter set into high and medium expressors, aligned the sequences based on the best binding site, and looked at the frequency of nucleotides in the initiator and discriminator sequences (Figure B.11). We found that higher expressors (Figure B.11C-D) show a slightly stronger binding preference of the -35 foot. Moreover, a slight preference for the previously reported extended -10 element [46] is identified in higher expressors, however, without significant predictive power. When looking at the initiator sequence, we do find virtually no signal whatsoever.

3.3 Discussion

σ^{70} is one of the most studied transcription factors. Being a housekeeping gene in *E. coli*, crucial for expression due to its role in the formation of RNAP holoenzyme transcription complex, with long defined and established consensus sequence, it seems like a perfect system to study the general productiveness of expression based on binding energy. In this work, we generated a high-resolution expression dataset consisting of random promoters regulated only by σ^{70}.

Following the work of several groups, we investigated promoter sequence features relying on the hypothesis that the binding energy of the σ^{70} alone can be sufficient for predicting the expression level. Results from this study suggest otherwise. That is, for a set of a little over 3000 random synthetic promoters characterized with high-resolution expression data, we could not find any predictive power of the binding affinity model we suggest. Having this in mind, we hypothesize what would be the next direction to continue with the investigation of this phenomenon.

One approach would be the generation of more specific promoters. In our dataset, promoter sequences were up to 150 base pairs long. In theory, that is enough for 3 RNAP holoenzyme complexes to "stack up" the promoter and "queue" for the start of transcription. Having shorter promoter regions would allow for a more specific, differential design of promoter sequences and subsequent expression measurement. Of course, generating a larger dataset of such promoters would allow for an even more systematic investigation of proposed promoter features driving the difference in expression.

All bacteria in our dataset were grown in the same conditions [166], however, there could be some subtle differences that would result in impairment of RNAP to initiate gene transcription. One such example is the non-DNA binding ppGpp which is stress-mediated TF [61]. We could hypothesize that, for some cells, the non-zero concentration of such TFs could introduce the concept of RNAPs molecules in different states - available and not available for σ^{70} mediated transcription initiation.

Lastly, one of the issues might be in the assumption of the thermodynamics model itself. These types of models rely on several assumptions which are critical for their application. The main assumption is the equilibrium of the

system[17]. It is known that bacteria consume energy (ATP) to melt the DNA and initiate transcription, and it is known that his process is not reversible which already means that there is no equilibrium in the gene transcription system.

In conclusion, from previous work, it is evident that tweaking promoter expression relative to a reference value previously measured is possible [22, 98]. However, accurately modeling gene expression in terms of promoter sequence features is still an open question.

4

Conclusion and future outlook

Understanding gene regulation in terms of gene regulatory networks is a central question of molecular biology. In the presented chapters we addressed this question on two levels: 1) by inferring a set of regulatory elements in a given organism and modeling expression in terms of genome-wide regulatory sites and 2) modeling expression of a constitutive promoter in terms of a single TF binding affinity.

In Chapter 2, we hypothesized that transcription factors that encode similar DNA-binding domains bind similar motifs. We confirmed our hypothesis by analyzing curated sets of TFs with mapped motifs in human and mouse. Having confirmed our initial hypothesis, we developed a general pipeline for inferring a gene regulation resource by 1. inferring a set of TFs with mapped motifs in a given, less studied, organism by leveraging its similarity to other, better-studied organisms, 2. annotating a promoter set from transcript data and aligning the promoter regions to related genomes, and 3. predicting genome-wide transcription factor binding sites. To demonstrate our pipeline, we inferred a set of 994 TFs with 552 unique motifs in zebrafish by leveraging its similarity to human and mouse. Furthermore, we annotated a set of 36259 promoters in the zebrafish genome which is aligned to related genomes of common carp, goldfish, and grass carp, and predicted transcription factor binding sites genome-wide. With generated resources, we employed ISMARA[14], a computational tool that models gene expression in terms of genome-wide regulatory sites, and analyzed several RNA-seq data sets and one single-cell RNA-seq dataset. Our analysis resulted in the prediction of known and novel regulators across zebrafish tissue. Notably, our results offer an expansion of the HNF-

family TFs regulatory relationship by predicting interaction between hnf1ba/b and hnf4/b TFs. We also predicted the novel role of zbtb14 in repressing neuronal genes in non-neuronal tissue and the previously unknown activator role of grhl1 in gill and blood cell types. In our analysis of the scRNA-seq dataset, we predicted the expected activator role of myod1 in adaxial cells cluster and, previously unknown, differential transcription regulation role of gata2a and gata3 in epidermis tissue. The next step would be to experimentally test the generated hypotheses about novel regulators in zebrafish gene expression.

Notably, in the work presented here, we focused mainly on the RNA-bulk data set and showed, as proof of concept, that ISMARA can analyze scRNA-seq datasets as well. Future efforts would be put in equipping ISMARA for streamlined analysis of scRNA-seq datasets since predicting regulation in terms of motif activities at a single-cell resolution is of great interest. However, there are challenges when it comes to the analysis of single-cell RNA-seq datasets [86, 132, 156]. Due to the high noise in single-cell datasets, even for highly studied organisms that have the majority of TFs annotated with sequence binding specificities, it is challenging to computationally predict novel regulators driving subtle changes in gene expression state between single cells. scRNA-seq datasets are subject to dropout events, where in the same sample a gene can be detected in moderate or high levels in some cells, but not at all in others. This can lead to the prediction of false negatives in terms of genome-wide regulation. There are, however, many efforts put in towards dealing with different caveats scRNA-seq datasets come with. For example, we now have correct ways to distinguish between technical and biological noise [21] and have more insights into what would be the right information to look for when intuitively thinking of and visualizing high-dimensional expression data[33]. Moreover, it is estimated that for every 3 published scRNA-seq datasets, there are 2 newly published methods [20], which shows that great efforts are put in for finding the right ways to answer biological questions from this exciting experimental output.

While we are still figuring out how to analyze scRNA-seq datasets, the advances are not stopping. We are already seeing single-cell multi-omics data which combine several readouts from a single-cell such as the combination of mRNA-genome, mRNA-DNA methylation, mRNA-chromatin accessibility, and mRNA-protein [95]. By expanding the information of a cell state in terms

of having both measured expression and chromatin state, we are facilitating computational methods to provide more accurate insights about gene regulatory networks in systems spanning from organism development to tissue heterogeneity in various diseases.

In parallel to elucidating genome-wide transcription regulation, understanding how regulatory sequences encode specific sites that in turn regulate different expression levels of the same TFs is of great interest. Understanding this question would allow for the design of tunable gene regulation. In Chapter 3 we zoomed in and investigated gene regulation on the level of a single TF on a set of random synthetic promoters. We focused on the system of σ^{70} mediated transcription initiation in *E. coli*. Using a set of random synthetic promoter sequences artificially evolved for expression regulated solely by the σ^{70}, we investigated modeling of gene expression in terms of σ^{70} binding affinity. The data set consists of over 3000 promoters distributed between two expression regimes: medium and high expressors. We modeled gene expression by assuming a thermodynamic system described by the binding probabilities of the RNAP holoenzyme transcription complex in terms of σ^{70} binding affinity. We showed that, with this approach, on our dataset, we do not find enough predictive power to infer changes in gene expression solely based on σ^{70} binding affinity and related features.

Even in a bacterial system, for a transcription factor that is as studied as σ^{70}, it is still unknown what is the best way to predict gene expression on an arbitrary promoter sequence. The situation, of course, becomes even more complicated when we look into eukaryotic organisms where gene expression gets regulated on more levels. Moreover, in eukaryotes, mRNA content and protein concentration do not necessarily tightly correlate. That is, the amount of mRNA available does not necessarily equal the amount of protein synthesized [24]. Furthermore, some transcription factors need to be activated to regulate gene expression[105], so even having the full protein content measurement of the cell might not provide all the necessary information for understanding gene expression regulation.

Alas, all is not grim. With advances in technologies, no matter how many caveats they come with, we are learning more and more with every new ex-

periment. So much so that, for some questions posed at the introduction of this thesis, we already had an answer 30 years ago - we can reprogram cells! Researchers managed to convert fibroblasts into myoblasts by transfecting the right piece of DNA sequence[43, 148, 150]. And we will most certainly find more ways to convert from one cell type to another. Or at least find pairs of cell types where that is not possible.

A

SI: A pipeline for genome-wide annotation of transcription factors, their sequence specificities, and binding site

A.1 Supplementary tables

Table A.1: List of manually curated Pfam domains.

Name	ID	Description
AATF-Che1	PF13339	Apoptosis antagonizing transcription factor
AKNA	PF12443	AT-hook-containing transcription factor
AlbA_2	PF04326	Putative DNA-binding domain
ARID	PF01388	ARID/BRIGHT DNA binding domain
AT_hook	PF02178	AT hook motif
B3	PF02362	B3 DNA binding domain
BET	PF17035	Bromodomain extra-terminal - transcription regulation
BTD	PF09270	Beta-trefoil DNA-binding domain
bZIP_1	PF00170	bZIP transcription factor
bZIP_Maf	PF03131	bZIP Maf transcription factor
Cdh1_DBD_1	PF18196	Chromodomain helicase DNA-binding domain 1
CENP-B_N	PF04218	CENP-B N-terminal DNA-binding domain
CEP1-DNA_bind	PF09287	CEP-1, DNA binding

Table A.1 continued from previous page

Name	ID	Description
Cep3	PF16846	Centromere DNA-binding protein complex CBF3 subunit B
Ciart	PF15673	Circadian-associated transcriptional repressor
COE1_DBD	PF16422	Transcription factor COE1 DNA-binding domain
CP2	PF04516	CP2 transcription factor
CSD	PF00313	'Cold-shock' DNA-binding domain
CTF_NFI	PF00859	CTF/NF-I family transcription modulation region
CUT	PF02376	CUT domain
CUTL	PF16557	CUT1-like DNA-binding domain of SATB
DBD_Tnp_Hermes	PF10683	Hermes transposase DNA-binding domain
DBINO	PF13892	DNA-binding domain
DM	PF00751	DM DNA binding domain
DMRT-like	PF15791	Doublesex-and mab-3-related transcription factor C1 and C2
Dmrt1	PF12374	Double-sex mab3 related transcription factor 1
E2F_CC-MB	PF16421	E2F transcription factor CC-MB domain
E2F_TDP	PF02319	E2F/DP family winged-helix DNA-binding domain
EAF	PF09816	RNA polymerase II transcription elongation factor
EBV-NA1	PF02905	Epstein Barr virus nuclear antigen-1, DNA-binding domain
efThoc1	PF11957	THO complex subunit 1 transcription elongation factor
eIF-5a	PF01287	Eukaryotic elongation factor 5A hypusine, DNA-binding OB fold
EKLF_TAD1	PF16832	Erythroid krueppel-like transcription factor, transactivation 1
EKLF_TAD2	PF16833	Erythroid krueppel-like transcription factor, transactivation 2
Elongin_A	PF06881	RNA polymerase II transcription factor SIII (Elongin) subunit A
EST1_DNA_bind	PF10373	Est1 DNA/RNA binding domain
Ets	PF00178	Ets-domain
ETS_PEA3_N	PF04621	PEA3 subfamily ETS-domain transcription factor N terminal domain
Filament_head	PF04732	Intermediate filament head (DNA binding) region
FLYWCH	PF04500	FLYWCH zinc finger domain
Forkhead	PF00250	Forkhead domain
FYVE	PF01363	FYVE zinc finger
GAGA_bind	PF06217	GAGA binding protein-like family
GATA	PF00320	GATA zinc finger
GATA-N	PF05349	GATA-type transcription activator, N-terminal

Table A.1 continued from previous page

Name	ID	Description
GCFC	PF07842	GC-rich sequence DNA-binding factor-like protein
GCM	PF03615	GCM motif protein
GTF2I	PF02946	GTF2I-like repeat
HLH	PF00010	Helix-loop-helix DNA-binding domain
HMG_box	PF00505	HMG (high mobility group) box
Homeobox_KN	PF05920	Homeobox KN domain
Homeodomain	PF00046	Homeodomain
HPD	PF05044	Homeo-prospero domain
HSF_DNA-bind	PF00447	HSF-type DNA-binding
IRF	PF00605	Interferon regulatory factor transcription factor
Jun	PF03957	Jun-like transcription factor
Kin17_mid	PF10357	Domain of Kin17 curved DNA-binding protein
LAG1-DNAbind	PF09271	LAG1, DNA binding
Lbh	PF15317	Cardiac transcription factor regulator, Developmental protein
MADF_DNA_bdg	PF10545	Alcohol dehydrogenase transcription factor Myb/SANT-like
MBD	PF01429	Methyl-CpG binding domain
Med1	PF10744	Mediator of RNA polymerase II transcription subunit 1
Med19	PF10278	Mediator of RNA pol II transcription subunit 19
Med8	PF10232	Mediator of RNA polymerase II transcription complex subunit 8
Med9	PF07544	RNA polymerase II transcription mediator complex subunit 9
MH1	PF03165	MH1 domain
MMS19_C	PF12460	RNAPII transcription regulator C-terminal
MMS19_N	PF14500	Dos2-interacting transcription regulator of RNA-Pol-II
mTERF	PF02536	mTERF
Myb_DNA-binding	PF00249	Myb-like DNA-binding domain
Myb_DNA-bind_3	PF12776	Myb/SANT-like DNA-binding domain
Myb_DNA-bind_4	PF13837	Myb/SANT-like DNA-binding domain
Myb_DNA-bind_5	PF13873	Myb/SANT-like DNA-binding domain
Myb_DNA-bind_6	PF13921	Myb-like DNA-binding domain
Myb_DNA-bind_7	PF15963	Myb DNA-binding like
MYT1	PF08474	Myelin transcription factor 1
NCU-G1	PF15065	Lysosomal transcription factor, NCU-G1
Neuro_bHLH	PF12533	Neuronal helix-loop-helix transcription factor
Nrf1_DNA-bind	PF10491	NLS-binding and DNA-binding and dimerisation domains of Nrf1

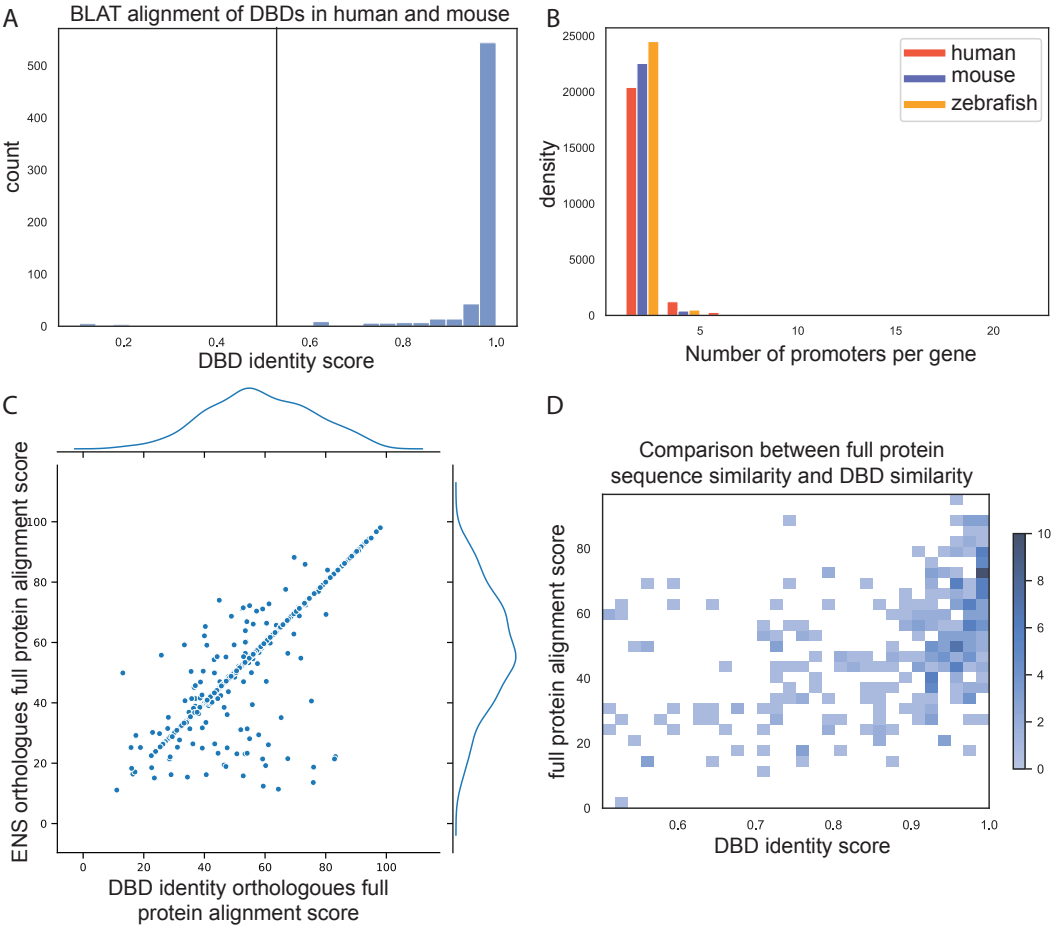
Table A.1 continued from previous page

Name	ID	Description
P53	PF00870	P53 DNA-binding domain
PAX	PF00292	'Paired box' domain
PheRS_DBD1	PF18552	PheRS DNA binding domain 1
PheRS_DBD2	PF18554	PheRS DNA binding domain 2
PheRS_DBD3	PF18553	PheRS DNA binding domain 3
Phospho_p8	PF10195	DNA-binding nuclear phosphoprotein p8
Phtf-FEM1B_bdg	PF12129	Male germ-cell putative homeodomain transcription factor
POT1	PF02765	Telomeric single stranded DNA binding POT1/CDC13
Pou	PF00157	Pou domain - N-terminal to homeobox domain
POU2F1_C	PF19536	POU domain, class 2, transcription factor 1 C-terminal
PSK_trans_fac	PF07704	Rv0623-like transcription factor
RFX1_trans_act	PF04589	RFX1 transcription activation region
RFX5_DNA_bdg	PF14621	RFX5 DNA-binding domain
RFX_DNA_binding	PF02257	RFX DNA-binding domain
RHD_DNA_bind	PF00554	Rel homology DNA-binding domain
RLL	PF10036	RNA transcription, translation and transport factor protein
Runt	PF00853	Runt domain
SAC3	PF12209	Leucine permease transcriptional regulator helical domain
SAND	PF01342	SAND domain
Sox17_18_mid	PF12067	Sox 17/18 central domain
SOXp	PF12336	SOX transcription factor
Spt5-NGN	PF03439	Early transcription elongation factor of RNA pol II, NGN section
Spt5_N	PF11942	Spt5 transcription elongation factor, acidic N-terminal
SRF-TF	PF00319	SRF-type transcription factor (DNA-binding and dimerisation domain)
STAT2_C	PF12188	Signal transducer and activator of transcription 2 C terminal
STAT_bind	PF02864	STAT protein, DNA binding domain
T-box	PF00907	T-box
T-box_assoc	PF16176	T-box transcription factor-associated
TBP	PF00352	Transcription factor TFIID (or TATA-binding protein, TBP)
TBX	PF12598	T-box transcription factor

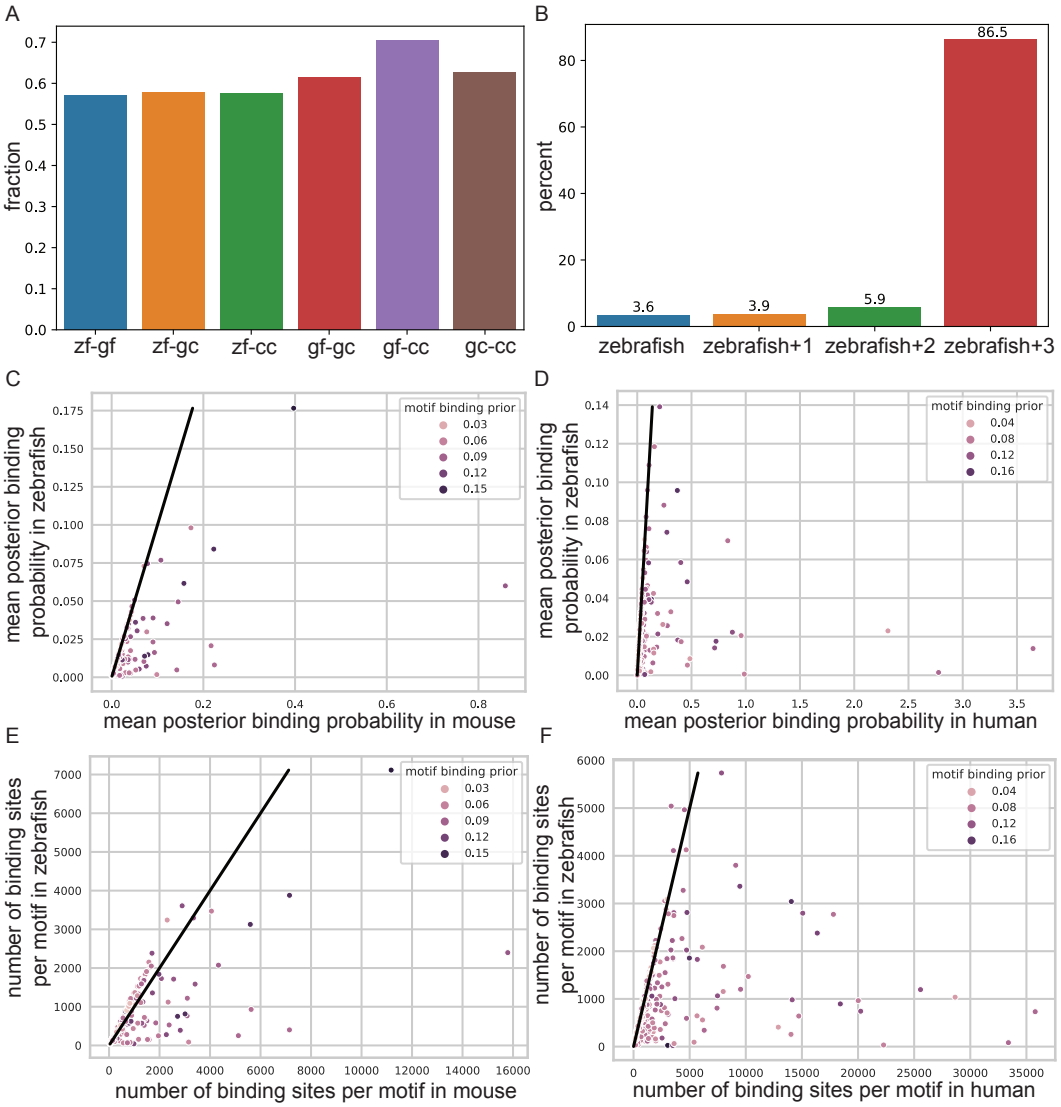
Table A.1 continued from previous page

Name	ID	Description
TCR	PF03638	Tesmin/TSO1-like CXC domain, cysteine-rich domain
TDP43_N	PF18694	Transactive response DNA-binding protein N-terminal domain
TFIIE-A_C	PF11521	C-terminal general transcription factor TFIIE alpha
TF_AP-2	PF03299	Transcription factor AP-2
TF_Otx	PF03529	Otx1 transcription factor
THAP	PF05485	THAP domain
TMF_DNA_bd	PF12329	TATA element modulatory factor 1 DNA binding
Tnp_DNA_bind	PF14706	Transposase DNA-binding
Topoisom_I_N	PF02919	Eukaryotic DNA topoisomerase I, DNA binding fragment
TRAUB	PF08164	Apoptosis-antagonizing transcription factor, C-terminal
Vert_HS_TF	PF06546	Vertebrate heat shock transcription factor
Vert_IL3-reg_TF	PF06529	Vertebrate interleukin-3 regulated transcription factor
Yippee-Mis18	PF03226	Yippee zinc-binding/DNA-binding /Mis18, centromere assembly
zf-BED	PF02892	BED zinc finger
zf-C2H2	PF00096	Zinc finger, C2H2 type
zf-C2HC	PF01530	Zinc finger, C2HC type
zf-C3Hc3H	PF13891	Potential DNA-binding domain
zf-C4	PF00105	Zinc finger, C4 type (two domains)
zf-CCCH	PF00642	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
zf-CSL	PF05207	CSL zinc finger
zf-CXXC	PF02008	CXXC zinc finger domain
zf-FCS	PF06467	MYM-type Zinc finger with FCS sequence motif
zf-H3C2	PF16721	Zinc-finger like, probable DNA-binding
zf-LYAR	PF08790	LYAR-type C2HC zinc finger
zf-NF-X1	PF01422	NF-X1 type zinc finger
zf-TFIIIC	PF12660	Putative zinc-finger of transcription factor IIIC complex
Zfx_Zfy_act	PF04704	Zfx / Zfy transcription activation region

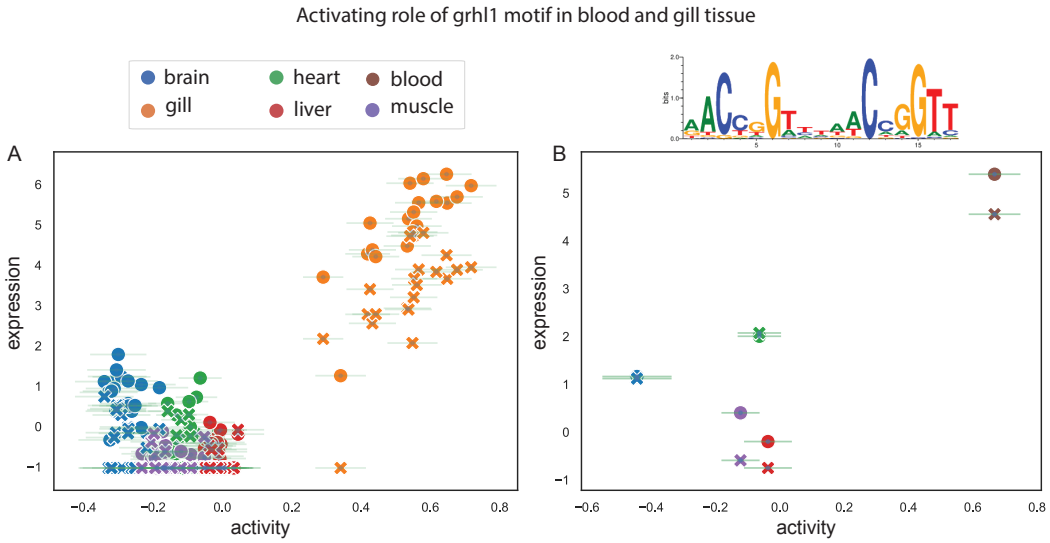
A.2 Supplementary figures



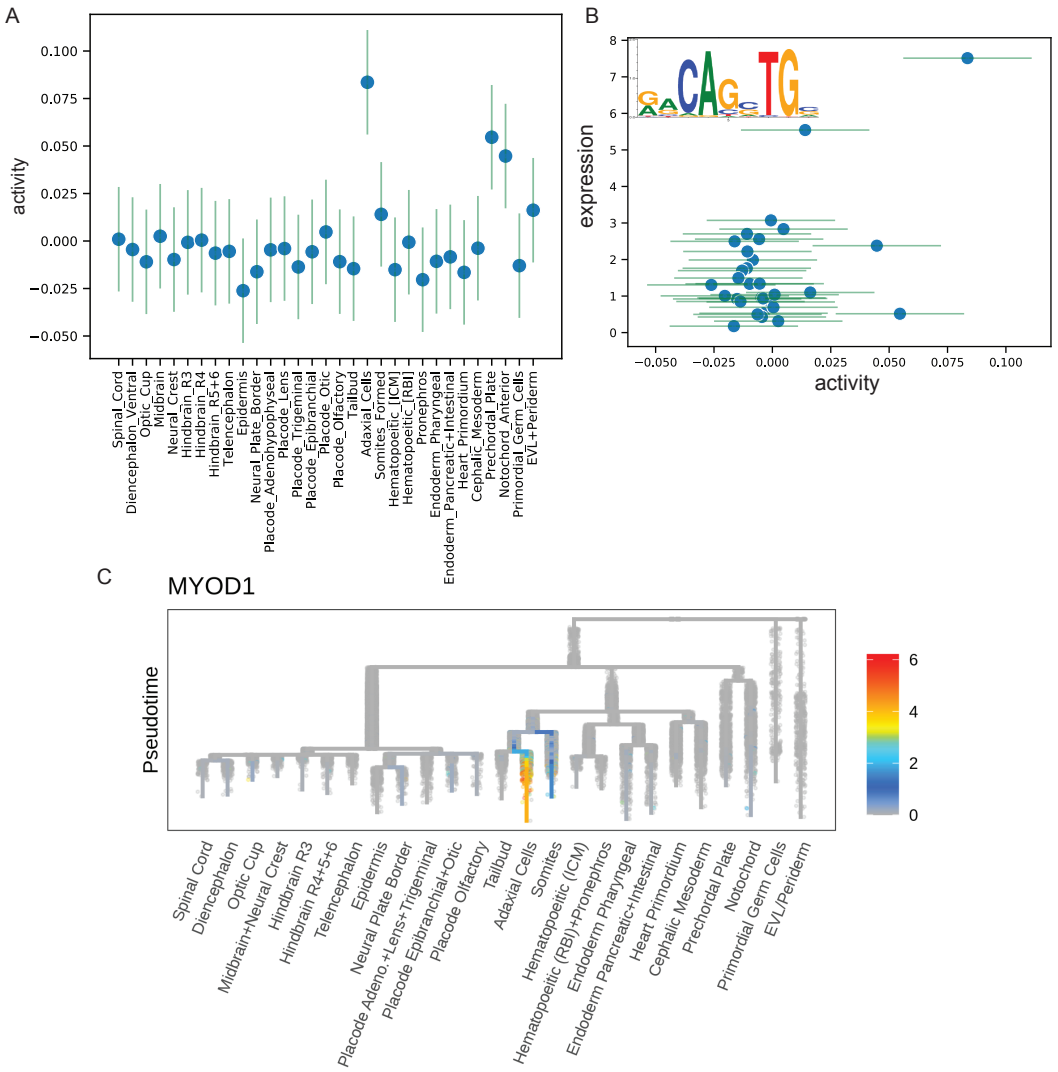
Supplementary Figure A.1: **DBD similarity correlates with motif similarity** (A) BLAT alignment of concatenated DBDs between human and mouse TFs with annotated binding sites retrieved from SWISSREGULON [119]. (B) Comparison of motif similarity and DBD identity score between TFs with annotated binding site information between human and mouse. (C) Comparison of needle alignment identity score between homologous genes annotated by Ensembl and TFs mapped by DBD identity score. (D) For 321 TFs which do not have an homologous genes in Ensembl annotation, comparison of DBD identity score and full sequence needle alignment identity score.



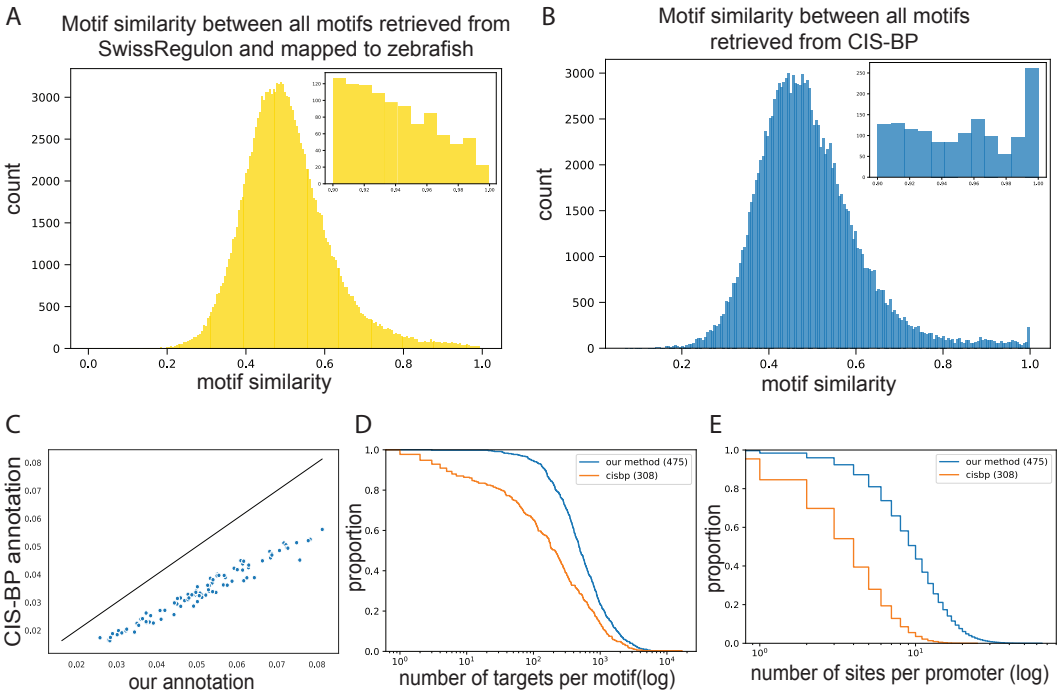
Supplementary Figure A.2: **Species alignment and transcription factor binding site distribution matching.** (A) Conservation information between each two pair of fish calculated from whole-genome alignments ("zf" - Zebrafish, "gc" - Grass Carp, "cc" - Common Carp, "gf" - Goldfish). (B) Percentage of conserved regions between zebrafish and n other fish. (C-F) Each dot represents a motif. Color scale represents the TFBS prior set in the MotEvo (for more details see Methods and Materials 2.4.4). (C-D) Comparison of mean posterior probability binding per motif in zebrafish (y-axis), mouse (x-axis left), and human (x-axis right). (E-F) Number of targets predicted per motif in zebrafish (y-axis), mouse (x-axis left), and human (x-axis right).



Supplementary Figure A.3: **Grlh1 targets are up-regulated in gill and blood cell types.** (A and C) Motif activity of grhl1 is up in gill (A) and blood (C) tissue in comparison with other tissues. (B and D) Grhl1 expression is positively correlated with its motif activity which indicates that grhl1 is likely performing and activator role in gill and blood tissue.



Supplementary Figure A.4: **Myod1 targets are up regulated in adaxial cells.** (A) Motif activity profile of myod1 indicates up-regulation of myod1 targets in adaxial cell. (B) Myod1 expression is positively correlated with its motif activity which indicates that myod1 is an activator. (C) Myod1 is almost exclusively expressed in cells leading to development of adaxial cells.



Supplementary Figure A.5: **CIS-BP motifs.** (A) Pairwise motif similarity between motifs used in our study. (B) Pairwise motif similarity betewen motifs retrieved from CIS-BP. (C) Comparison of Fraction of explained Variance (FoV) between 2 ISMARA runs: using set of motifs inferred in our study (x-axis) and using set of motifs retrieved from CIS-BP (y-axis). (D) Comparison of distribution of number of targets between TF-motif set from CIS-BP and TF-motif mapping from our method. (E) Same as (D) just with distribution of number of sites per promoter.

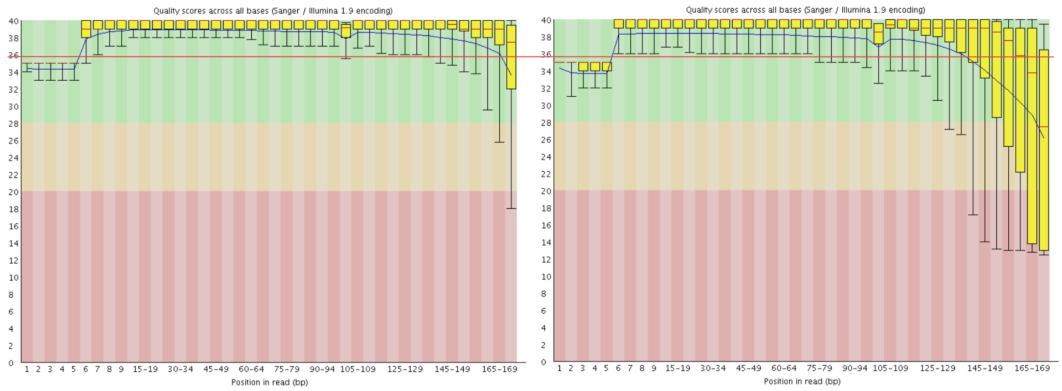
B

SI: Modelling constitutive promoter expression in *Escherichia coli*

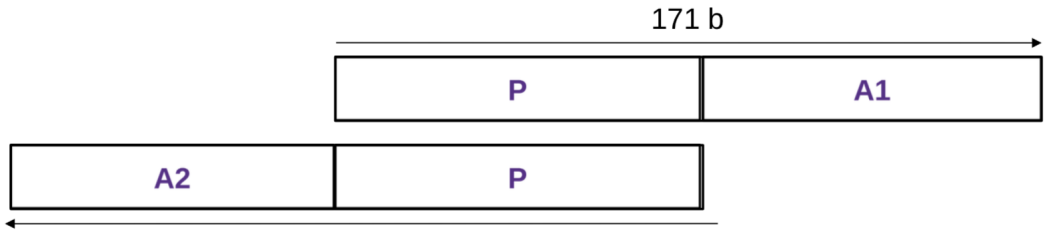
B.1 Acquiring high resolution expression data

Promoter expression characterization coming from [166] offered broad distributions of expression levels for each promoter. To increase the resolution of the gene expression measurements, we performed additional experiments where for each round and each of the expression regimes, promoters were additionally selected into 8 expression ranges (Figure B.9). Furthermore, promoter sequences were re-sequenced using NGS which allowed for higher resolution of the expression library.

From resulting sequences, as we know they are coming from an artificially induced evolutionary experiment, we generated sequence lineages we call clusters. That is, we clustered sequences based on similarity and essentially generated lineage clusters spanning from 1st (initial) round to 5th (final) round (detailed processing can be seen in Appendix B.2). The resulting dataset included 97462 cluster lineages.



Supplementary Figure B.1: Forward and reverse read quality coming from the NGS enhanced experiment.



Supplementary Figure B.2: Illustration of adapters ligated to promoter sequence. Where A1 is the forward adapter and A2 is the reverse adapter.

B.2 Generating sequence lineages

From the NGS experiment, we got 2 x 100 sequencing libraries of paired-end reads. Each sequence in a .fastq file is 171 base pairs long and, next to the promoter sequence, contains the adapter sequence and a plasmid complementary sequence (Figure B.2). Quality of reads is presented in Figure B.1.

To clean and prepare read count data for further analysis, we had to apply the following steps:

1. **Adapter trimming** - detecting the adapter sequences and cutting them out
2. **Merging** - taking forward and reverse reads and merging them into one single read by aligning one with the reverse complement of the other
3. **Clustering** - organizing promoters into clusters which, once observed through rounds, should represent a lineage of promoters

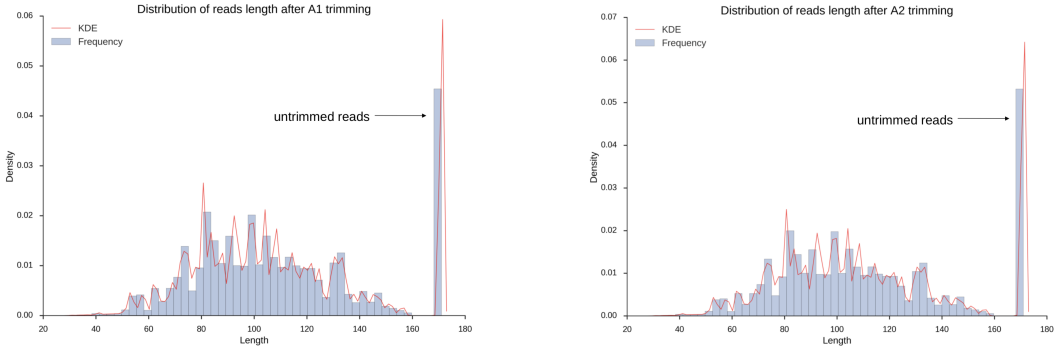
Adapter cutting

Adapter cutting is the processing of forwarding and reverse reads which result in cutting out the primer part of the sequence. These parts of sequences are identical for all promoters and do not hold any biological information. In this case, for defining the adapter sequence, we also took into consideration the plasmid complementary part. Adapters are defined as follows:

a1 = GGATCCTCTGGATGTAAGAAGGAGCTGTCTCTTATACACATC
TCCGAGCCCACGAGACNNNNNNNNATCTCGTATGCCGTCTTCT

a2 = CTCGAGGTGAAGACGAAAGGGCCTGTCTCTTATACACATCTG
ACGCTGCCGACGANNNNNNNNGTGTAGATCTCGGTGGTCGCCG

where *a1* is the adapter of the forward read and *a2* is the adapter of the reverse read.

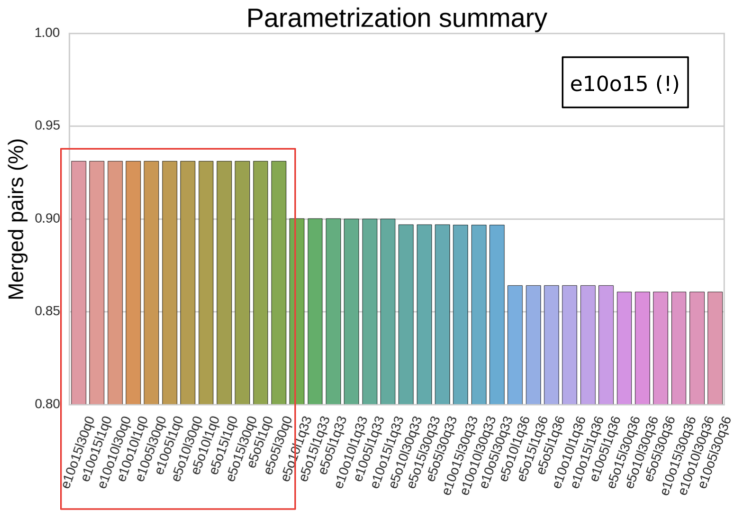


Supplementary Figure B.3: **Adapter trimming.** Statistics on adapter trimming of forward reads (left) and reverse reads (right).

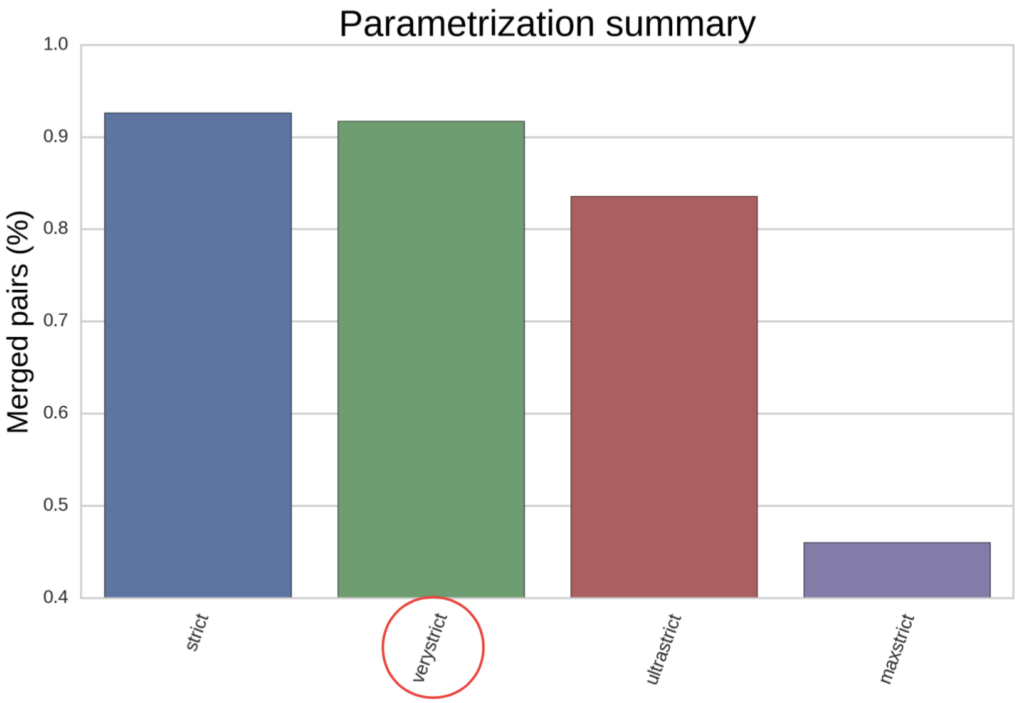
The tool used for adapter cutting is cutadapt[103]. Part of the adapter *NNNNNNN* is defined as "wildcard characters". The wildcard character N is useful for trimming adapters with an embedded variable barcode. Parameters that were used with are: error rate = 0.10 and overlap = 15. These settings mean that the if adapter length is 86, as it is in our case if there is a match of at least 13.5 base pairs, an adapter (or its part) will be cut.

Sequence merging

Software used for merging is BBmerge[29]. Several parameters were tested and one picked was *very strict*. In Figure B.5 are shown all different parameters



Supplementary Figure B.4: **Cutadapt parameters parameter setup.**



Supplementary Figure B.5: **BBmerge parameter setup**

tested and the number of reads that were merged as a result of them.

Clustering

After adapter trimming and pair-end read merging we are left with 100 .fasta files with promoter sequences. Next, we collapsed these files into 1

bigger file, while keeping the unique sequence IDs. Then, promoters were partitioned into a cluster to select groups of sequences that originate from the same ancestor (i.e. products of the synthetic evolution). The clustering assumes that a sequence is always more similar to its ancestor/its mutants than to any other ancestor or mutant of any other ancestor.

In general idea behind the algorithm of clustering may be presented as follows:

1. Having the set of all sequences S: pre-cluster all the sequences into a set of disjoint sets: $C = \{c_i | i \in I\}$ and $\bigcup C = S$; c_i is an over-clustered group that might require further partitioning

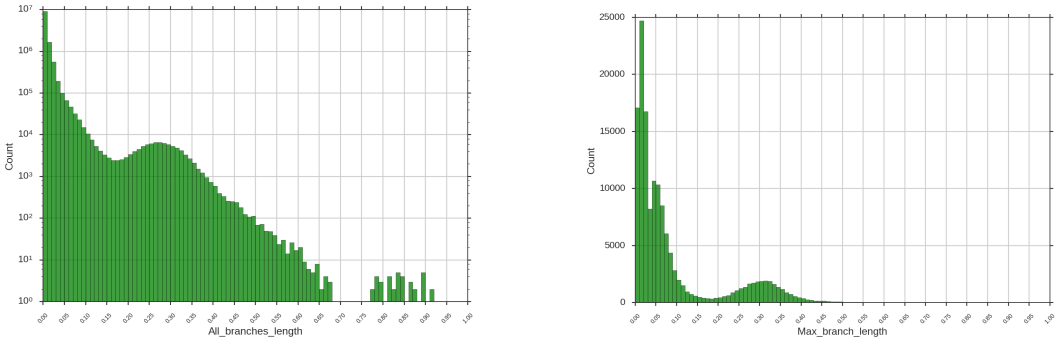
2. Iterate over I and for every pre-cluster c_i create a multiple sequence alignment tree; now we have $T = \{t_i | i \in I\}$ as this step calculates alignment scores and reshapes the data into a new structure: $t_i = MSA(c_i)$

3. Iterate over I and for every MSA tree t_i, if it is necessary, divide the tree into a number of sub-trees. For every tree we have extract only the partitioning as we do not need the tree structure anymore. After such adjustment operation f we obtain final clusters: $\overline{C} = \bigcup\{f(t_i) | i \in I\}$

The set of all distinct promoters (i.e. identical sequences collapsed into a capture number in the ID) is way too big to cluster at once, especially if we care to get precise and accurate results. Firstly, we pre-clustered the sequences to reduce the computational complexity of the problem. We used CD-HIT[54] to obtain initial clusters. CD-HIT employs greedy incremental clustering. Briefly, sequences are first sorted by length. The longest sequence is selected as the representative of the cluster. Next, each of the remaining sequences is compared to the representative sequence. If it falls below the threshold, it is added to the cluster, otherwise, a new cluster is formed.

The parameters for CD-HIT were chosen such that we will not introduce any unnecessary divisions between loosely related promoters at this point. The global sequence identity (GSI) was set to 0.8. We iterated over the clusters and for every set of sequences we constructed a multiple sequence alignment tree. Cases with a single distinct sequence in a cluster were omitted since these could be already considered as final clusters. For the multiple sequence alignment, we used MAFFT[80].

Having all clusters with more than one distinct sequence reformatted into MSA trees we refined the initial clusters into a set of final clusters. Maximum branch length was set such that above it a tree should be divided into two sub-trees. Trimming is iterative until there is no branch above the length threshold (for a given cluster). To select the cutoff value we inspected the distributions of lengths of all branches across all trees as well as only the lengths of the longest branches per tree:



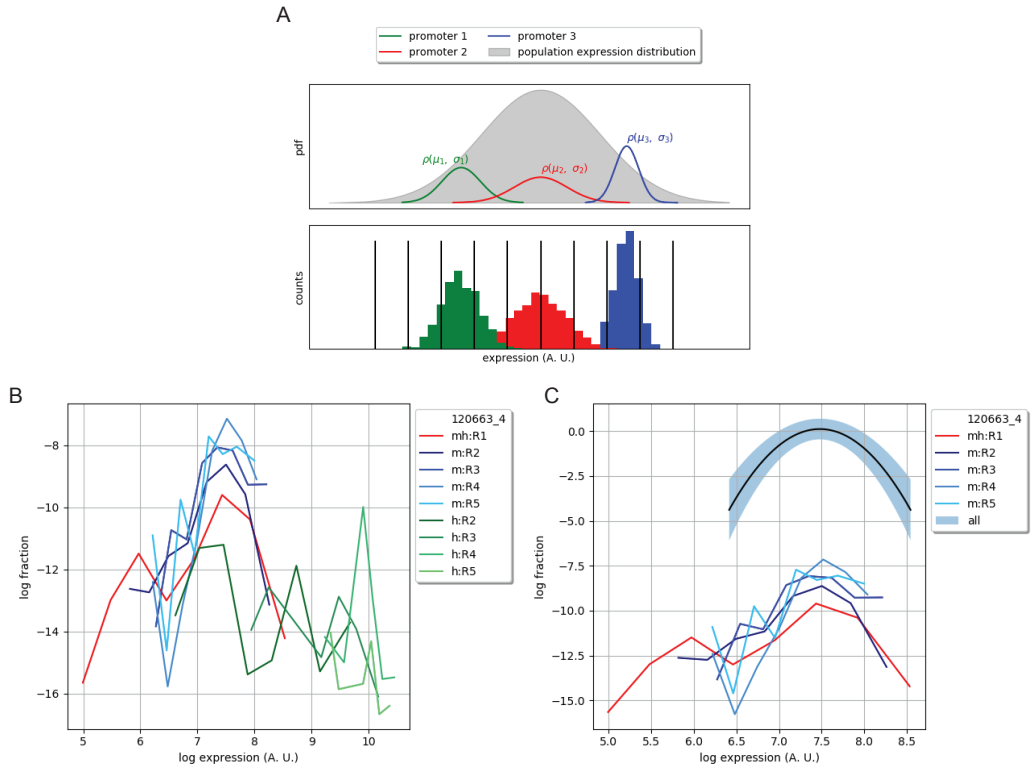
Supplementary Figure B.6: **Branch cutting in phylogenetic clustering of sequences.** Distribution of branch length before cutting (A) and after selecting for longest branches (B).

Tree branching cutoff value to 0.17, as it is a valley in both distributions (Figure B.6). It suggests that it is the best choice for a value that would separate clusters of sequences that truly originate from a common ancestor from promoters that share a substantial fraction of similarity, yet come from distinct ancestors. Given that we selected a cutoff for maximum branch length in an alignment tree we could iteratively trim all the trees. We have divided the trees into smaller sub-trees, for every one of the sub-trees we stored sequence IDs into a separate text file and recovered the .fasta entries that corresponded to that IDs. In the special case when a tree did not require division all the IDs were saved into one file. Combining these clusters with the ones with only a single distinct sequence (filtered at the beginning) resulted in the final clusters.

As a result, we have in total 97462 clusters which essentially represent lineages present in sequence evolution.

B.2.1 Estimating promoter expression mean and variance

With the newly generated dataset, we were allowed to statistically infer expression for each of the clusters. The initial hypothesis was that promoter expression falls into a Gaussian distribution (Figure B.7A). To test our hy-



Supplementary Figure B.7: **Enhanced experiment for gene expression selection.** (A) Each of the 5 evolutionary rounds is segmented into 8 expression bins. The hypothesis is that each of the promoters will have a normal distribution of expression across neighboring expression gates. Empirical analysis of resulting expression values shows that indeed, the medium expression cluster in figure (B) shows a parabolic shape in log space which implies that it follows a Gaussian distribution. (C) We infer the mean and variance of expression for each of the promoter clusters, together with error bars.

pothesis, we started from the initial model which, for a promoter of choice we have the expression distribution $\widetilde{P}$ defined as:

$$\widetilde{P} = f_b^r \cdot q^r \tag{B.1}$$

where f_b^r is the probability that a given promoter p is in round r and bin b, and q^r is the fraction of promoter p in round r before selection. Now, since we assume that in each round, the probability for a promoter to be selected is constant, we know that this distribution is proportional to f_b^r. That is, we can empirically calculate f_b^r with

$$\widetilde{P} = f_b^r \cdot q^r \propto f_b^r = \frac{n_b^r}{N_b^r} F_b \tag{B.2}$$

where n_b^r is the number of reads of given promoter in bin b and round r, N_b^r is total number of reads of all promoters in in b and round r and F_b is the fraction of the whole population that was selected in bin b. F_b we can estimate from FACS measurements.

After examining the initial results (Figure B.7B), we concluded that the promoter expression has a Gaussian distribution form centered around on μ and variance σ^2, $P(x|\mu, \sigma^2)$:

$$P(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{B.3}$$

We define $\rho_b^r(x)$ as the probability that a cell that expresses at level x is selected in round r, bin b. We assume that the selection probability function has a Gaussian form, centered on μ_{br} (the middle of the bin) and with width τ_{br}^2 (the width of the bin). That is, we assume

$$\rho_b^r(x) = \exp\left(-\frac{(x-\mu_{br})^2}{2\tau_{br}^2}\right). \tag{B.4}$$

Furthermore, we define $\rho_b^r(\mu, \sigma^2)$ as the fraction of cells with a promoters whose expression distribution has mean μ and variance σ^2 that are selected in round r bin b. We have:

$$\rho_b^r(\mu, \sigma^2) = \int \rho_b^r(x) P(x|\mu, \sigma^2) dx \tag{B.5}$$

The integral can be performed analytically (details in Appendix B.3) and we obtain:

$$\rho_b^r(\mu, \sigma^2) = \sqrt{\frac{\tau_{br}^2}{\tau_{br}^2 + \sigma^2}} \exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2 + \sigma^2)}\right). \tag{B.6}$$

We note ρ_b^r as the average probability to be selected in bin b in round r, averaged over all cells in the population in round r. Note that this can be written as

$$\rho_b^r = \int_{-\infty}^{\infty} \rho_b^r(x) P^r(x) dx \tag{B.7}$$

where $P^r(x)$ is the expression distribution in the population in round r and we read out the values from the FCS measurement files.

Having $P^r(x)$ being a discrete set of values $(x_i, c_i) \in X, i \in \{1, ..., n\}$ where x_i is the measured GFP expression and c_i is the number of times measuring expression x_i occurred, we in fact have ρ_b^r in the following form:

$$\rho_b^r = \sum_{i=1}^{n} \rho_b^r(x_i) P^r(x_i) = \sum_{i=1}^{n} \rho_b^r(x_i) \frac{c_i}{\sum_j c_j} \tag{B.8}$$

Note that with this notation, the probability $\hat{P}$ that a randomly selected cell from bin b round r comes from our promoter is given by

$$\hat{P} = \frac{\rho_b^r(\mu, \sigma^2) q^r}{\rho_b^r} \tag{B.9}$$

Now, we have that the probability distribution for selection of a random promoter to bin b in round r for mean μ and variance σ^2 given for our promoter according for our data is Binomial distribution:

$$P(D|\mu, \sigma^2) = \binom{N_b^r}{n_b^r} \left(\hat{P}\right)^{n_b^r} \left(1 - \hat{P}\right)^{N_b^r - n_b^r} \tag{B.10}$$

Since we have that the probability $\hat{P}$ takes small values in the limit and the number of reads is large, we can define the probability of the data (i.e. all counts for our promoter of interest) as given by a product of Poisson sampling distributions:

$$P(D|\mu, \sigma^2) = \prod_{r,b} \frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu, \sigma^2) q^r}{\rho_b^r} \right)^{n_b^r} \exp\left(-N_b^r \frac{\rho_b^r(\mu, \sigma^2) q^r}{\rho_b^r} \right). \tag{B.11}$$

To make further progress, we integrate over the unknown fractions q^r. Using a scale-prior of the form $1/q^r$ we obtain (Equation (B.14)):

$$P(D|\mu, \sigma^2) \propto \prod_{b,r} \frac{1}{n_b^r!} \left[\left(N_b^r \frac{\rho_b^r(\mu, \sigma^2)}{\rho_b^r} \right)^{n_b^r} \left(\sum_b N_b^r \frac{\rho_b^r(\mu, \sigma^2)}{\rho_b^r} \right)^{-n_r} \right]. \tag{B.12}$$

For further calculations, the most useful form is the log-likelihood. Keeping only the terms depending on μ and σ^2, we find (Equation (B.14))

$$L(D|\mu,\sigma^2) = \sum_{b,r} n_b^r \log\left[\rho_b^r(\mu,\sigma^2)\right] - \sum_r n^r \log\left[\sum_b N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right]. \quad \text{(B.13)}$$

Fitting expression data for a given promoter p with Equation (B.13) allows us to estimate mean μ and variance σ^2 of promoter expression including error estimation on both values (Figure B.7C). We applied the Expectation-Maximization algorithm and processed 97462 promoter clusters. For downstream analysis we selected promoters which have:

- 1000 reads in R3 or R4 or R5 for medium expressors
- 1000 reads in R4 or R5 for high expressors

3228 clusters were kept as fit for further analysis. Selected 3228 clusters to contain 86.6% of all the reads in the data set. 165 clusters satisfy the conditions in both medium and high expression regimes and those were considered separately. Out of 3228 clusters, there are 2204 medium expressor clusters and 1024 high expressor clusters. Values inferred for mean and variance were filtered so that they meet the following requirements: mean $\in [4.868, 10.537]$, mean error $\in [0.0, 1.0]$, variance $\in [0, 1.0]$ and variance error $\in [0.0, 0.5]$. The lower bound for the mean is set to be equal to the lowest value of all gate interval bounds in the FACS sorting experiments and, similarly, the high bound is set to be the highest value of all gate interval bounds. Other cutoffs seemed reasonable to set. That left us with 2969/97462 (2.9%) clusters which contain 81.2% of reads. Out of 2969 clusters, we have 2048 clusters from medium expression regime (92.9% from fit medium clusters) and 821 clusters from high expression regime (80.2% from fit high clusters).

B.3 Detailed Calculations

Inferring $\rho_b^r(\mu,\sigma^2)$ (Equation (B.5) to Equation (B.6))

$\rho_b^r(\mu,\sigma^2)$ is the fraction of cells with a promoters whose expression distribution has mean μ and variance σ^2 and are selected in round r bin b.

$$\rho_b^r(\mu,\sigma^2) = \int_{-\infty}^{\infty} \rho_b^r(x) P(x|\mu,\sigma^2) dx \tag{B.14}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(x-\mu_{br})^2}{2\tau_{br}^2}\right) dx$$

$$= \Big/ \text{substitute variables } 1 : p = \frac{1}{\tau_{br}^2}, q = \frac{1}{\sigma^2} \Big/$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[p(x-\mu_{br})^2 + q(x-\mu)^2\right]\right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[(p+q)x^2 - 2x(p\mu_{br}+q\mu) + p\mu_{br}^2 + q\mu^2\right]\right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[(p+q)\left(x - \frac{p\mu_{br}+q\mu}{p+q}\right)^2 + p\mu_{br}^2 + q\mu^2 - \frac{(p\mu_{br}+q\mu)^2}{p+q}\right]\right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[(p+q)\left(x - \frac{p\mu_{br}+q\mu}{p+q}\right)^2 + p\mu_{br}^2 + q\mu^2 - \frac{p^2\mu_{br}^2 + 2pq\mu_{br}\mu + q^2\mu^2}{p+q}\right]\right) dx$$

$$= \Big/ \text{substitute variables } 2 : t = x - \frac{p\mu_{br}+q\mu}{p+q} \rightarrow dt = dx \Big/$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p+q)}{2}t^2\right) \exp\left(-\frac{1}{2}\left[\frac{pq\mu_{br}^2 - 2pq\mu_{br}\mu + pq\mu^2}{p+q}\right]\right) dt$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(pq\frac{(\mu_{br}^2-\mu)^2}{p+q}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(p+q)}{2}t^2\right) dt$$

$$= \Big/ \text{table integral} : \int_{-\infty}^{\infty} \exp(-cx^2) dx = \sqrt{\frac{\pi}{c}} \Big/$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{\frac{\pi}{\frac{(p+q)}{2}}} \exp\left(pq\frac{(\mu_{br}^2-\mu)^2}{p+q}\right)$$

$$= \Big/ \text{returning substitute } 1 \Big/$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{\frac{2\pi}{\frac{1}{\tau_{br}^2}+\frac{1}{\sigma^2}}} \exp\left(\frac{1}{\tau_{br}^2\sigma^2}\frac{(\mu_{br}^2-\mu)^2}{\frac{1}{\tau_{br}^2}+\frac{1}{\sigma^2}}\right)$$

$$= \sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}} \exp\left(\frac{(\mu_{br}^2-\mu)^2}{\tau_{br}^2+\sigma^2}\right)$$

Integrating Poisson's sampling distribution over unknown q^r (Equation (B.11) to Equation (B.12))

Using prior $\frac{1}{q^r}$

$$
\begin{aligned}
&\int_0^\infty \frac{1}{q^r} P(D|\mu,\sigma^2) dq^r = \\
&= \int_0^\infty \frac{1}{q^r} \prod_{r,b} \frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu,\sigma^2) q^r}{\rho_b^r} \right)^{n_b^r} \exp\left(-N_b^r \frac{\rho_b^r(\mu,\sigma^2) q^r}{\rho_b^r} \right) dq^r \\
&= \int_0^\infty \frac{1}{q^r} \left(\prod_{b,r} \frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_b^r} \right) \prod_r (q^r)^{\sum_b n_b^r} \exp\left(-q^r \left(\sum_b N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right) \right) dq^r \\
&= \prod_{b,r} \left(\frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_b^r} \right) \prod_r \int_0^\infty \frac{1}{q^r} (q^r)^{n_r} \exp\left(-q^r \left(\sum_b N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right) \right) dq^r \\
&= \Big/ \text{substitute variables} : \alpha = n_r, \beta = \sum_b \frac{N_b^r \rho_b^r(\mu,\sigma^2)}{\rho_b^r} \Big/ \\
&= \prod_{b,r} \left(\frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_b^r} \right) \prod_r \int_0^\infty (q^r)^{\alpha-1} \exp\left(-q^r \beta\right) dq^r \\
&= \Big/ \text{Gamma integral} : \Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx \Big/ \\
&= \prod_{b,r} \left(\frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_b^r} \right) \prod_r \frac{1}{\beta^{\alpha-1}} \int_0^\infty (q^r \beta)^{\alpha-1} \exp\left(-q^r \beta\right) dq^r \beta \\
&= \Big/ \text{return substitution} \Big/ \\
&= \prod_{b,r} \left(\frac{1}{n_b^r!} \left(N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_b^r} \right) \prod_r \frac{\Gamma(n_r)}{\left(\sum_b \frac{N_b^r \rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_r - 1}} \\
&\propto \prod_{b,r} \frac{1}{n_b^r!} \left[\left(N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{n_b^r} \left(\sum_b N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r} \right)^{-n_r} \right]
\end{aligned}
\tag{B.15}
$$

Likelihood to log-likelihood (Equation (B.12) to Equation (B.13))

$$
\begin{aligned}
&L(D|\mu,\sigma^2) = \log\left[P(D|\mu,\sigma^2)\right] = \\
&= \log\left[\prod_{b,r}\frac{1}{n_b^r!}\left[\left(\sum_b N_b^r\frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right)^{-n_r}\left(N_b^r\frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right)^{n_b^r}\right]\right] \\
&= \sum_{b,r} n_b^r \log\left[N_b^r\frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right] - \sum_r n_r\log\left[\sum_b N_b^r\frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right] - \sum_{b,r} n_b^r! \\
&= \sum_{b,r} n_b^r\left[\log N_b^r + \log\rho_b^r(\mu,\sigma^2) - \log\rho_b^r\right] - \sum_r n_r\log\left[\sum_b N_b^r\frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right] - \sum_{b,r} n_b^r! \\
&= \Big/\text{keeping only terms that depend on } \mu \text{ and } \sigma^2\Big/ \\
&= \sum_{b,r} n_b^r\log\left[\rho_b^r(\mu,\sigma^2)\right] - \sum_r n_r\log\left[\sum_b N_b^r\frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right]
\end{aligned}
\tag{B.16}
$$

The reason why we only keep the terms depending on μ and σ^2 is the following: if we asume a uniform prior over μ and σ^2 and $L(D|\mu,\sigma^2)$ is as given in Equation (B.16) we have the posterior given with:

$$
P(\mu,\sigma^2|D) = \frac{\exp(L(D|\mu,\sigma^2))}{\int_0^\infty \exp(L(D|\bar{\mu},\bar{\sigma^2})d\bar{\mu}d\bar{\sigma^2}} \tag{B.17}
$$

and we see that for $L(D|\mu,\sigma^2) \to L(D|\mu,\sigma^2) + \lambda$ the posterior in Equation (B.17) is unchanged.

Derivative calculations

Derivative over μ on $\rho_b^r(\mu,\sigma^2)$

$$
\begin{aligned}
\frac{\partial\rho_b^r(\mu,\sigma^2)}{\partial\mu} &= \frac{\partial}{\partial\mu}\sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}}\exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right) \\
&= \sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}}\exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right)\frac{-1}{2(\tau_{br}^2+\sigma^2)}2(\mu_{br}-\mu)(-1) \\
&= \rho_b^r(\mu,\sigma^2)\frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2}
\end{aligned}
\tag{B.18}
$$

Second derivative over μ on $\rho_b^r(\mu, \sigma^2)$

$$\begin{aligned}
\frac{\partial^2 \rho_b^r(\mu,\sigma^2)}{\partial^2 \mu} &= \frac{\partial}{\partial \mu}\rho_b^r(\mu,\sigma^2)\frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2} \\
&= \frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2}\frac{\partial}{\partial \mu}\rho_b^r(\mu,\sigma^2) + \rho_b^r(\mu,\sigma^2)\frac{\partial}{\partial \mu}\frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2} \\
&= \left(\frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2}\right)^2 \rho_b^r(\mu,\sigma^2) - \rho_b^r(\mu,\sigma^2)\frac{1}{\tau_{br}^2+\sigma^2} \\
&= \frac{\rho_b^r(\mu,\sigma^2)}{\tau_{br}^2+\sigma^2}\left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2} - 1\right)
\end{aligned} \tag{B.19}$$

Derivative over σ^2 on $\rho_b^r(\mu, \sigma^2)$

$$\begin{aligned}
\frac{\partial \rho_b^r(\mu,\sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}}\exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right) \\
&= \exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right)\frac{\partial}{\partial \sigma^2}\sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}} \\
&\quad + \sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}}\frac{\partial}{\partial \sigma^2}\exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right) \\
&= \exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right)\frac{-\sqrt{\tau_{br}^2}}{2}\left(\frac{1}{\tau_{br}^2+\sigma^2}\right)^{3/2} \\
&\quad + \sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}}\exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right)\frac{-(\mu_{br}-\mu)^2}{2}\frac{-1}{(\tau_{br}^2+\sigma^2)^2} \\
&= \frac{1}{2}\sqrt{\frac{\tau_{br}^2}{\tau_{br}^2+\sigma^2}}\exp\left(-\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)}\right)\left(\frac{(\mu_{br}-\mu)^2}{(\tau_{br}^2+\sigma^2)^2} - \frac{1}{\tau_{br}^2+\sigma^2}\right) \\
&= \frac{\rho_b^r(\mu,\sigma^2)}{2(\tau_{br}^2+\sigma^2)}\left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2} - 1\right)
\end{aligned} \tag{B.20}$$

It's interesting to note that $\frac{\partial \rho_b^r(\mu,\sigma^2)}{\partial \sigma^2} = \frac{1}{2}\frac{\partial^2 \rho_b^r(\mu,\sigma^2)}{\partial^2 \mu}$

Second derivative over σ^2 on $\rho_b^r(\mu, \sigma^2)$

$$
\begin{aligned}
&\frac{\partial^2 \rho_b^r(\mu,\sigma^2)}{\partial \sigma^2} \frac{1}{2} \frac{\partial}{\partial \sigma^2} \left(\rho_b^r(\mu,\sigma^2) \frac{(\mu_{br}-\mu)^2}{(\tau_{br}^2+\sigma^2)^2} - \frac{\rho_b^r(\mu,\sigma^2)}{\tau_{br}^2+\sigma^2} \right) \\
&= \frac{1}{2} \left(\frac{\rho_b^r(\mu,\sigma^2)}{2(\tau_{br}^2+\sigma^2)} \left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2} - 1 \right) \frac{(\mu_{br}-\mu)^2}{(\tau_{br}^2+\sigma^2)^2} - \rho_b^r(\mu,\sigma^2) \frac{2(\mu_{br}-\mu)^2}{(\tau_{br}^2+\sigma^2)^3} \right. \\
&\qquad \left. - \frac{\frac{\rho_b^r(\mu,\sigma^2)}{2(\tau_{br}^2+\sigma^2)} \left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2} - 1 \right) (\tau_{br}^2+\sigma^2) - \rho_b^r(\mu,\sigma^2)}{(\tau_{br}^2+\sigma^2)^2} \right) \\
&= \frac{\rho_b^r(\mu,\sigma^2)}{2} \left(\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)^3} \left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2} - 1 \right) - \frac{2(\mu_{br}-\mu)^2}{(\tau_{br}^2+\sigma^2)^3} - \frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)^3} \right. \\
&\qquad \left. + \frac{3}{2(\tau_{br}^2+\sigma^2)^2} \right) \\
&= \frac{\rho_b^r(\mu,\sigma^2)}{2} \left(\frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)^3} \left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2} - 1 \right) - \frac{4(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)^3} - \frac{(\mu_{br}-\mu)^2}{2(\tau_{br}^2+\sigma^2)^3} \right. \\
&\qquad \left. + \frac{3(\tau_{br}^2+\sigma^2)}{2(\tau_{br}^2+\sigma^2)^3} \right) \\
&= \frac{\rho_b^r(\mu,\sigma^2)}{4(\tau_{br}^2+\sigma^2)^3} \left(\frac{(\mu_{br}-\mu)^4}{\tau_{br}^2+\sigma^2} - 6(\mu_{br}-\mu)^2 + 3(\tau_{br}^2+\sigma^2) \right)
\end{aligned}
\tag{B.21}
$$

Mixed derivative over on $\rho_b^r(\mu, \sigma^2)$

Since all the functions in the composition of $\frac{\partial^2 \rho_b^r(\mu,\sigma^2)}{\partial \mu \partial \sigma^2}$ are continuous, we have:

$$
\begin{aligned}
\frac{\partial^2 \rho_b^r(\mu,\sigma^2)}{\partial\mu\partial\sigma^2} &= \frac{\partial^2 \rho_b^r(\mu,\sigma^2)}{\partial\sigma^2\partial\mu} \\
&= \frac{\partial}{\partial\sigma^2}\left(\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)\right) \\
&= \frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)\frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2} \\
&= \frac{\rho_b^r(\mu,\sigma^2)}{2(\tau_{br}^2+\sigma^2)}\left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2}-1\right)\frac{\mu_{br}-\mu}{\tau_{br}^2+\sigma^2} - \rho_b^r(\mu,\sigma^2)\frac{(\mu_{br}-\mu)}{(\tau_{br}^2+\sigma^2)^2} \\
&= \rho_b^r(\mu,\sigma^2)\left(\frac{(\mu_{br}-\mu)^3}{2(\tau_{br}^2+\sigma^2)^3} - \frac{\mu_{br}-\mu}{2(\tau_{br}^2+\sigma^2)^2} - \frac{\mu_{br}-\mu}{(\tau_{br}^2+\sigma^2)^2}\right) \\
&= \rho_b^r(\mu,\sigma^2)\left(\frac{(\mu_{br}-\mu)^3}{2(\tau_{br}^2+\sigma^2)^3} - \frac{3(\mu_{br}-\mu)}{2(\tau_{br}^2+\sigma^2)^2}\right) \\
&= \rho_b^r(\mu,\sigma^2)\frac{\mu_{br}-\mu}{2(\tau_{br}^2+\sigma^2)^2}\left(\frac{(\mu_{br}-\mu)^2}{\tau_{br}^2+\sigma^2}-3\right)
\end{aligned}
\tag{B.22}
$$

Derivative over μ on $L(D|\mu,\sigma^2)$

$$
\begin{aligned}
&\frac{\partial L(D|\mu,\sigma^2)}{\partial\mu} = \\
&= \frac{\partial}{\partial\mu}\sum_{b,r} n_b^r \log\left[\rho_b^r(\mu,\sigma^2)\right] - \sum_r n^r \log\left[\sum_b N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right] \\
&= \sum_{b,r} n_b^r \frac{\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)}{\rho_b^r(\mu,\sigma^2)} - \sum_r n^r \frac{\sum_b N_b^r \frac{\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)}{\rho_b^r}}{\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}}
\end{aligned}
\tag{B.23}
$$

Second derivative over μ on $L(D|\mu,\sigma^2)$

$$\begin{aligned}
&\frac{\partial^2 L(D|\mu,\sigma^2)}{\partial^2\mu} = \\
&= \frac{\partial}{\partial\mu}\sum_{b,r} n_b^r \frac{\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)}{\rho_b^r(\mu,\sigma^2)} - \sum_r n^r \sum_b \frac{N_b^r \frac{\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)}{\rho_b^r}}{\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}} = \\
&= \sum_{b,r} n_b^r \frac{\frac{\partial}{\partial^2\mu}\rho_b^r(\mu,\sigma^2)\rho_b^r(\mu,\sigma^2) - (\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2))^2}{(\rho_b^r(\mu,\sigma^2))^2} \\
&- \sum_r n^r \sum_b \frac{N_b^r}{\rho_b^r} \frac{\frac{\partial}{\partial^2\mu}\rho_b^r(\mu,\sigma^2)\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right) - \frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\frac{\partial}{\partial\mu}\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right)}{\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right)^2}
\end{aligned} \tag{B.24}$$

Derivative over σ^2 on $L(D|\mu,\sigma^2)$

$$\begin{aligned}
&\frac{\partial L(D|\mu,\sigma^2)}{\partial\sigma^2} \\
&= \frac{\partial}{\partial\sigma^2}\sum_{b,r} n_b^r \log\left[\rho_b^r(\mu,\sigma^2)\right] - \sum_r n^r \log\left[\sum_b N_b^r \frac{\rho_b^r(\mu,\sigma^2)}{\rho_b^r}\right] \\
&= \sum_{b,r} n_b^r \frac{\frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)}{\rho_b^r(\mu,\sigma^2)} - \sum_r n^r \frac{\sum_b N_b^r \frac{\frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)}{\rho_b^r}}{\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}}
\end{aligned} \tag{B.25}$$

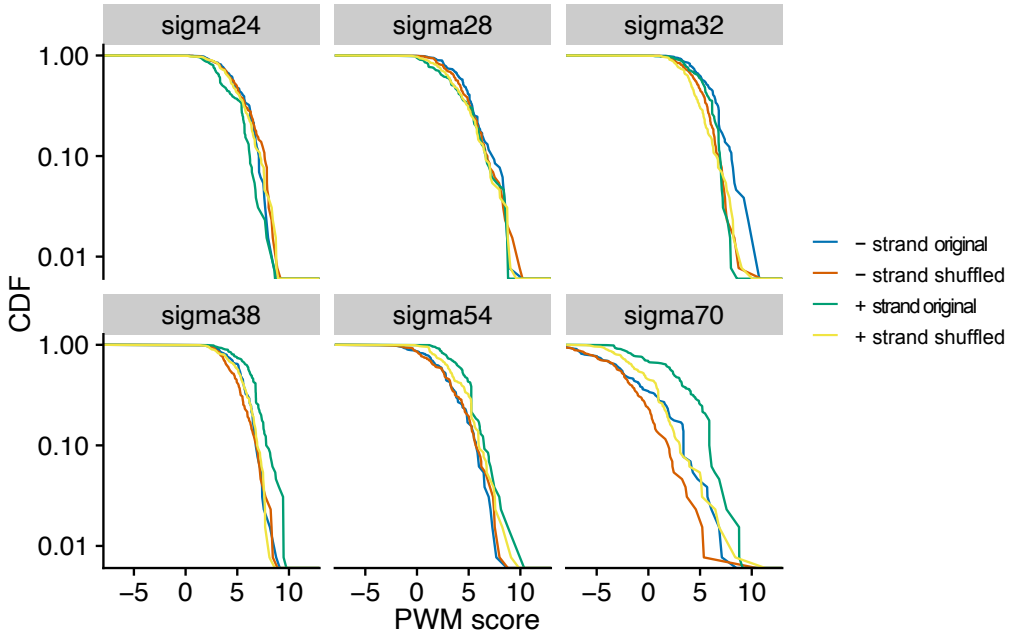
Second derivative over σ^2 on $L(D|\mu, \sigma^2)$

$$\begin{aligned}
&\frac{\partial^2 L(D|\mu,\sigma^2)}{\partial^2\sigma^2} = \\
&= \frac{\partial}{\partial\sigma^2}\sum_{b,r} n_b^r \frac{\frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)}{\rho_b^r(\mu,\sigma^2)} - \sum_r n^r \sum_b \frac{\frac{N_b^r}{\rho_b^r}\frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)}{\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}} \\
&= \sum_{b,r} n_b^r \frac{\frac{\partial}{\partial^2\sigma^2}\rho_b^r(\mu,\sigma^2)\rho_b^r(\mu,\sigma^2) - (\frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2))^2}{(\rho_b^r(\mu,\sigma^2))^2} \\
&- \sum_r n^r \sum_b \frac{N_b^r}{\rho_b^r} \frac{\frac{\partial}{\partial^2\sigma^2}\rho_b^r(\mu,\sigma^2)\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right) - \frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\frac{\partial}{\partial\sigma^2}\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right)}{\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right)^2}
\end{aligned} \tag{B.26}$$

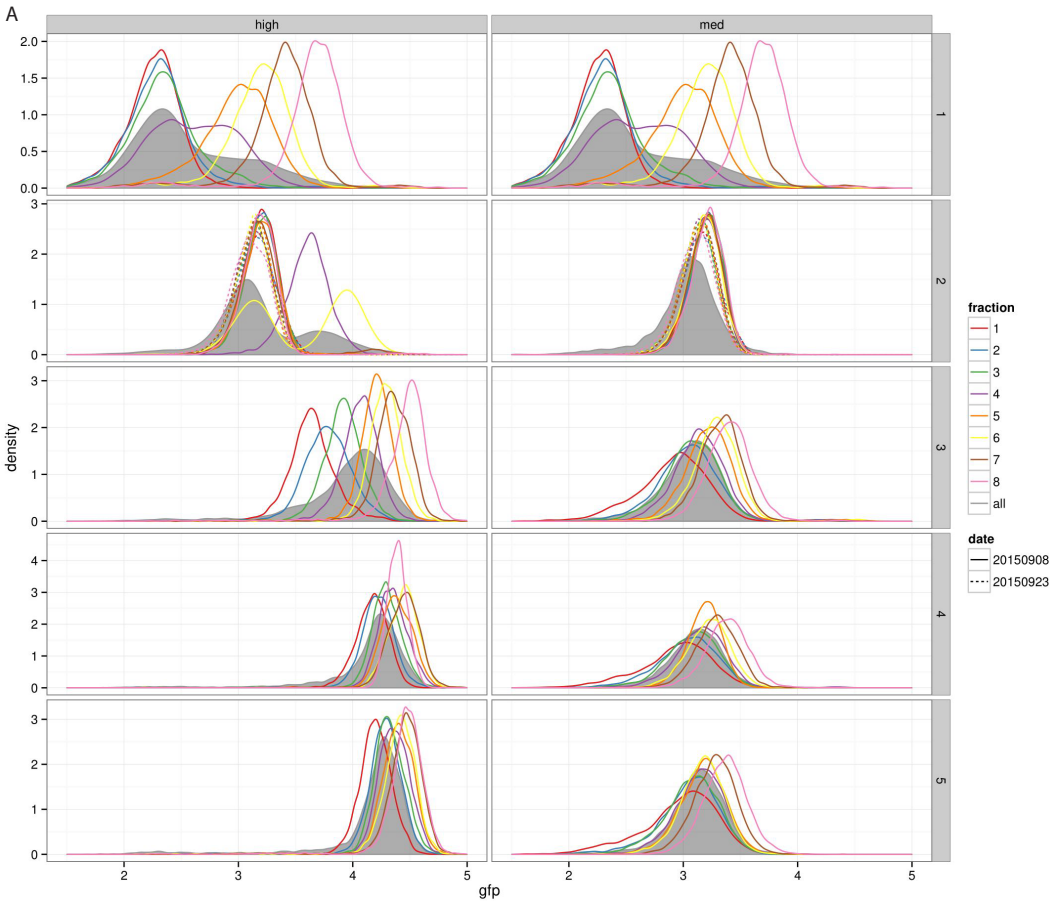
Mixed derivative on $L(D|\mu, \sigma^2)$

$$\begin{aligned}
&\frac{\partial^2 L(D|\mu,\sigma^2)}{\partial\mu\partial\sigma^2} = \\
&= \frac{\partial}{\partial\sigma^2}\sum_{b,r} n_b^r \frac{\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)}{\rho_b^r(\mu,\sigma^2)} - \sum_r n^r \sum_b \frac{\frac{N_b^r}{\rho_b^r}\frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)}{\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}} \\
&= \sum_{b,r} n_b^r \frac{\frac{\partial}{\partial\mu\partial\sigma^2}\rho_b^r(\mu,\sigma^2)\rho_b^r(\mu,\sigma^2) - \frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)\frac{\partial}{\partial\sigma^2}\rho_b^r(\mu,\sigma^2)}{(\rho_b^r(\mu,\sigma^2))^2} \\
&- \sum_r n^r \sum_b \frac{N_b^r}{\rho_b^r} \frac{\frac{\partial}{\partial\mu\partial\sigma^2}\rho_b^r(\mu,\sigma^2)\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right) - \frac{\partial}{\partial\mu}\rho_b^r(\mu,\sigma^2)\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\frac{\partial}{\partial\sigma^2}\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right)}{\left(\sum_{\tilde{b}} N_{\tilde{b}}^r \frac{\rho_{\tilde{b}}^r(\mu,\sigma^2)}{\rho_{\tilde{b}}^r}\right)^2}
\end{aligned} \tag{B.27}$$

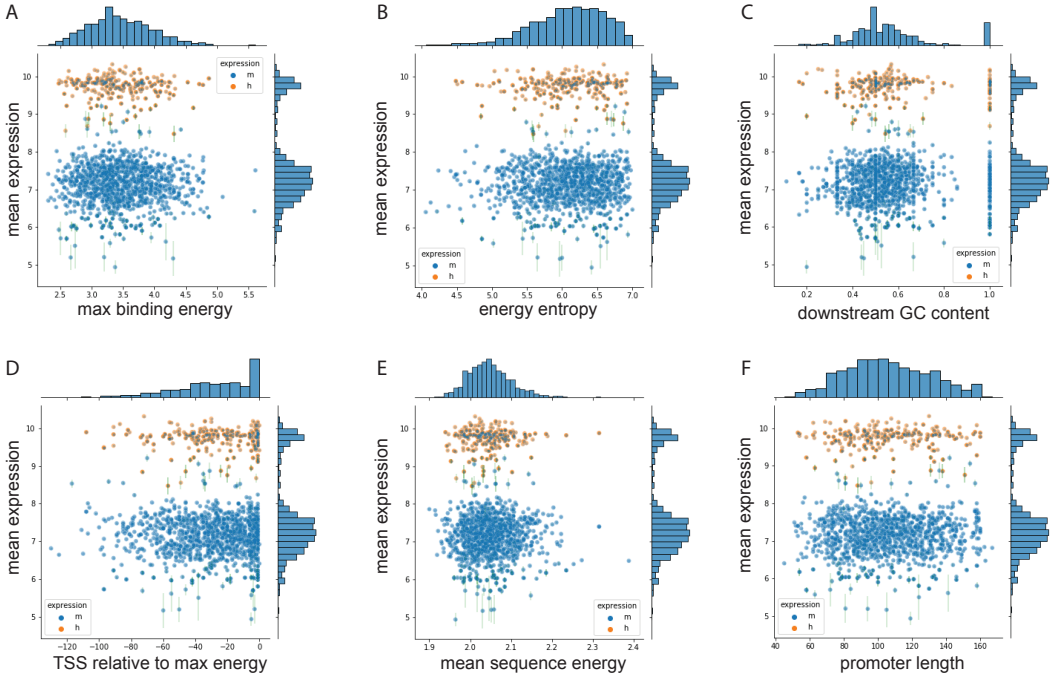
B.4 Supplementary figures



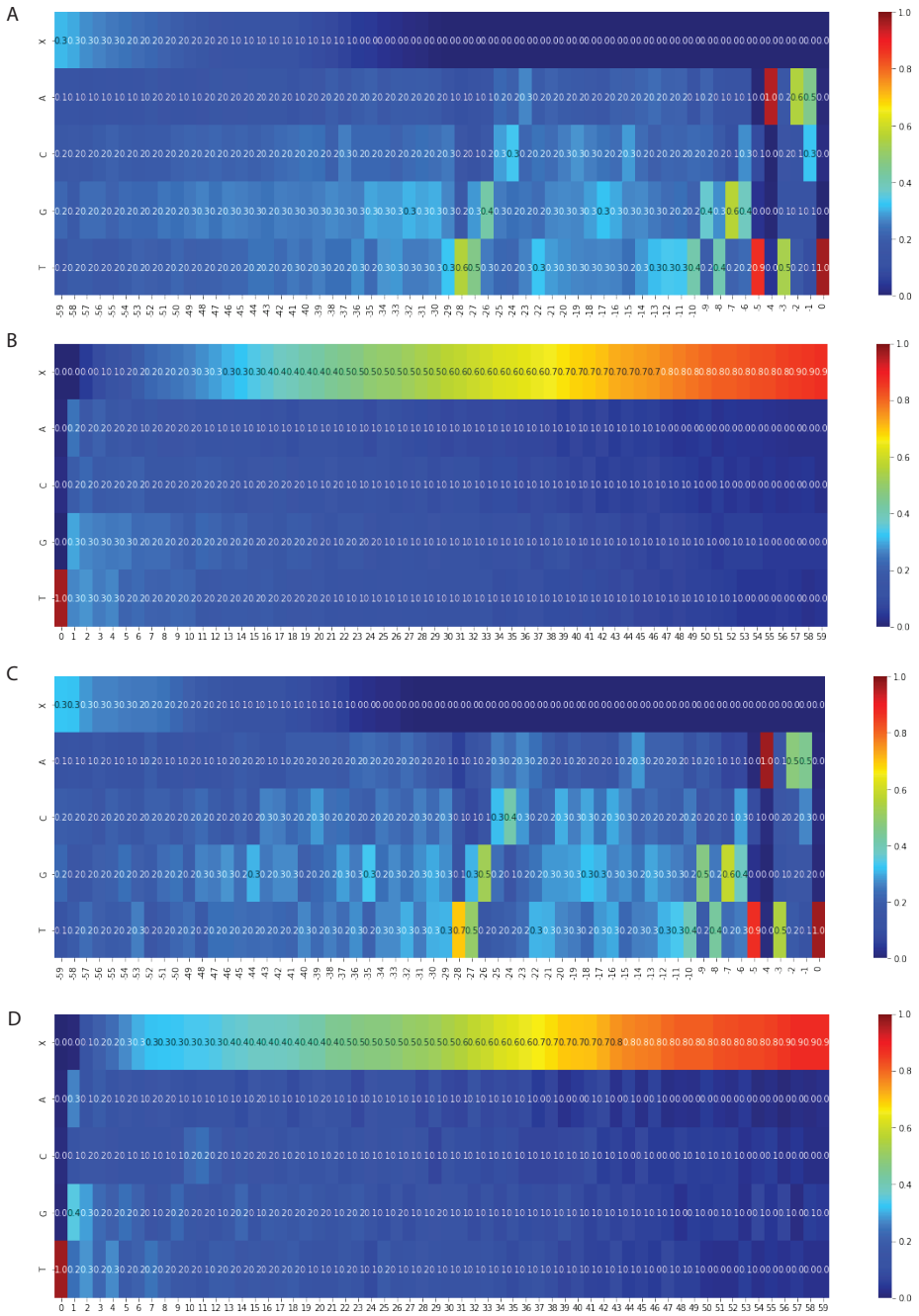
Supplementary Figure B.8: **Binding affinity of σ family of TFs**. For our synthetic promoters, we tested the σ^{70} specificity by running MotEvo[8] on all sequences with all known σ factors in *E. coli*. On the x-axis, we have the PWM score which represents the sum of PWM values for the best binding site. For each of the promoters, we checked the σ factor specificities for positive and negative strands, as well as for the shuffled strands. σ^{70} factor is the only one exhibiting an increased PWM score in the non-shuffled positive strand of the plasmid vector to which the random promoter was loaded.



Supplementary Figure B.9: **Enhanced promoter selection.** Each of the five selection rounds generated previously in [166] was segmented into 8 bins. Promoters from each of the rounds were FACS sorted into each of the 8 bins and subsequently sequenced using NGS.



Supplementary Figure B.10: **Promoter features**. In each of the panels, on the y-axis is the mean expression of a given promoter. Each dot is a promoter with orange dots being high expressors and blue dots being expressors. On each of the panels, we inspect the predictive power of promoter sequence features inferred from the σ^{70} binding affinity model.



Supplementary Figure B.11: **Initiator and discriminator sequence.** Promoters were stratified into two groups: medium (A-B) and high (C-D) expressors. Relative to the best binding site (hypothetical TSS), sequences were aligned and nucleotide composition was computed. High expressors exhibit a slightly stronger preference towards the -35 foot (middle of the figures), with no apparent preference in the discriminator sequence composition (A and C). Initiator sequences (B and D) are indistinguishable for medium and high expressors.

Bibliography

[1] The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

[2] Kete Ai, Kai Luo, Youshen Li, Wei Hu, Weihua Gao, Liu Fang, Guangming Tian, Guoliang Ruan, and Qiaoqing Xu. Expression pattern analysis of irf4 and its related genes revealed the functional differentiation of irf4 paralogues in teleost. *Fish & shellfish immunology*, 60:59–64, 2017.

[3] Kete Ai, Kai Luo, Lihai Xia, Weihua Gao, Wei Hu, Zhitao Qi, and Qiaoqing Xu. Functional characterization of interferon regulatory factor 5 and its role in the innate antiviral immune response. *Fish & shellfish immunology*, 72:31–36, 2018.

[4] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11): 1083–1086, 2017.

[5] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential cell biology*. Garland Science, 2015.

[6] James C Alwine, David J Kemp, and George R Stark. Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences*, 74(12):5350–5354, 1977.

[7] Olivier Armant, Martin März, Rebecca Schmidt, Marco Ferg, Nicolas Diotel, Raymond Ertzer, Jan Christian Bryne, Lixin Yang, Isabelle

Baader, Markus Reischl, et al. Genome-wide, whole mount in situ analysis of transcriptional regulators in zebrafish embryos. *Developmental biology*, 380(2):351–362, 2013.

[8] Phil Arnold, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, 28(4):487–494, 2011.

[9] Patrick K Arthur, Maike Claussen, Susanne Koch, Katsiaryna Tarbashevich, Olaf Jahn, and Tomas Pieler. Participation of xenopus elr-type proteins in vegetal mrna localization during oogenesis. *Journal of Biological Chemistry*, 284(30):19982–19992, 2009.

[10] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[11] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of experimental medicine*, 79(2):137–158, 1944.

[12] Yael Avissar, Jung Choi, Jean DeSaix, Vladimir Jurukovski, Robert Wise, Connie Rye, et al. Biology: Openstax. 2013.

[13] Timothy L Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[14] Piotr J Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J Gruber, Mihaela Zavolan, and Erik van Nimwegen. Ismara: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome research*, 24(5):869–884, 2014.

[15] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11):1429–1435, 2006.

[16] Severin Berger, Mikhail Pachkov, Phil Arnold, Saeed Omidi, Nicholas Kelley, Silvia Salatino, and Erik van Nimwegen. Crunch: Integrated processing and modeling of chip-seq data in terms of regulatory motifs. *Genome research*, 29(7):1164–1177, 2019.

[17] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current opinion in genetics & development*, 15 (2):116–124, 2005.

[18] Valentina Boeva. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in genetics*, 7:24, 2016.

[19] WA Bonner, HR Hulett, RG Sweet, and LA Herzenberg. Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3):404–409, 1972.

[20] Sina Booeshaghi and Lior Pachter. An introduction to single-cell rna-seq, 2021. URL https://tinyurl.com/fd5ps86h.

[21] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell rna-seq data. *Nature Biotechnology*, pages 1–9, 2021.

[22] Robert C Brewster, Daniel L Jones, and Rob Phillips. Tuning promoter strength through rna polymerase binding site design in escherichia coli. *PLoS computational biology*, 8(12):e1002811, 2012.

[23] James A Briggs, Caleb Weinreb, Daniel E Wagner, Sean Megason, Leonid Peshkin, Marc W Kirschner, and Allon M Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(6392), 2018.

[24] Christopher Buccitelli and Matthias Selbach. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020.

[25] P Burda, P Laslo, and T Stopka. The role of pu. 1 and gata-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia*, 24(7):1249–1257, 2010.

[26] Richard R Burgess and Larry Anthony. How sigma docks to rna polymerase and what sigma does. *Current opinion in microbiology*, 4(2): 126–131, 2001.

[27] RICHARD R Burgess and ANDREW A Travers. Escherichia coli rna polymerase: purification, subunit structure, and factor requirements. In *Federation proceedings*, volume 29, pages 1164–1169, 1970.

[28] Stephen K Burley and Katsuhiko Kamada. Transcription factor complexes. *Current opinion in structural biology*, 12(2):225–230, 2002.

[29] Brian Bushnell, Jonathan Rood, and Esther Singer. Bbmerge–accurate paired shotgun read merging via overlap. *PloS one*, 12(10):e0185056, 2017.

[30] Patrick Cahan, Hu Li, Samantha A Morris, Edroaldo Lummertz Da Rocha, George Q Daley, and James J Collins. Cellnet: network biology applied to stem cell engineering. *Cell*, 158(4):903–915, 2014.

[31] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[32] Laurie F Caslake, Shovon I Ashraf, and Anne O Summers. Mutations in the alpha and sigma-70 subunits of rna polymerase affect expression of the mer operon. *Journal of Bacteriology*, 179(5):1787–1795, 1997.

[33] Tara Chari, Joeyta Banerjee, and Lior Pachter. The specious art of single-cell genomics. *bioRxiv*, 2021.

[34] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, 17(6):1009–1023, 2016.

[35] Zelin Chen, Yoshihiro Omori, Sergey Koren, Takuya Shirokiya, Takuo Kuroda, Atsushi Miyamoto, Hironori Wada, Asao Fujiyama, Atsushi Toyoda, Suiyuan Zhang, et al. De novo assembly of the goldfish (carassius auratus) genome and the evolution of genes after whole-genome duplication. *Science advances*, 5(6):eaav0547, 2019.

[36] Biswanath Chowdhury and Gautam Garai. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5-6):419–431, 2017.

[37] Mei-I Chung, Sara M Peyrot, Sarah LeBoeuf, Tae Joo Park, Kriston L McGary, Edward M Marcotte, and John B Wallingford. Rfx2 is broadly required for ciliogenesis during vertebrate development. *Developmental biology*, 363(1):155–165, 2012.

[38] FS Collins, ES Lander, J Rogers, RH Waterston, and IHGS Conso. Finishing the euchromatic sequence of the human genome. *Nature*, 431 (7011):931–945, 2004.

[39] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

[40] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[41] Ralf Dahm. Discovering dna: Friedrich miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581, 2008.

[42] Charles Darwin. *The origin of species*. PF Collier & son New York, 1909.

[43] Robert L Davis, Harold Weintraub, and Andrew B Lassar. Expression of a single transfected cdna converts fibroblasts to myoblasts. *Cell*, 51 (6):987–1000, 1987.

[44] Bernardo P de Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of enhancers. *bioRxiv*, 2021.

[45] Stephen H Devoto, Ellie Melançon, Judith S Eisen, and Monte Westerfield. Identification of separate slow and fast muscle precursor cells in vivo, prior to somite formation. *Development*, 122(11):3371–3380, 1996.

[46] Marko Djordjevic. Redefining escherichia coli σ70 promoter elements:-15 motif as a complement of the- 10 motif. *Journal of bacteriology*, 193 (22):6305–6314, 2011.

[47] René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database: expansion of epdnew and new promoter analysis tools. *Nucleic acids research*, 43(D1):D92–D96, 2015.

[48] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[49] Sean Eddy. Hmmer user's guide. *Department of Genetics, Washington University School of Medicine*, 2(1):13, 1992.

[50] Csaba Erö, Marc-Oliver Gewaltig, Daniel Keller, and Henry Markram. A cell atlas for the mouse brain. *Frontiers in neuroinformatics*, 12:84, 2018.

[51] Martin J Evans and Matthew H Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *nature*, 292(5819):154–156, 1981.

[52] Jeffrey A Farrell, Yiqun Wang, Samantha J Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360 (6392):eaar3131, 2018.

[53] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39 (suppl_2):W29–W37, 2011.

[54] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[55] ArchibaldE Garrod. The incidence of alkaptonuria: a study in chemical individuality. *The Lancet*, 160(4137):1616–1620, 1902.

[56] Jay D Gralla, Agamemnon J Carpousis, and James E Stefano. Productive and abortive initiation of transcription in vitro at the lac uv5 promoter. *Biochemistry*, 19(25):5864–5869, 1980.

[57] Lenhard group. Common carp genome, 2021. URL http://data.genereg.net/gtan/Carp/assembly/cypCar1.fa.

[58] Roderic Guigo. An introduction to position specific scoring matrices. *Retrieved August*, 30, 2016.

[59] Mohamed-Ali Hakimi, Daniel A Bochar, Josh Chenoweth, William S Lane, Gail Mandel, and Ramin Shiekhattar. A core–braf35 complex containing histone deacetylase mediates repression of neuronal-specific genes. *Proceedings of the National Academy of Sciences*, 99(11):7420–7425, 2002.

[60] Sridhar Hannenhalli. Eukaryotic transcription factor binding sites—modeling and integrative search methods. *Bioinformatics*, 24(11): 1325–1331, 2008.

[61] Vasili Hauryliuk, Gemma C Atkinson, Katsuhiko S Murakami, Tanel Tenson, and Kenn Gerdes. Recent functional insights into the role of (p) ppgpp in bacterial physiology. *Nature Reviews Microbiology*, 13(5): 298–309, 2015.

[62] John Hawkins, Charles Grant, William Stafford Noble, and Timothy L Bailey. Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, 25(12):i339–i347, 2009.

[63] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

[64] Sarah Herberg, Krista R Gert, Alexander Schleiffer, and Andrea Pauli. The ly6/upar protein bouncer is necessary and sufficient for species-specific fertilization. *Science*, 361(6406):1029–1033, 2018.

[65] Kay Hofmann and Philipp Bucher. The rsp5-domain is shared by proteins of diverse functions. *FEBS letters*, 358(2):153–157, 1995.

[66] India G Hook-Barnard and Deborah M Hinton. The promoter spacer influences transcription initiation via $\sigma 70$ region 1.1 of escherichia coli

rna polymerase. *Proceedings of the National Academy of Sciences*, 106 (3):737–742, 2009.

[67] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021.

[68] Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. Animaltfdb 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research*, 47(D1):D33–D38, 2019.

[69] Timothy R Hughes. *A handbook of transcription factors*, volume 52. Springer Science & Business Media, 2011.

[70] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

[71] Randall Hyde. *The art of assembly language*. No Starch Press, 2010.

[72] Christian Iseli, Giovanna Ambrosini, Philipp Bucher, and C Victor Jongeneel. Indexing strategies for rapid searches of short words in genome sequences. *PloS one*, 2(6):e579, 2007.

[73] Makiko Iwafuchi-Doi and Kenneth S Zaret. Cell fate control by pioneer transcription factors. *Development*, 143(11):1833–1837, 2016.

[74] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[75] Ernest Jay, Robert Bambara, Rt Padmanabhan, and Ray Wu. Dna sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping. *Nucleic Acids Research*, 1(3):331–354, 1974.

[76] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, 337(6096):816–821, 2012.

[77] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316 (5830):1497–1502, 2007.

[78] Kenji Kamimoto, Christy M Hoffmann, and Samantha A Morris. Celloracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*, 2020.

[79] Florian A Karreth, Yvonne Tay, and Pier Paolo Pandolfi. Target competition: transcription factors enter the limelight. *Genome biology*, 15 (4):1–3, 2014.

[80] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[81] Charles K Kaufman, Ping Zhou, H Amalia Pasolli, Michael Rendl, Diana Bolotin, Kim-Chew Lim, Xing Dai, Maria-Luisa Alegre, and Elaine Fuchs. Gata-3: an unexpected regulator of cell lineage determination in skin. *Genes & development*, 17(17):2108–2122, 2003.

[82] Kriti Kaushik, Vincent Elvin Leonard, Shamsudheen Kv, Mukesh Kumar Lalwani, Saakshi Jalali, Ashok Patowary, Adita Joshi, Vinod Scaria, and Sridhar Sivasubbu. Dynamic expression of long non-coding rnas (lncrnas) in adult zebrafish. *PloS one*, 8(12):e83616, 2013.

[83] Hatice S Kaya-Okur, Steven J Wu, Christine A Codomo, Erica S Pledger, Terri D Bryson, Jorja G Henikoff, Kami Ahmad, and Steven Henikoff. Cut&tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1):1–10, 2019.

[84] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.

[85] W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[86] Peter V Kharchenko. The triumphs and limitations of computational methods for scrna-seq. *Nature Methods*, pages 1–10, 2021.

[87] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

[88] Yusuke Kijima, Wang Wantong, Yoji Igarashi, Kazutoshi Yoshitake, Shuichi Asakawa, Yutaka Suzuki, Shugo Watabe, and Shigeharu Kinoshita. Age-associated different transcriptome profiling in zebrafish and rat: insight into diversity of vertebrate aging. *bioRxiv*, page 478438, 2018.

[89] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[90] Hisanori Kiryu, Taku Oshima, and Kiyoshi Asai. Extracting relations between promoter sequences and their strengths from microarray data. *Bioinformatics*, 21(7):1062–1068, 2005.

[91] Wenjun Kong, Yuheng C Fu, and Samantha A Morris. Capybara: A computational tool to measure cell identity and fate transitions. *bioRxiv*, 2020.

[92] Jana Krtková and Alexander R Paredez. Use of translation blocking morpholinos for gene knockdown in giardia lamblia. In *Morpholino Oligomers*, pages 123–140. Springer, 2017.

[93] Andrew T Kwon, David J Arenillas, Rebecca Worsley Hunt, and Wyeth W Wasserman. opossum-3: advanced analysis of regulatory motif over-representation across genes or chip-seq datasets. *G3: Genes—Genomes— Genetics*, 2(9):987–1002, 2012.

[94] Hwee Hui Lau, Natasha Hui Jin Ng, Larry Sai Weng Loo, Joanita Binte Jasmen, and Adrian Kee Keong Teo. The molecular functions of hepatocyte nuclear factors–in and beyond the liver. *Journal of hepatology*, 68 (5):1033–1048, 2018.

[95] Jeongwoo Lee, Daehee Hwang, et al. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9): 1428–1442, 2020.

[96] Sylvain Lemeille, Marie Paschaki, Dominique Baas, Laurette Morlé, Jean-Luc Duteyrat, Aouatef Ait-Lounis, Emmanuèle Barras, Fabien Soulavie, Julie Jerber, Joëlle Thomas, et al. Interplay of rfx transcription factors 1, 2 and 3 in motile ciliogenesis. *Nucleic acids research*, 48 (16):9019–9036, 2020.

[97] Huan Liu, Kaylia Duncan, Annika Helverson, Priyanka Kumari, Camille Mumm, Yao Xiao, Jenna Colavincenzo Carlson, Fabrice Darbellay, Axel Visel, Elizabeth Leslie, et al. Analysis of zebrafish periderm enhancers facilitates identification of a regulatory variant near human krt8/18. *Elife*, 9:e51325, 2020.

[98] Shumo Liu and Albert Libchaber. Some aspects of e. coli promoter evolution observed in a molecular evolution experiment. *Journal of molecular evolution*, 62(5):536–550, 2006.

[99] Tao Liu, Jorge A Ortiz, Len Taing, Clifford A Meyer, Bernett Lee, Yong Zhang, Hyunjin Shin, Swee S Wong, Jian Ma, Ying Lei, et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome biology*, 12(8):1–10, 2011.

[100] Hei-Yong G Lo, Ramon U Jin, Greg Sibbel, Dengqun Liu, Anju Karki, Matthew S Joens, Blair B Madison, Bo Zhang, Valerie Blanc, James AJ Fitzpatrick, et al. A single transcription factor is sufficient to induce and maintain secretory cell architecture. *Genes & development*, 31(2): 154–171, 2017.

[101] Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian RN Tivey, Simon C Potter, Robert D Finn, et al. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636–W641, 2019.

[102] Hiroto Maeda, Nobuyuki Fujita, and Akira Ishihama. Competition among seven escherichia coli σ subunits: relative binding affinities to the core rna polymerase. *Nucleic acids research*, 28(18):3497–3503, 2000.

[103] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[104] Hideo Matsumura, Akiko Ito, Hiromasa Saitoh, Peter Winter, Günter Kahl, Monika Reuter, Detlev H Krüger, and Ryohei Terauchi. Supersage. *Cellular microbiology*, 7(1):11–18, 2005.

[105] Bernhard Mayr and Marc Montminy. Transcriptional regulation by the phosphorylation-dependent factor creb. *Nature reviews Molecular cell biology*, 2(8):599–609, 2001.

[106] Abhishek Mazumder and Achillefs N Kapanidis. Recent advances in understanding σ70-dependent transcription initiation mechanisms. *Journal of molecular biology*, 2019.

[107] Barbara McClintock. Chromosome organization and genic expression. In *Cold Spring Harbor symposia on quantitative biology*, volume 16, pages 13–47. Cold Spring Harbor Laboratory Press, 1951.

[108] WR McClure, DK Hawley, P Youderian, and MM Susskind. Dna determinants of promoter selectivity in escherichia coli. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 47, pages 477–481. Cold Spring Harbor Laboratory Press, 1983.

[109] Ilona Miko. Gregor mendel and the principles of inheritance. *Nature Education*, 1(1):134, 2008.

[110] Rashid Minhas, Aleksandra Paterek, Maciej Łapiński, Michał Bazała, Vladimir Korzh, and Cecilia L Winata. A novel conserved enhancer at zebrafish zic3 and zic6 loci drives neural expression. *Developmental Dynamics*, 248(9):837–849, 2019.

[111] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021.

[112] Yirong Mo. Probing the nature of hydrogen bonds in dna base pairs. *Journal of molecular modeling*, 12(5):665–672, 2006.

[113] Alan M Moses, Derek Y Chiang, Daniel A Pollard, Venky N Iyer, and Michael B Eisen. Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome biology*, 5(12):1–15, 2004.

[114] Weipeng Mu, Joshua Starmer, Della Yee, and Terry Magnuson. Ezh2 variants differentially regulate polycomb repressive complex 2 in histone methylation and cell differentiation. *Epigenetics & chromatin*, 11(1): 1–14, 2018.

[115] Chinese Academy of Sciences National Center for Gene Research. Grass carp genome, 2021. URL http://www.ncgr.ac.cn/grasscarp/.

[116] Marshall W Nirenberg and J Heinrich Matthaei. The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences*, 47(10):1588–1602, 1961.

[117] Kazuhiro R Nitta, Arttu Jolma, Yimeng Yin, Ekaterina Morgunova, Teemu Kivioja, Junaid Akhtar, Korneel Hens, Jarkko Toivonen, Bart Deplancke, Eileen EM Furlong, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *elife*, 4:e04837, 2015.

[118] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, 2000.

[119] Mikhail Pachkov, Piotr J Balwierz, Phil Arnold, Evgeniy Ozonov, and Erik Van Nimwegen. Swissregulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic acids research*, 41(D1): D214–D220, 2012.

[120] R Padmanabhan, Raji Padmanabhan, and Ray Wu. Nucleotide sequence analysis of dna: Ix. use of oligonucleotides of defined sequence as primers in dna sequence analysis. *Biochemical and biophysical research communications*, 48(5):1295–1302, 1972.

[121] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680, 2009.

[122] Marilyn K Parra, Sherry Gee, Narla Mohandas, and John G Conboy. Efficient in vivo manipulation of alternative pre-mrna splicing events using antisense morpholinos in mice. *Journal of Biological Chemistry*, 286(8):6033–6039, 2011.

[123] Rafael Riudavets Puig, Paul Boddie, Aziz Khan, Jaime Abraham Castro-Mondragon, and Anthony Mathelier. Unibind: maps of high-confidence direct tf-dna interactions across nine species. *bioRxiv*, pages 2020–11, 2021.

[124] Bushra Raj, Jeffrey A Farrell, Jialin Liu, Jakob El Kholtei, Adam N Carte, Joaquin Navajas Acedo, Lucia Y Du, Aaron McKenna, Đorđe Relić, Jessica M Leslie, et al. Emergence of neuronal diversity during vertebrate brain development. *Neuron*, 108(6):1058–1074, 2020.

[125] Nikolaus Rajewsky, Massimo Vergassola, Ulrike Gaul, and Eric D Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. *BMC bioinformatics*, 3 (1):1–13, 2002.

[126] Wu Ray, D Tu Chen-pei, and R Padmanabhan. Nucleotide sequence analysis of dna xii. the chemical synthesis and sequence analysis of a dodecadeoxynucleotide which binds to the endolysin gene of bacteriophage lambda. *Biochemical and Biophysical Research Communications*, 55(4): 1092–1099, 1973.

[127] Jose Manuel Rodriguez, Juan Rodriguez-Rivas, Tomás Di Domenico, Jesús Vázquez, Alfonso Valencia, and Michael L Tress. Appris 2017: principal isoforms for multiple gene sets. *Nucleic acids research*, 46(D1): D213–D217, 2018.

[128] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.

[129] Frederick P Roth, Jason D Hughes, Preston W Estep, and George M Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

[130] Kumiko Samejima, Hiromi Ogawa, Carol A Cooke, Damien Hudson, Fiona MacIsaac, Susana A Ribeiro, Paola Vagnarelli, Stefano Cardinale, Alastair Kerr, Fan Lai, et al. A promoter-hijack strategy for conditional shutdown of multiply spliced essential cell cycle genes. *Proceedings of the National Academy of Sciences*, 105(7):2457–2462, 2008.

[131] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[132] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021.

[133] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[134] Christopher J Schoenherr and David J Anderson. The neuron-restrictive silencer factor (nrsf): a coordinate repressor of multiple neuron-specific genes. *Science*, 267(5202):1360–1363, 1995.

[135] Maxwell ER Shafer, Ahilya N Sawh, and Alexander F Schier. Gene family evolution underlies cell type diversification in the hypothalamus of teleosts. *BioRxiv*, 2020.

[136] Rahul Siddharthan, Eric D Siggia, and Erik Van Nimwegen. Phylogibbs: a gibbs sampling motif finder that incorporates phylogeny. *PLoS computational biology*, 1(7):e67, 2005.

[137] Stefan Siebert, Jeffrey A Farrell, Jack F Cazet, Yashodara Abeykoon, Abby S Primack, Christine E Schnitzler, and Celina E Juliano. Stem cell differentiation trajectories in hydra resolved at single-cell resolution. *Science*, 365(6451), 2019.

[138] Olin K Silander, Nela Nikolic, Alon Zaslaver, Anat Bren, Ilya Kikoin, Uri Alon, and Martin Ackermann. A genome-wide analysis of promoter-mediated phenotypic noise in escherichia coli. *PLoS genetics*, 8(1): e1002443, 2012.

[139] Saurabh Sinha, Erik Van Nimwegen, and Eric D Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(suppl_1):i292–i301, 2003.

[140] Saurabh Sinha, Mathieu Blanchette, and Martin Tompa. Phyme: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC bioinformatics*, 5(1):1–17, 2004.

[141] Julie Soutourina. Transcription regulation by the mediator complex. *Nature reviews Molecular cell biology*, 19(4):262–274, 2018.

[142] Gary D Stormo, Thomas D Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, 10(9):2997–3011, 1982.

[143] Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nature Methods*, pages 1–9, 2021.

[144] James E Summerton. Invention and early history of morpholinos: from pipe dream to practical products. *Morpholino Oligomers*, pages 1–15, 2017.

[145] Xiao-Jian Sun, Zhanxin Wang, Lan Wang, Yanwen Jiang, Nils Kost, T David Soong, Wei-Yi Chen, Zhanyun Tang, Tomoyoshi Nakadai, Olivier Elemento, et al. A stable transcription factor complex nucleated by oligomeric aml1–eto controls leukaemogenesis. *Nature*, 500(7460): 93–97, 2013.

[146] Sowmya Swaminathan. Gfp: the green revolution. *Nature Cell Biology*, 11(1):S20–S20, 2009.

[147] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[148] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

[149] Kimiko Takebayashi-Suzuki, Hidenori Konishi, Tatsuo Miyamoto, Tomoko Nagata, Misa Uchida, and Atsushi Suzuki. Coordinated regulation of the dorsal-ventral and anterior-posterior patterning of xenopus embryos by the btb/poz zinc finger protein zbtb14. *Development, growth & differentiation*, 60(3):158–173, 2018.

[150] Stephen J Tapscott, Robert L Davis, Mathew J Thayer, Pei-Feng Cheng, Harold Weintraub, and Andrew B Lassar. Myod1: a nuclear phosphoprotein requiring a myc homology region to convert fibroblasts to myoblasts. *Science*, 242(4877):405–411, 1988.

[151] Kathryn E Tiller and Peter M Tessier. Advances in antibody design. *Annual review of biomedical engineering*, 17:191–216, 2015.

[152] Pablo Emiliano Tomatis, Marco Schütz, Elina Umudumov, and Andreas Plückthun. Mutations in sigma 70 transcription factor improves expression of functional eukaryotic membrane proteins in escherichia coli. *Scientific reports*, 9(1):1–14, 2019.

[153] Andrew A Travers and Richard R Burgess. Cyclic re-use of the rna polymerase sigma factor. *Nature*, 222(5193):537–540, 1969.

[154] Craig Tuerk and Larry Gold. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *science*, 249(4968):505–510, 1990.

[155] Vladimir N Uversky. Intrinsically disordered proteins and their "mysterious"(meta) physics. *Frontiers in Physics*, 7:10, 2019.

[156] Koen Van den Berge, Katharina M Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I Love, Rob Patro, and Mark D Robinson. Rna sequencing data: Hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science*, 2:139–173, 2019.

[157] Erik van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC bioinformatics*, 8(6):1–26, 2007.

[158] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[159] Todd Wasson and Alexander J Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome research*, 19(11): 2101–2112, 2009.

[160] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[161] Matthew T Weirauch and TR Hughes. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In *A handbook of transcription factors*, pages 25–73. Springer, 2011.

[162] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6): 1431–1443, 2014.

[163] Travis J Wheeler, Jody Clements, and Robert D Finn. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden markov models. *BMC bioinformatics*, 15(1):1–9, 2014.

[164] Edgar Wingender, Peter Dietze, Holger Karas, and Rainer Knüppel. Transfac: a database on transcription factors and their dna binding sites. *Nucleic acids research*, 24(1):238–241, 1996.

[165] Luise Wolf. *Evolution of transcriptional regulation in" Escherichia coli"*. PhD thesis, University_of_Basel, 2014.

[166] Luise Wolf, Olin K Silander, and Erik van Nimwegen. Expression noise facilitates the evolution of gene regulation. *Elife*, 4:e05856, 2015.

[167] Yujian Wu, Xiangjing Hu, Zhen Li, Min Wang, Sisi Li, Xiuxia Wang, Xiwen Lin, Shangying Liao, Zhuqiang Zhang, Xue Feng, et al. Transcription factor rfx2 is a key regulator of mouse spermiogenesis. *Scientific reports*, 6(1):1–13, 2016.

[168] Xuefei Yuan, Mengyi Song, Patrick Devine, Benoit G Bruneau, Ian C Scott, and Michael D Wilson. Heart enhancers with deeply conserved regulatory activity are established early in zebrafish development. *Nature communications*, 9(1):1–14, 2018.

[169] Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25 (21):2227–2241, 2011.

[170] Alon Zaslaver, Anat Bren, Michal Ronen, Shalev Itzkovitz, Ilya Kikoin, Seagull Shavit, Wolfram Liebermeister, Michael G Surette, and Uri Alon.

A comprehensive library of fluorescent transcriptional reporters for escherichia coli. *Nature methods*, 3(8):623–628, 2006.

[171] Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job Van Der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

[172] Qiong Zhang, Wei Liu, Hong-Mei Zhang, Gui-Yan Xie, Ya-Ru Miao, Mengxuan Xia, and An-Yuan Guo. htftarget: a comprehensive database for regulations of human transcription factors and their targets. *Genomics, proteomics & bioinformatics*, 18(2):120–128, 2020.

ĐORĐE RELIĆ

PhD in Bioinformatics, MSc in Computer Science

@ dorde.relic@pm.me brlauuu dorderelic

EXPERIENCE

Data Scientist & Software Engineer
Roche, Pharma Research and Early Development (pRED)
06/2022-present Basel, CH

Research Scientist, PhD Candidate & Postdoc
Biozentrum
09/2017-05/2022 Basel, CH

Research Collaborator
Harvard
09/2018, 05/2019 Cambridge, MA, USA

Teaching Assistant in Databases & Programming
University of Basel
02/2016-01/2021 Basel, CH

Co-founder & CIO
CAD Xpediters
05/2016-present Basel, CH/Sombor, SRB

Co-creator & Team Lead
Sonochrome
11/2013-12/2016 Belgrade, SRB

Software Developer Intern
Gecko Solutions
05/2015–07/2015 Belgrade, SRB

PUBLICATIONS

Journal Articles

- **Relić Đ** Pachkov M, Schier AF and van Nimwegen E (2023). "Inferring gene regulatory networks in *Danio rerio*". In: *In preparation*.
- Baranasic D. Hörtenhuber M. Balwierz P.J. Zehnder T. Mukarram A.K. Nepal C. Várnai C. Hadzhiev Y. Jimenez-Gonzalez A. Li N. Wragg J. D'Orazio F.M., **Relic D.** et al. (2022). "Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements". In: *Nature genetics* 54.7, pp. 1037–1050.
- Raj B. Farrell J.A. Liu J. El Kholtei J. Carte A. N. Du L.Y. McKenna A. **Relić Đ.** Leslie J.M., Schier A.F. (2020). "Emergence of neuronal diversity during vertebrate brain development". In: *Neuron*.

EDUCATION

PhD Bioinformatics
Biozentrum, University of Basel, CH
09/2017 – 12/2021

Thesis: *Modelling Gene Expression in Terms of DNA Sequence*

MSc Computer Science
University of Basel, CH
09/2015 – 06/2017

Thesis: *Learning Heuristic Functions Through Supervised Learning*

BSc Mathematics
University of Belgrade, SRB
09/2010 – 02/2015

CERTIFICATIONS

Scrum Master™ I (PSM I)
Scrum Product Owner™ I (PSPO I)

LANGUAGES

English | Serbian | German | Spanish

HONORS & AWARDS

Fellowship For Excellence
3 years fully funded PhD Fellowship at Biozentrum, provided by Siemens, class of 2017

Swiss MSP Representative
One of 13 Microsoft Student Partners from the world to be invited and hosted at the 2016 MVP Summit

MSc Studies Stipend
2 year scholarship from Kanton Basel-Stadt, CH, 2015-2016

Microsoft Imagine Cup
Imagine Cup 2014 World Finalist with project Sonochrome, held at Seattle, WA, USA

BSc Studies Stipend
1 year stipend provided by University of Belgrade, Serbia, 2014