Computational Perspectives on Transcription Regulation Throughout Evolution and Development

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 2 oktober 2024
om 16.30 uur precies

door

Maarten van der Sande
geboren op 13 augustus 1993
te Renkum

Promotor:

- Prof. Dr. G.J.C. Veenstra

Copromotor:

- Dr. S.J. van Heeringen

Manuscriptcommissie:

- Prof. dr. M.A. Huijnen,
- Prof. dr. C.F.H.A. Gilissen,
- Prof. dr. M. Richardson, Universiteit Leiden

Contents

Chapter 1

General Introduction

1.1 The central dogma of molecular biology

All cells within our body share the same DNA, yet they display remarkable diversity and specialization. How then, can a single set of genetic instructions (DNA) give rise to such diverse cell types? Consider the complexity of the human body, composed of an estimated 37 trillion cells[1,2]. Each of these cells contains approximately 2 meters of DNA, collectively forming an astounding 74,000,000,000 kilometers of genetic material, equivalent to nearly 250 round trips to the sun! This genetic information is distributed across 23 distinct structures known as chromosomes and contains roughly 20,000 genes. What is the role of evolution in this process, and consequently, what are the differences in instructions between species? To better understand these fascinating phenomena, it is crucial to understand the central dogma of molecular biology (Fig. 1.1)[3].

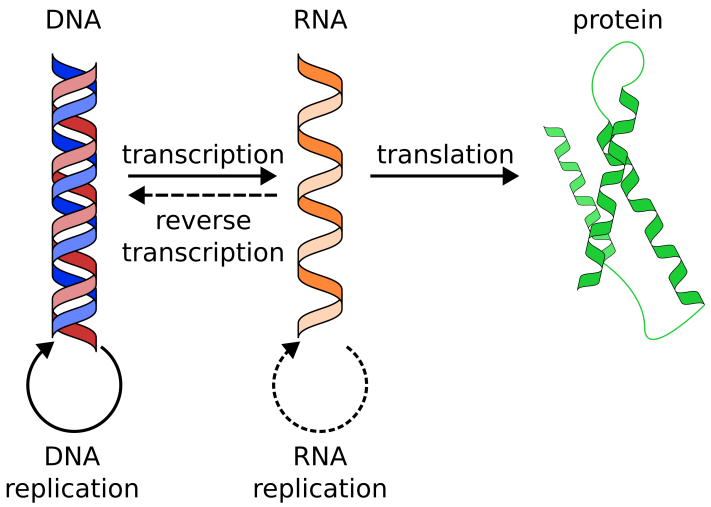


Figure 1.1: **The central dogma of molecular biology.** DNA transcribes to RNA, and RNA translates to protein. Solid arrows indicate the general flow of information in the system, and dotted arrows are special cases.

The central dogma of molecular biology describes the flow of genetic information within a biological system. Whereas in a computer information is stored in bits, which can be either zero (0) or one (1), genetic information is stored in nucleotides, which can be either adenine (A), cytosine (C), guanine (G), or thymine (T). DNA is composed of two large strands of these four nucleotides that together form a double helix. Both strands contain the same information, but where there is a nucleotide A on one strand, there is always a corresponding nucleotide T on its complementary strand, and similarly for C and G. RNA, on the other hand, is a similar molecule but is typically single-stranded. It is transcribed from DNA and shares a similar nucleotide composition but replaces thymine (T) with uracil (U). RNA serves as the bridge between DNA and proteins. Through the process of translation, the information encoded in RNA is decoded and used to assemble chains of amino acids, known as proteins. Proteins carry out various tasks in our bodies, such as enabling chemical reactions, transporting other molecules, providing structure, and acting as regulators of DNA transcription and RNA translation.
Far from all human DNA is transcribed to RNA, and not even all RNA translates to protein. As a general rule of thumb, we distinguish DNA sequences that get transcribed into RNA which in turn translate for proteins as genes. The human genome contains approximately 20,000 protein-

coding genes, and these genes cover only 1.5% of the genome[4]. Early molecular biologists mainly focused on protein-coding genes, thus the remaining 98.5% of the DNA got known as "junk DNA", a controversial term in the field[5]. More recent research has revealed that around 10% of human DNA is functional[5], although some estimates suggest that as much as 80% of DNA plays a role in at least one biochemical process[6]. Functional DNA is DNA that affects the overall fitness of an individual, and thus is conserved or under selective pressure[7].

1.2 Gene expression regulation

Even though all cells of an organism contain identical DNA, they use completely different sets of genes. This is possible due to the tight regulation of gene expression, for which cells have a wide array of tools at their disposal. This thesis mainly focuses on transcription factors and chromatin context in the context of gene expression regulation.

1.2.1 Transcription Factors

A typical gene consists of its coding regions (exons), noncoding regions (introns), and the start site (promoter). The promoter functions as a location where general transcription factors (GTFs) bind, which in turn recruit RNA polymerase II (RNAPII). RNAPII is the protein complex responsible for transcription and is responsible for reading out the DNA of a gene and converting it to RNA. Even though the presence of general transcription factors and a promoter is generally enough for transcription to occur[8], its regulation occurs through the interplay of transcription factors and enhancers (Fig. 1.2). Enhancers are small regions on the DNA, typically a few hundred base pairs in length. TFs bind in these enhancers, based on specific DNA sequences (motifs). Different TFs have different functions, and some help with the recruitment of RNAPII and GTFs, whilst others help with making the promoter or other DNA sequences more accessible. Transcription factors thus act like switches, determining whether transcription should be turned up or down in response to various signals and conditions.

Gene expression is regulated through the combined action of many (cis-)regulatory elements, such as the promoter and enhancers, but also silencers and insulators[9]. Silencers have a similar but opposite function to enhancers, they downregulate transcription. Insulators regulate the distances between enhancers/silencers and the promoter. Insulators function by recruiting proteins that create a physical barrier, effectively blocking interactions between regulatory elements located on opposite sides of the insulator. An insulator in between an enhancer and promoter can, for instance, block the TFs bound on the enhancer on the one side from recruiting GTFs to the promoter on the other side.

There are usually multiple transcription factors binding at a single enhancer, and the function of an enhancer should be considered at the combinatorial level[9]. The positioning of motifs in an enhancer is known as the motif grammar, and includes, for instance, the presence (or absence) of certain motifs, the order of these motifs, their orientation, and spacing[10]. The most strict motif grammar model is the enhanceosome, where for an enhancer to be functional, the motif presence, order, orientation, and spacing all need to be correct[11]. Alternatively, the billboard model predicts that only specific motifs need to be present, with little requirements for their order, orientation and spacing. The most loose model is the TF collective model, where the motif presence, order, orientation, and spacing are all dynamic[12]. Studying motif grammar is a hard problem, as there are many permutations of motifs possible in a single enhancer. Moreover, motif grammar changes based on the cell context, and it is costly to experimentally validate predictions.

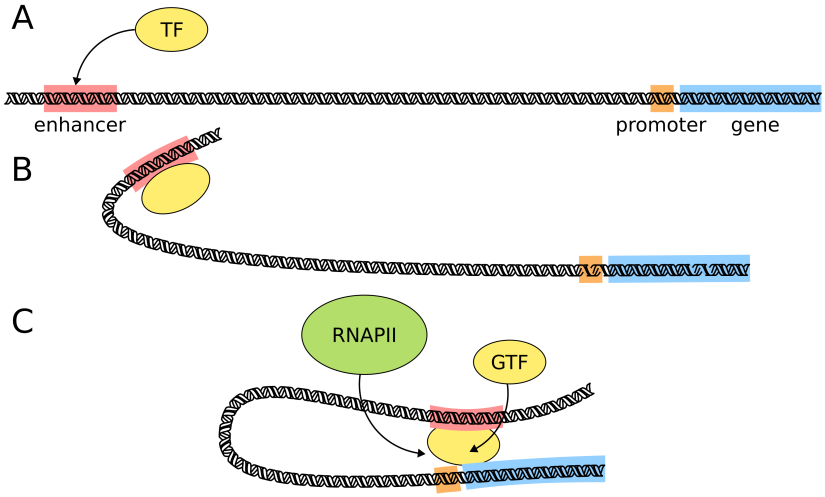


Figure 1.2: **Gene regulation by transcription factors. (A)** A gene (blue area) with a promoter upstream (orange) and a "distal" enhancer (red). A transcription factor (TF) binds to the enhancer. **(B)** The TF causes the DNA to "loop", bringing the enhancer and TF closer to the promoter. **(C)** Because of the DNA loop, the distal enhancer and promoter are in close proximity, and the TF can increase transcription by recruiting general transcription factors (GTFs) or RNA polymerase II (RNAPII). Usually, multiple enhancers and multiple transcription factors per enhancer are involved, and their combined action regulates gene expression.

1.2.2 Chromatin context

To be able to fit two meters of DNA into each cell, DNA needs to be carefully folded. Chromatin is the term for the structure of DNA, which consists of DNA and proteins. The main proteins in chromatin are histones. Histones are protein complexes around which DNA tightly wraps, forming a structure called a nucleosome. Nucleosomes are the basic units of chromatin (Fig. 1.3). Next to the structural role of chromatin, it also plays an important part in gene expression regulation, and can both promote and inhibit gene expression.

Histones have long protruding tails, that serve as a binding scaffold for protein complexes. These complexes in turn remodel the chromatin, for instance, to condense it, which blocks access of TFs to the DNA. The tails are chemically modifiable, such as by acetylation, methylation, and phosphorylation. The way to describe the modifications is by adding the histone number, amino acid, and type of modification. A well-studied modification is H3K27ac, which is the acetylation (ac) of the 27th amino acid, which is Lysine (K), of histone 3 (H3). H3K27ac tends to be associated with an open chromatin structure, allowing easier access to the DNA, and thus promoting the expression of nearby genes[13]. Another example is the H3K9me3 modification, which is the trimethylation (me3) of the 9th amino acid, Lysine (K), of histone 3 (H3). H3K9me3 is associated with a high nucleosome density and is generally a mark for repressed DNA[14].

Nucleosomes are not evenly spread out along the DNA. There are stretches with barely any nucleosomes, called nucleosome-free regions or open chromatin, or regions with a high density of nucleosomes, called closed chromatin or heterochromatin. The general idea is that closed chromatin is so tightly folded that it is difficult for proteins such as TFs to bind to the DNA. The other extreme is open chromatin, where (almost) no nucleosomes are present. Here, there is nothing to block proteins from accessing the DNA, and open chromatin is generally associated with active

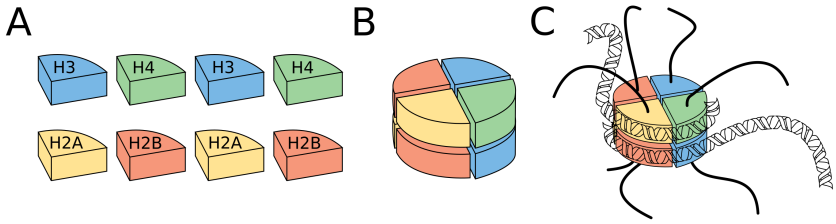


Figure 1.3: **The nucleosome consists of a histone complex with DNA wrapped around it. (A)** The eight subunits of a histone complex. **(B)** An assembled histone complex. **(C)** A histone with DNA wrapped around it. DNA wraps twice around a histone in 146 base pairs. Chemically modifiable tails stick out of the histones, which are important markers and regulators of chromatin state.

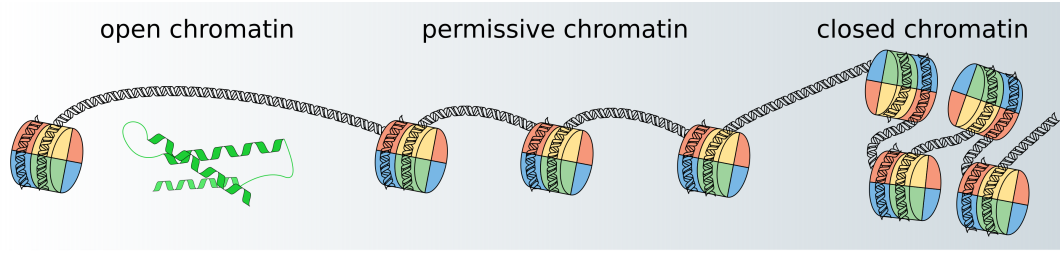


Figure 1.4: **Schematic representation of how chromatin context regulates DNA accessibility**. When no histones are bound (open chromatin) transcription factors can easily scan the DNA for motifs and bind. Closed chromatin on the other hand is so tightly bound by histones that the DNA has become inaccessible for transcription factors. In between lies permissive chromatin, which, depending on the histone modifications is accessible for transcription factors.

DNA[15]. In between open and closed chromatin is permissive chromatin, where nucleosomes are present, but the DNA is not folded so tightly that it has become inaccessible.
Another way chromatin context regulates gene expression is by the specific positioning of nucleosomes. A sequence of 146 base pairs wraps twice around a histone complex. By wrapping around a histone, two motifs at +/- 73 bp distance can suddenly become adjacent. It is thus not surprising that nucleosome positioning is far from random, and plays an important role in gene expression regulation[16].
Finally, another important aspect of gene regulation is DNA methylation, where methyl groups are added to certain regions of DNA. DNA methylation can cause genes to be "silenced" or turned off, as it prevents the cellular machinery from binding to the DNA and initiating gene transcription[17,18].

1.2.3 Other gene regulatory modes

There are many other ways gene expression is regulated, such as by regulating RNA degradation, post-translational modifications, and signal transduction. After transcription, RNA exists in the cell and gets translated into protein by ribosomes. Mammalian RNA has a lifetime of several minutes to days[19]. But to be able to react to environmental changes, a cell is able to mark RNAs for degradation, for instance by poly-A tail removal.
In addition, the activity and localization of proteins can be regulated by post-translational modifications[20]. These modifications play an important role in fine-tuning the functionality of proteins and their

interactions with other molecules. Many different post-translational modifications exist, such as phosphorylation, acetylation, sumoylation, and methylation. These modifications usually influence the protein's stability, shape, activity, and localization. Phosphorylation, for instance, can serve as a switch to activate or deactivate specific protein functions[21], ubiquitination marks proteins for degradation, and sumoylation is involved in proteins' stability and localization[22].
Finally, chemical and physical signals also play an important role in regulating gene expression. Our bodies react to light by keeping a day-night schedule, produce melanin after exposure to UV light, and grow our muscles after training. A particularly interesting example of how such signals regulate gene expression is the electrical signal that flatworms use during regeneration. A flatworm is a highly regenerative animal. You can cut off its head and tail, and the worm will regrow back its head and tail. It turns out that flatworms make use of an electrical gradient along their body that helps cells decide where in the body axis they are. When one cuts off the head and tail of a flatworm, whilst simultaneously interfering with this electrical gradient, you get flatworms that grow back two heads or tails[23].

1.2.4 Gene regulatory networks

Mapping how protein products of genes (*e.g.* transcription factors) influence the protein level of other genes is crucial for understanding how cellular processes are regulated. A common abstraction to understand gene-gene interactions is a gene regulatory network.
The first concept of gene regulatory networks was proposed by Roy Britten and Eric Davidson in 1969[24]. They observed that a cell (i) responds to an external signal; (ii) then produces its own signal as a response; (iii) transmits its own signal to receptors that do not perceive the external signal; (iv) then responds to its own signal, and finally (v) produces a protein as a response. Without modern knowledge of gene regulation, they then predicted what the minimum requirements are for such a system. One of the things they correctly predicted was the existence of transcription factors and promoter sequences. However, it took Eric Davidson more than 30 years to experimentally validate his original predictions[25].
Gene regulatory networks come in many forms and shapes but are often modeled and visualized with direct gene-gene interactions (Fig. 1.5A). This means for instance that the protein product of gene α directly regulates the protein product of gene β, ignoring all steps in between, such as transcription and DNA binding. Even though this is a gross oversimplification of gene-gene interactions, these simple models can already exhibit complex behavior. One of the older and more well-known examples of complex behavior from a simple model is a Turing pattern, first described by the famous computer scientist Alan Turing[26]. One specific Turing model, the Gierer-Meinhardt model, which consists of only two genes, gene α, and gene β, where gene α upregulates itself and gene β, but gene β inhibits gene α (Fig. 1.5B). When modeling this simple two-gene network in a spatial setting it produces complex patterns (Fig. 1.5C), of which similar patterns have been observed in nature, such as the skin pattern of a pufferfish (Fig. 1.5D).

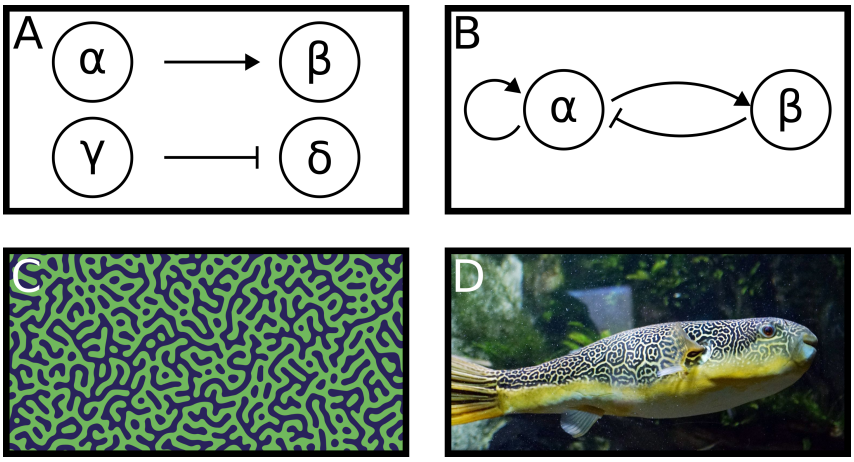


Figure 1.5: **Basic gene regulatory networks.** (**A**) the standard schematic way of representing gene-gene interactions. Gene α upregulates gene β, and gene γ downregulates gene δ. (**B**) Gierer-Meinhardt gene regulatory network, where gene α upregulates itself and gene β, and gene β downregulates gene α. (**C**) Simulation of the Gierer-Meinhardt gene regulatory network in a spatial context. Image taken from https://www.nature.com/articles/s43588-022-00306-0. (**D**) A Mbuna pufferfish with Turing pattern. Pufferfish photo taken by Tiia Monto: https://en.wikipedia.org/wiki/Mbu_pufferfish#/media/File:Tetraodon_mbu_2.jpg

Mapping the relations between genes is a hard problem and unfeasible to do by hand. Not only can potentially any gene have an (in)direct effect on any other gene, but the potential interactions also depend on the context, such as cell type or cell state. Theoretically, the 20,000 human protein-coding genes have $20{,}000^2$ = 400,000,000 potential direct interactions, and if we consider indirect interactions, cell type-specific interactions, and combinatorial interactions the number of potential gene networks explodes. To infer gene interactions at scale and make sense of the resulting networks we *have* to make use of computational tools. The most common approach so far has been to measure RNA expression in a combination of conditions and correlate the RNA expression of genes over these conditions[27,28]. Correlations above a certain threshold would indicate an (in)direct gene regulation. Adding more information to these networks, such as the chromatin state around genes[29,30], improves the resulting networks. In **Chapter 2** I discuss the latest developments in gene regulatory network inference.

1.3 Genomics analysis

In order to study the regulation of gene expression one needs to experimentally measure the molecular changes between conditions. Genomic experiments can usually be broken down into two main stages: wet lab work, and computational analysis. The wet lab phase involves hands-on experimentation with physical materials like chemicals and biological samples, DNA isolation, and DNA sequencing. In contrast, the computational analysis is concerned with the interpretation and analysis of the data generated by the wet lab. In this thesis, three different types of sequencing assays have been used and analyzed (ATAC-seq, ChIP-seq, and RNA-seq), and here I will briefly explain their wet lab and computational aspects.

1.3.1 Wet lab

The Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a technique employed to assess genome-wide chromatin accessibility[31]. Chromatin accessibility is closely linked to DNA activity; accessible DNA regions tend to be associated with active gene expression, while less accessible regions often correspond to inactive DNA. ATAC-seq operates by introducing a protein known as Tn5 transposase into a biological sample. This transposase enzyme semi-randomly cleaves the genomic DNA. However, in cases where DNA is inaccessible, for instance, due to obstructive histone modifications, the Tn5 transposase is unable to reach and cleave the DNA. Consequently, smaller DNA fragments represent accessible regions, as they could be cleaved. Subsequently, the smaller DNA fragments, typically less than 500 base pairs in length, are isolated and prepared for sequencing.
Chromatin ImmunoPrecipitation sequencing (ChIP-seq) is a technique used to pinpoint the locations of specific proteins in close proximity to DNA[32]. These proteins are often transcription factors or histone modifications. The method typically starts with the fixation of chromatin with formaldehyde, effectively preserving its structure. Subsequently, the DNA is fragmented into small segments, typically around 200 base pairs in length. Next, the proteins of interest are bound by specific antibodies. These antibody-protein-DNA complexes are then isolated, for instance by using magnetic antibodies. In the final stage, the DNA, proteins, and antibodies are separated, leaving behind the DNA fragments of interest. These fragments are then prepared for sequencing.
RNA sequencing (RNA-seq) is a technique to assess the presence and abundance of gene transcripts within a biological sample[33]. The quantity of RNA transcripts serves as a proxy for gene regulation as well as the potential amount of corresponding protein products. The RNA-seq process begins by isolating all the RNA molecules from the sample. Subsequently, these RNA molecules are converted into complementary DNA (cDNA) through a process known as reverse transcription (Fig. 1.1). The resulting cDNA then represents the abundance of gene transcripts within the sample. This cDNA then is ready for sequencing.

Sequencing

Sequencing is used to determine the precise order of nucleotides within a DNA sequence. In this introduction, our primary focus is on the sequencing by synthesis method[34], although several other techniques, such as chain-termination sequencing[34], Nanopore Sequencing[35], and Single-Molecule Real-Time Sequencing[36], are also available. Sequencing by synthesis takes place on a specialized surface referred to as a flow cell. The process begins with the addition of adapters to both ends of each DNA strand. Subsequently, the DNA is evenly distributed across the flow cell, with the adapters serving as anchors that firmly attach the DNA to the flow cell. Through polymerase chain reaction amplification (PCR), the DNA strands, along with the adapters, are duplicated, resulting in the duplicated sequences anchored in proximity to their original. Repeating the process of PCR amplification leads to the formation of clusters of identical DNA sequences concentrated at various spots on the flow cell. Next, the flow cell undergoes a heating step, causing the DNA strands to become single-stranded. In a sequential manner, fluorescently labeled nucleotides are introduced into the flow cell. These nucleotides are each associated with a distinct color that is emitted upon their incorporation into the DNA strand. Throughout this process, a camera captures the emitted colors from each spot. The nucleotide sequence of DNA is deduced by observing the order of colors emitted from each spot on the flow cell.

1.3.2 Dry lab

The outcome of sequencing is a data file, structured with as many lines as there were spots on the flow cell, with each line representing a DNA sequence (in FASTQ format). First, it is necessary to computationally remove the sequencing adapters from these sequences to prepare them for further analysis. The subsequent step involves aligning these sequences to a reference genome, a computationally intensive task given that FASTQ files typically contain several millions of reads, while genomes can be billions of nucleotides long. Following this alignment, a file is generated that maps each read to its corresponding position within the genome. In the context of RNA-seq experiments, the subsequent stages usually involve associating these genomic alignments with their genes. This process is relatively straightforward since the positions of genes are typically known. Researchers can simply count the number of reads aligned within each gene for each sample. However, in the case of ATAC-seq and ChIP-seq experiments, the regions of interest are usually not known in advance. Therefore, the initial task is to identify these regions. This is typically done through peak-calling algorithms[37–39]. These algorithms search for regions (peaks) where the number of mapped reads exceeds a statistically defined threshold. The results from multiple samples in ATAC-seq, ChIP-seq, and RNA-seq experiments are generally aggregated into a table format, where each row corresponds to a gene or peak, and each column represents a different sample. The numerical values within each cell indicate the number of reads mapped to a particular gene/peak for a specific sample.
After obtaining a table with values for each sample, the subsequent steps depend on the research objectives. If one is interested in the difference between two relatively similar conditions, such as a healthy vs. a disease condition, a typical approach is to perform a differential expression analysis on the genes or peaks[40]. One can, for instance, apply a gene set enrichment analysis on the differentially expressed genes, a method to identify functional classes of genes that are over-represented in either condition. Other approaches include looking for specific mutations that only occur in one of the two groups. A more advanced approach would be to try to infer the liver gene regulatory network based on these samples. This network can then help by identifying dysfunctional genes and possible treatments. In RNA-seq experiments, the inference of gene regulatory networks is often achieved by examining correlations in gene expression over multiple samples. Gene-gene correlations that surpass a predefined threshold are considered indicative of causality. In ATAC-seq or ChIP-seq experiments, regulatory network inference frequently involves a transcription factor motif analysis on the identified peaks. Ideally, both RNA-seq and ATAC/ChIP-seq data are combined to construct more accurate gene networks.
In **Chapter 3** I introduce the bioinformatics tool seq2science. Seq2science is a preprocessing pipeline, that requires FASTQ files as input and runs a standard analysis. Seq2science streamlines a significant portion of the dry lab workflow, resulting in standardized analyses and freeing up time for computational analysts for data exploration.

1.3.3 Single-cell

A significant drawback of traditional sequencing is that it aggregates the signal of all the cells present in a sample. If we were to take a biopsy of a liver, it would consist of a combination of liver cells (hepatocytes), blood cells, fat cells, etc., and sequencing this combination would result in a compound signal over cell types. This is why this type of sequencing has become known as bulk sequencing. Relatively recently, new techniques have been developed to separate all cells in a sample, give them cell-specific sequencing adapters and thus perform RNA- or ATAC-seq on each cell separately[41,42]. This type of data gives unprecedented resolution in analyses and has been welcomed as a major improvement over bulk sequencing.
A major downside of single-cell sequencing, however, is that the data analysis is significantly

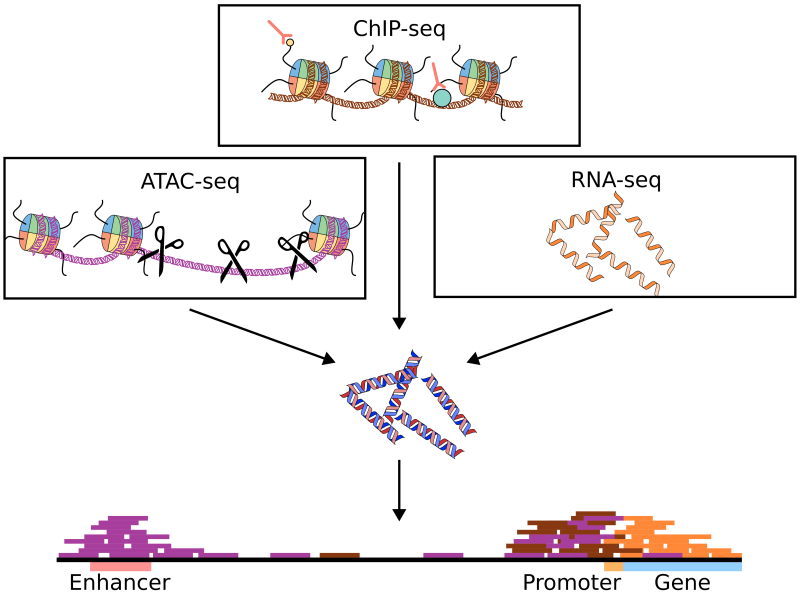


Figure 1.6: **Schematic overview of ATAC-seq, ChIP-seq, and RNA-seq and their computational analysis.** ATAC-seq is a technique that semi-randomly cleaves accessible DNA. Short DNA sequences thus represent accessible DNA. ChIP-seq is a way to isolate DNA sequences with specific proteins (usually TFs or histone modifications) in proximity. By filtering out the DNA after sonication with the bound antibodies (orange) you are left with DNA that was in close proximity to the protein of interest. RNA-seq is a technique for the estimation of the number of transcripts per gene. By reverse transcription the RNA gets converted into cDNA. The DNA these techniques produce can then be sequenced. After sequencing the DNA is computationally mapped to the genome by the dry lab, after which the real analysis starts.

harder. Whereas originally one would compare for instance a healthy liver vs. a diseased liver, now there are thousands of liver cells separated over multiple cell types to compare. Moreover, the hardware that was used for the original bulk analyses is suddenly not powerful enough to deal with the enormous increase in data. Furthermore, statistical methods developed for bulk analyses are sometimes not appropriate anymore. For instance, when studying the accessibility of a sequence of DNA, you can only measure each chromosome once, which results in either zero, one, or two hits per region. As a consequence, the assumption of a continuous distribution of hits per region is broken. Similarly, single-cell RNA-seq count tables are sparse, which means they are filled with zeroes. Biological reasons can cause this, the gene is simply not expressed, or by technical reasons, where insufficient transcripts are sequenced in a cell[43]. Moreover, single-cell sequencing has the problem that it is extremely sensitive to batch effects. With for instance major differences between who did the sequencing and when.
Regardless of these downsides, single-cell sequencing is a promising technique that has been quickly adopted in the field of genomics. It has given the field a new appreciation for the complexity of biology. In **Chapter 6** I introduce SCEPIA, a tool for the inference of TF activity based on single-cell RNA-seq.

1.4 Evolutionary development (evo-devo)

A single fertilized egg cell multiplies and develops into a complex collection of trillions of cells by the time we reach adulthood. How does each cell know where it is positioned in the body and what to develop into? The field of evolutionary development (evo-devo) studies development from an evolutionary point of view. In the 1980s scientists discovered a set of genes in fruit flies, which when mutated, were responsible for strange bodily transformations. A mutation in these genes caused flies to grow legs instead of antennae from their mouths[44] or flies to develop a second pair of wings[45]. This work was revolutionary at the time as it showed that (precursor) antennae cells contain all the information necessary to build legs. The group of genes responsible for these mutations are transcription factors now known as HOX genes.
The embryonic development of a fruit fly starts as a worm-like larva consisting of multiple repeated segments. The gene expression of HOX genes determines the identity of each segment and guides its growth. For example, the *lab* HOX gene is expressed at the frontal part of the larva, indicating surrounding cells this area will develop into the mouth of the fly[46]. Similarly, the *Antp* HOX gene indicates leg formation and is expressed in the mid-section of the larva. Interestingly, HOX genes are found in nearly all animals, where they play an important role in the spatial organization of the organism. What makes Hox genes particularly fascinating is that their order on the chromosome is the same as the spatial ordering of their gene expression along the embryo. This concept is called spatial colinearity, something we as yet have no clear explanation for[47].
Another remarkable family of transcription factors is the family of PAX transcription factors. While the HOX transcription factors are generally responsible for the anterior-posterior (head-to-tail) axis, the PAX transcription factors are involved in the development of specific structures like the eyes, ears, nervous system, and other organs. The most-studied transcription factor of the family is PAX6, which is involved in eye development and is highly conserved for bilateria (*e.g.* fruit flies, frogs, and humans). Injecting PAX6 into a developing frog embryo results in an extra eye on the spot of the injection[48].
The observation that single mutations can cause such large changes in body plans, in combination with the fact that the responsible genes are deeply conserved among species, shifted the way we think about evolution. Speciation does not have to happen through a combination of many small incremental mutations, but only a few mutations, for instance, in the HOX or PAX genes or their enhancers, can cause major changes. This is now the basis of the scientific field of evo-devo.

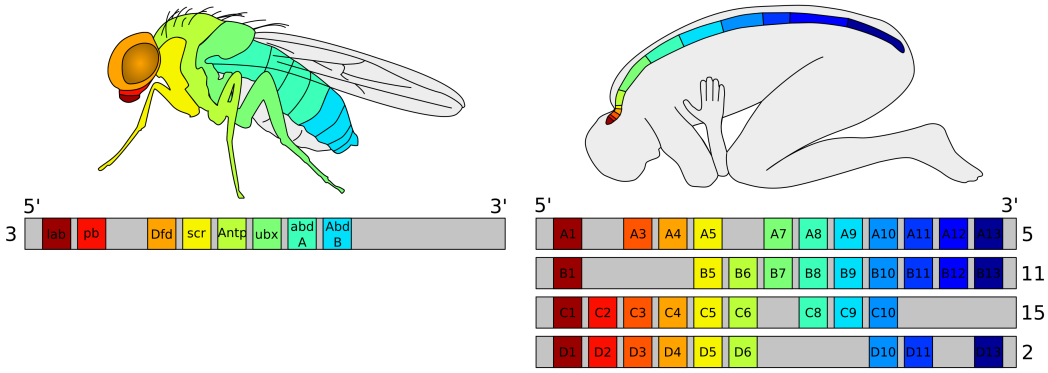


Figure 1.7: **The genomic ordering of HOX genes in fruit fly and human, and their colinearity in expression.** All the HOX genes for the fruit fly lay on chromosome 3. The human genome has had two genome duplications since the last common ancestor with fruit flies, so the HOX genes are spread over four chromosomes (5, 11, 15, and 2). Even though some HOX genes got lost on some chromosomes, their ordering has remained the same.

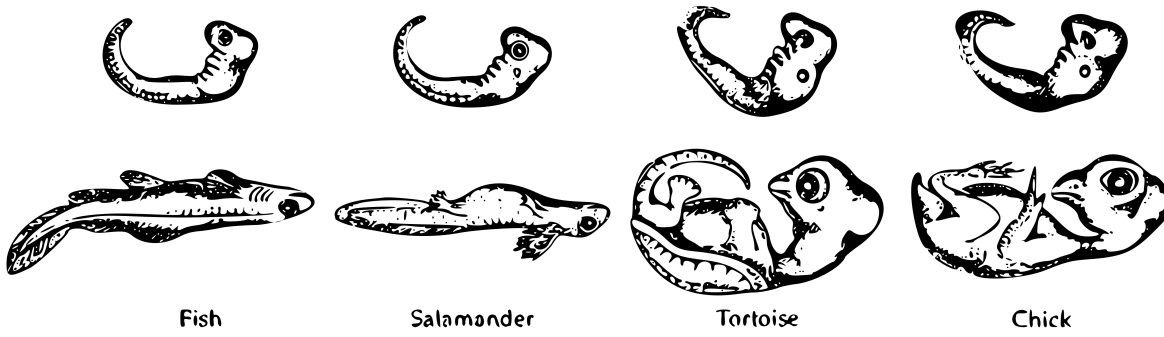
1.4.1 The hourglass model and the phylotypic stage

The molecular observations of the HOX and PAX genes are a relatively recent addition to the field of evo-devo. However, some of the earliest and most profound observations in this field are rooted in the morphological (shape-based) study of embryos. Ernst Haeckel, through a series of observations and drawings in the 19th century[49], noticed a stage early in development where all vertebrates appear morphologically similar. This stage is now known as the phylotypic stage (see Fig. 1.8 for an adaptation of some of his drawings). These observations came only a decade after publication of The Origin of Species, *the* book by Charles Darwin in which he proposes the evolution theory. At the time, the idea of a scala naturæ (the ordering of species into higher and lower species) was still prevalent. Haeckel attempted to integrate his observations of a conserved embryonic stage with the scala naturæ and the evolution theory, and came with a refinement of the recapitulation theory. Haeckel proposed that embryos of higher species consecutively develop from embryos of lower species into embryos of higher species. For example, a human embryo, which clearly would be the highest and most developed species, would first develop into a fish embryo, then a reptile embryo, and finally into a mammalian embryo. This would explain why, for instance, gills and tails develop in human embryos, to then later disappear. On top of that, Haeckel was a strong proponent of eugenics and believed such orderings to also exist in human races, where the Germanic race, coincidentally the race Haeckel belongs to, was listed all the way at the top[50]. The recapitulation theory was already controversial at the time it was proposed and is now refuted by the contemporary scientific community. Nevertheless, the notion of a morphologically conserved stage early in development is still prevalent to this day.
In the same time period, Karl Ernst von Baer proposed an alternative theory of embryonic development. Von Baer was strongly opposed to Haeckel's recapitulation theory. He noted, for instance, that a yolk sac is present during bird embryonic development, but not for frog embryonic development. This is inconsistent with the recapitulation theory as frogs were considered a higher species than birds. Von Baer's opposing theory consisted of four main rules, or laws[53]. His first law, for instance, states that *the more general characters of a large group appear earlier in the embryo than the more special characters.* This means that as an embryo develops, it first develops its oldest phylum-specific features, to then respectively develop its class, order, family, and species-specific features.

Simply put, embryos of related species become increasingly diverse as development proceeds. Von Baer did not believe in the idea of a single common ancestor for all life on earth, currently a widely accepted scientific concept, as he believed the differences between some species, *e.g.* humans, plants, and sponges, to be too large to be bridged by evolution.
Even though the recapitulation theory of Haeckel and of Von Baer's laws are generally dismissed by contemporary biologists, their observations of a morphologically conserved stage in development remain intriguing to the field and have led to the formulation of the hourglass model of development. The hourglass model is based on the model proposed by Paul Medawar in 1954[54]. Medawar argues that somewhere mid-embryogenesis is the most morphologically conserved stage for vertebrates. This stage corresponds to Haeckel's phylotypic stage, but, different from Haeckel's recapitulation theory, different species are thought to be more diverse both before and after the mid-embryogenesis state. The phylotypic stage coincides with the formation of the basic body plan, so it is a popular hypothesis that HOX genes, the genes responsible for the general layout of an organism, are responsible for this conserved stage[55]. Recently, molecular evidence has been generated that suggests that gene expression between different species is most similar at the phylotypic stage[56–64]. This has led to the idea that the phylotypic stage is not just morphologically conserved, but also conserved on the level of gene expression and regulation. In **Chapters 4 and 5** I discuss the current statistical methods that estimate this (molecular) conservation, and by careful re-analysis, I demonstrate that the methodologies applied are inadequate for substantiating its conclusions. In turn, this means that the conclusions the original authors draw about the phylotypic stage and the molecular hourglass model are unfounded.

1.5 Thesis overview

This thesis focuses on the computational analysis of (evolutionary-) developmental processes. The current chapter (**Chapter 1**) serves as a general introduction to the scientific fields of computational genomics, gene regulation, and evolutionary development.
In **Chapter 2** I review the current computational approaches to model and understand gene regulatory networks in development. Current computational gene regulatory network inference methods perform poorly, and thus new and improved approaches are needed. I highlight three recent developments for gene regulatory network inference which I expect to improve the power of these methods; multi-omics networks, single-cell data, and artificial neural networks. Multi-omics data partially solves the curse of dimensionality by constraining the number of possible networks and providing more information on regulatory interactions. Bulk sequencing measures the compound signal of multiple cell types. Single-cell sequencing, on the other hand, separates the signal per cell type, so cell-type-specific networks can be made and the data contains a purer signal. Artificial neural networks can model more complex gene-gene interactions than the current approaches. I expect that future gene regulatory network inference methods will have increased accuracy by combining these three developments.
In **Chapter 3** I discuss the implementation of seq2science, a next-generation pre-processing workflow for functional genomics. Seq2science supports some of the most common assays, such as RNA-seq, ChIP-seq, and ATAC-seq, integrates with public databases, and reports an extensive quality control report. Seq2science has been tested on a wide array of different species and genome assemblies, and I show examples of common analyses that seq2science supports out of the box. Seq2science has an extensive user base[29,65–82], is downloaded over 50K times through Bioconda, and has more than 130 "stars" on GitHub.
In **Chapter 4** I discuss the molecular basis of the phylotypic stage and its related models. I explain how the current definition and analyses of the phylotypic stage are ambiguous, as they do not distinguish within-species effects from between-species effects. For this reason, I propose that



any study of the phylotypic stage includes at least a within-species comparison, a within-phylum comparison, and a between-phyla comparison. By systematically applying these comparisons to earlier studies I find important flaws in their interpretation. I highlight three examples where the within-species pattern alone explains the between-species pattern. Moreover, I find that a supposed between-phyla effect, the mid-developmental transition, is a statistical artifact. Finally, I question the general validity of the current approaches to studying the phylotypic stage, as they are gross oversimplifications of the biological complexity during development.

Chapter 5 is a short re-analysis of a quantitative study of morphological features in relation to the phylotypic stage. The original study finds support for an inverse hourglass model for mammalian embryonic development. However, with simulated data with no particular temporal conservation, I find practically identical results. As such I see no support for the claim of an inverse hourglass model of conservation for morphological features.

In **Chapter 6** I present the computational tool SCEPIA. SCEPIA infers TF activities by computationally linking single-cell RNA-seq with a reference epigenomic database. I show that the inferred motif activities based on the epigenomic reference are more accurate than on transcriptomic data alone. Moreover, by combining transcriptomic information with motif activities, we can infer differential TF motifs between cells. Finally, I show the effectiveness of SCEPIA on a dataset of the human heart, where SCEPIA recovers several transcription factors that are known to be markers for specific cell types in the heart.

Finally, in **Chapter 7** I summarize and discuss the results described in this thesis, and give future perspectives on the field.

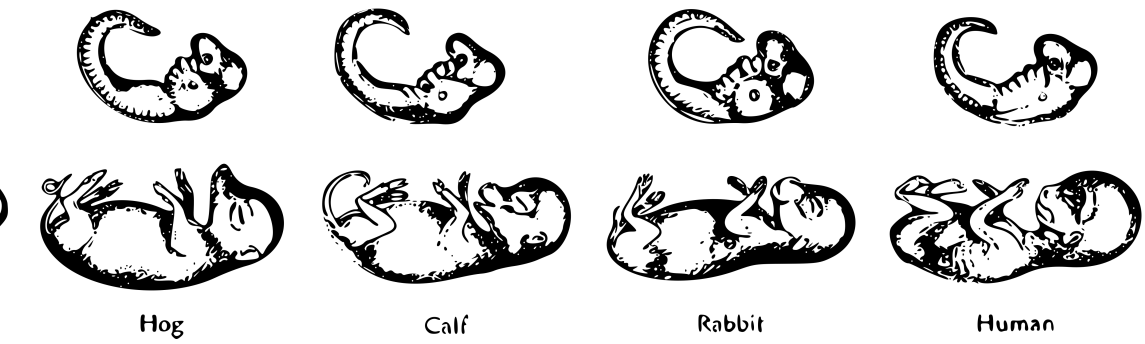


Figure 1.8: **Adaptation of George Romanes's 1892 copy of Ernst Haeckel's embryo drawings.** The upper row shows early embryos in the phylotypic stage, while the bottom row shows a later stage. Note how all the embryos appear similar in the phylotypic stage. Haeckel has been accused of fraud, amplifying the similarities at the phylotypic stage, with both supporters[51] and opponents of his drawings[52].

Chapter 2

Computational Approaches to Understand Transcription Regulation in Development

Maarten van der Sande*, Siebren Frölich*, Simon J. van Heeringen

2.1 Abstract

Gene regulatory networks (GRNs) serve as useful abstractions to understand transcriptional dynamics in developmental systems. Computational prediction of GRNs has been successfully applied to genome-wide gene expression measurements with the advent of microarrays and RNA-sequencing. However, these inferred networks are inaccurate and mostly based on correlative rather than causative interactions. In this review, we highlight three approaches that significantly impact GRN inference: (1) moving from one genome-wide functional modality, gene expression, to multi-omics, (2) single cell sequencing, to measure cell type-specific signals and predict context-specific GRNs, and (3) neural networks as flexible models. Together, these experimental and computational developments have the potential to significantly impact the quality of inferred GRNs. Ultimately, accurately modeling the regulatory interactions between transcription factors and their target genes will be essential to understand the role of transcription factors in driving developmental gene expression programs and to derive testable hypotheses for validation.

2.2 Introduction

Multicellular organisms develop from a single fertilized egg, guided by the genetic information encoded in the genome. Cell lineages diverge and form tissues and organs, based on the interplay between signaling pathways, biomechanical forces[83] and the regulation of gene expression programs[84]. While development is controlled on many levels, transcription regulation is crucial[85]. To better understand these regulatory principles in development and evolution, it is essential to construct informative models of gene regulation.

Transcription is regulated by transcription factors (TFs) within the chromatin context[86]. TFs bind the DNA either directly, mostly in a sequence-specific manner[87], or indirectly via other TFs[88]. They can recruit various other proteins, such as co-activators, RNA polymerase, chromatin remodelers and histone modifying enzymes, to remodel or stabilize the chromatin or to activate or repress transcription[89,90]. In metazoans, TFs form up to 8% of the known proteome[91,92], with DNA binding domains and affinities being highly conserved between metazoans[93–95]. They bind specific DNA motifs that are clustered in relatively short cis-regulatory elements (CREs) that can be categorized as promoters, enhancers and insulators[96]. The exact function of an element depends on the combination of bound transcription factors, which is influenced by motif specificity, distance between motifs and motif directionality[97–101]. Core regulatory modules and pathways involved in germ layer and axis formation are deeply conserved in metazoans[102].

A useful abstraction to study transcription regulation is a network of transcription factors and their target genes. This concept of a gene regulatory network (GRN) was introduced in 1969 by Roy Britten and Eric Davidson and later experimentally demonstrated in sea urchin embryos[24,25]. GRNs serve to predict the effect of transcription factor expression on gene transcription and to derive testable hypotheses for validation. More generally, they function to model cell type specification and differentiation in development as well as regulatory perturbations in disease. GRNs have been constructed, mostly based on experimental loss-of-function and gain-of-function studies, for a variety of developmental models. Examples include germ layer formation in echinoderms[103–105] and frogs[106–109], neural crest formation[110,111], the Drosophila gap gene network[112] and hematopoietic development[113–115]. However, experimental elucidation of a limited number of interactions is hard to scale. Regulatory interactions are highly context-specific[99,116] and most remain unknown[90,117].

Computational inference of genome-wide GRNs was made possible with the advent of expression microarrays. Expression levels between transcription factors and their target genes tend to correlate[118] and genes with similar mRNA expression patterns are more likely to be regulated by a common transcription factor[119,120]. This led to the conception of gene co-expression net-

works, where functional connections between genes are inferred by expression pattern similarity. WGCNA[27] and ARACNe[28] were among the first gene co-expression-based tools and remain popular. Presently, a multitude of GRN inference methods exists. Reviews on the technical details can be found here[96,121–123]. Recent advances in experimental and computational techniques means that GRN inference has progressed beyond simple co-expression. In this review, we will highlight three approaches that have the potential to significantly impact GRN modeling: (1) moving from one modality, gene expression, to multi-omics, (2) single cell sequencing for cell type-specific signal and (3) neural networks as flexible gene regulatory models (Fig. 2.1).

2.3 Multi-omics to capture gene regulation

Gene regulation by TFs is mediated through CREs including promoters and enhancers. By incorporating TF binding at enhancers, regulatory networks can be constrained by direct, causal relationships. Ideally, binding of TFs would be determined experimentally with chromatin immunoprecipitation followed by sequencing (ChIP-seq)[32] or related techniques[124–126]. While large compendia of TF binding profiles in different cell types have been collected for humans[117], this effort remains unfeasible for less well-studied organisms, including most developmental model systems. With sufficient training data, TF binding can be computationally imputed[89,127–141], however, this does not necessarily generalize across species[142]. As a result, most current approaches use relatively simple models that combine experimentally measured CRE activity with TF binding motifs to computationally predict TF binding.

Putative CREs and their activity can be mapped genome-wide using chromatin accessibility assays, such as DNase I hypersensitive sites sequencing (DNAse-seq)[143] and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[144]. The number of reads in an element can then be used as a measure for CRE activity in the experimental system[31]. ATAC-seq especially has been widely applied in developmental model systems, as it is experimentally relatively straightforward [57,64,65,92,111,145–149]. The chromatin environment can supply additional information on CRE location, function and activity. For instance, the transcriptional co-activator p300 is a histone acetyltransferase and can acytelate lysine 27 of histone H3 (H3K27ac). ChIP-seq using antibodies specific to p300 or H3K27ac can therefore identify active enhancers and promoters[13,150]. Other histone modifications that can be linked to CRE activity include H3K4me1 (enhancers) and H3K4me3 (promoters)[151].

CRE activity is determined by (in)direct binding of several TFs[152,153]. Therefore, characterizing TF binding at enhancers can identify their relative importance to the function of an enhancer. One approach to infer TF binding from genome-wide DNA accessibility is digital genomic footprinting[154], which has been used to directly infer GRNs[155]. However, sequence bias of the enzymes needs to be taken into account and TFs with more dynamic binding kinetics, such as some nuclear receptors, are not detected by footprint analysis[156–158]. Regardless, footprint analysis using cleavage bias correction can still be informative, especially in differential conditions[159,160]. A more routinely applied approach is to combine TF binding probabilities derived from TF motif scores with DNA accessibility. In some approaches, these are used as priors or constraints on network topology, where the network is inferred from gene expression measurements[161–163]. In alternative approaches, TF motif scores and accessibility are combined with RNA expression using regression models or co-variation of accessibility and expression[29,30,164–167].

Enhancers regulate transcription via context-dependent enhancer-promoter interactions[168], usually within a transcriptionally active domain[169]. Combined with TF binding data, these interactions allow for the inference of directed GRNs. Enhancer-promoter interactions can be identified experimentally with Chromatin Conformation Capture techniques[170–172], although this is still uncommon in non-model systems. Inferring interaction between enhancers to target genes is an active field

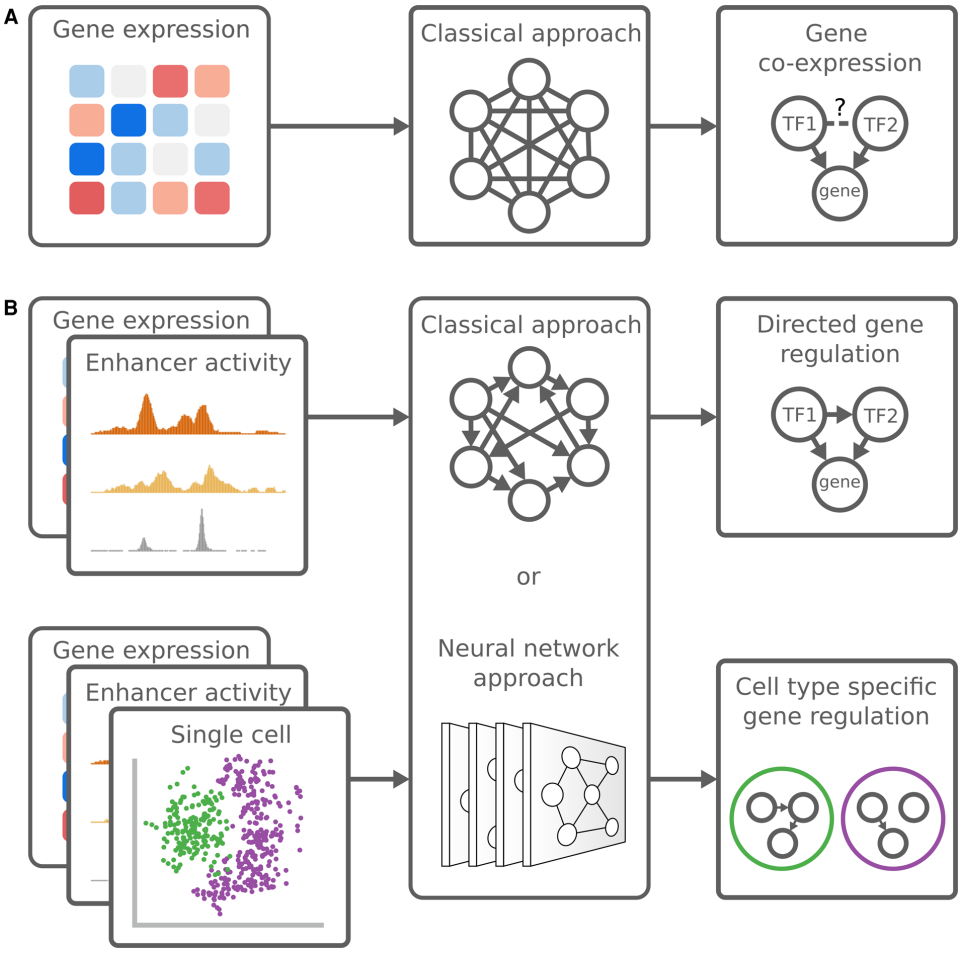


Figure 2.1: **Schematic overview of different gene regulatory network inference approaches.** **(A)** Classical approaches, *e.g.* correlation, regression or mutual information, can be applied on gene expression data to generate undirected co-expression networks. With prior knowledge about TFs the directionality between TF and target gene can be inferred, however, the directionality between two TFs cannot be established. **(B)** More recent approaches combine multiple types of genome-wide functional data (multi-omics), with either a classical approach or neural networks to identify directed gene regulatory networks. Single cell sequencing allows for the identification of cell type specific regulatory networks.

of research. The most commonly used heuristic is to link enhancers to the nearest gene. However, this heuristic is often still incorrect[173,174]. Accuracy can be improved by combining enhancer to gene distance with TF-target gene co-expression[175]. Finally, Activity-by-Contact based models significantly outperform the nearest gene heuristic by using enhancer to gene distance and enhancer activity[176].
By combining gene expression data with at least one source of enhancer data (*e.g.* accessibility or interaction data), directed regulatory networks may be inferred with significantly higher accuracy compared with traditional co-expression approaches[29,123,177]. Not only does the combined approach filter out spurious interactions and add causality, but it also reduces the biases introduced by singular approaches. Therefore, we believe that the use of multiple omics will become dominant in all modalities of GRN inference approaches.

2.4 Single cell sequencing for cell type specific regulation

Developmental transcription regulation has mainly been studied by either in situ hybridization[178], which maps the spatial distribution of gene expression of a small set of genes, or bulk gene expression studies[179]. The latter measures the whole transcriptome as a compound signal of all the different cells present in the sample. Single cell sequencing is a fast developing technique to measure the gene expression of individual cells separately, with newer techniques even capable of tagging cells to their spatial coordinates[180,181]. These techniques increase the number of measurements from a handful to several (tens to hundreds of) thousands. This substantial increase in data allows for interesting new ways of GRN inference, but poses new challenges as well.
The output of a single cell experiment generally consists of count tables containing several thousands of cells with low coverage, *e.g.* only a few thousand of measured transcripts per cell. The low coverage makes the detection of relations between lowly expressed genes difficult. Although it is possible to artificially increase the sequencing depth by simulation (imputation), this does not seem to improve GRN inference[182,183]. Furthermore, it is important to note that cells are repeated measures[184], meaning that the cells come from the same environmental and genetic background, which breaks most statistical assumptions. Computationally clustering related cells, called pseudobulk or meta-cells[185], and using their combined signal solves the issues of low coverage and repeated measures, and still yields cell-type specific signals.
Since fundamentally there are small differences between bulk and pseudobulk data, it is not uncommon to apply bulk GRN inference approaches, such as gene co-expression, ARACNE[28] and GENIE3[186], to pseudobulk data without much adjustment. The large number of cells, however, allows for specialized single cell GRN approaches. These include mutual information in combination with partial information decomposition[187], gene coexpression[188], self-organizing maps[189], or a combination of single cell RNA-seq and single cell ATAC-seq coexpression and/or bayesian ridge regression[190–192]. Other approaches first order cells by their inferred temporal ordering and then infer the gene-gene relations on this pseudotime, with the assumption that these orderings, also called trajectories, represent cell lineages[193]. Pseudotime can be estimated by simply following the first principal component, or finding the minimal spanning tree between clusters[194], where more advanced methods smoothen the tree[195,196]. A downside of these techniques is that they can not infer the directionality of the relationships. To computationally obtain this directionality, the ratio between spliced and unspliced transcripts per gene can be used as a proxy for whether or not a gene is actively transcribed. By applying this logic across all genes and all cells, one can infer a vector field of velocities of cells which then can be used to get a temporal cell ordering with a start and end[197,198]. These orderings then allow for inferring ordinary differential equations[199,200], Granger causality[201–203], boolean networks[204] or autoregressive models[205]. Most of these methods assume Gaussian noise for gene expression, even though transcription occurs in bursts[206,207], a

phenomenon that can only be captured on a single cell level. These dynamics can be modeled as a Markov process including transcriptional bursting and degradation[208]. Theoretically these mechanistic models could be great tools for hypothesis generation, but more work is needed to prove their practical usefulness. Even though the aforementioned GRN inference methods were developed for single cell data specifically, many fail to show consistent improvement over methods that were developed for bulk data, and are seemingly barely any better than purely random models[183,209,210]. Moreover, the added complexity and number of cells leads to computational scaling issues, with some methods taking several days to weeks to finish[183].

Single cell sequencing has the advantage that it disentangles the composite signal present in all biological tissues. The increased number of measurements allows for more complex GRN definitions and inference. Finally, it allows for the inference of fine grained temporal orderings necessary for GRN inference. Even though single cell GRN inference methods have not yet brought the improvements over bulk methods we hoped for, we still expect single cell GRN inference to become the new standard of the field.

2.5 Neural networks as flexible gene regulatory models

Computational inference of a GRN depends on a lot of implicit assumptions. For example, a common assumption is that the relationship between genes is additive, which means that the effect on a gene equals the sum of the effects of two regulators separately, but in reality, gene-gene relationships are more complex and for example can include multiplicative effects[211]. A type of model that requires little explicit specification about the possible relationships in the data, but automatically learns these relationships, is an Artificial Neural Network (ANN). ANNs have been successfully applied in a variety of settings, with famously complex problems such as protein folding[212], image recognition[213], and the board game Go[214]. The successes of ANNs in these unrelated fields show great promise for application in the field of gene regulatory inference.

Just like GRNs, ANNs consist of nodes and edges. Each edge multiplies the signal from the previous node to the next, and by applying a function to the sum of all the incoming edges the value in the next node is calculated. By adding multiple layers of nodes in between the in- and output nodes (this is where the term deep neural network comes from), a network is formed that is capable of learning more and more complex interactions. Learning happens by giving the model examples of input data and expected output, and based on this information the model iteratively updates (learns) its edge weights. After training, hypotheses can easily be tested by systematically querying the model for the predicted effect of certain changes. See[215] for an excellent review on the topic applied to genomics.

ANNs in genomics were first applied to predict the output of a genomic assay, for instance, histone modifications in a certain cell type, by using only the DNA sequence as input. Early models showed that convolutional neural networks are capable of predicting functional effects of noncoding variants from short (10–1000 bp) genomic sequences alone[216,217]. These types of models can be used to discover composite motifs and periodic binding[97]. Additionally, these models are capable of learning complex and distal biological relations, as increasing the input sequence to 131 kb still improves accuracy[218].

Whereas ANNs in genomics have mainly been popularized on sequence data, adoption for GRN inference has been relatively slow. Different approaches consist of self-organizing maps[189], variational autoencoders[219], extreme learning machines[220], or graph convolutional neural networks[221,222]. Even though these networks differ in architectural designs, they all report higher levels of accuracy over non-ANN approaches. However, without independent benchmark studies, it is hard to verify these results.

The main strength of ANNs is that they can approximate any continuous relationship in the data[223,224],

with the downside that large amounts of training data are required. This makes the combination of single cell sequencing and ANNs promising, as current single cell GRN inference approaches have scaling issues[183] and ANNs train relatively fast with the use of GPUs (graphics cards). Fundamentally, understanding how ANNs work is, however, much harder than understanding the classical models typically used for GRN inference. This causes ANNs to be met with skepticism and the persistent misconception that ANNs only function as a black box for predictions and its logic can not be interpreted[225]. We expect ANNs to become commonplace in the field of GRN inference due to their successes in other fields, ease of implementation with high-level programming libraries[226,227], and availability of sufficient training data due to single cell sequencing.

2.6 Discussion

Traditional GRNs, mostly based on gene co-expression, have so far served as a useful abstraction to understand regulatory dynamics in developmental systems. However, the way GRNs are currently derived suffers from two fundamental problems. First, the classic GRN that describes TF to target gene relations remains a simplified model and, by design, cannot properly reflect the full complexity of gene regulation. In addition, they are mostly based on mRNA expression as a measure of protein expression, even though this relation is not always linear[228]. In addition, any other types of regulation between transcript and protein product, such as mRNA degradation and post-translational modification, are usually ignored. Second, experiments generally have more features (*i.e.* genes measured) than samples which is also known as 'the curse of dimensionality'. In this underdetermined system, many different models can potentially fit to the data, and it is both practically and theoretically impossible to identify the correct model with certainty[229]. It then should not come as a surprise that benchmarks consistently demonstrate that the quality of the inferred GRNs is low[209,210,230–234]. Based on these observations it is clear that our current approach to infer GRN is not sustainable and design changes are needed. Ultimately, we expect the field to move towards GRNs inferred from neural networks trained on single cell multi-omics data.

Having said that, it is not enough to just naively apply single cell multi-omics ANNs. By adding more modalities, and making GRNs more complex, networks become even more underdetermined. This is why most multi-omics approaches use the new modalities to prune the possible TF-target gene relations, which actually reduces the degrees of freedom[29,188,191,192]. Moreover, one can use time-series data to further prune TF-target gene interactions[235], although time-series multi-omics GRN inference tools are still relatively uncommon[236–239]. In addition, computational methods such as regularization[240] and dropout[241] constrain the problem in such a way that you end up with the simplest fit out of likely possible fits. In addition, recent developments have made it possible to measure multiple modalities in the same cell, such as combined ATAC-seq and RNA-seq[242–244], which offers new, exciting opportunities for combining single cell sequencing with multi-omics data. ANNs, finally, have been made relatively easy to implement, can learn any type of interaction, and make no assumptions about the data (such as normality), which makes them extremely powerful GRN tools. However, it is not yet clear what the optimal architecture is for these networks, and interpreting the learned network from the ANN remains difficult.

GRN inference has become a data science, and it is time that we start treating it as such. Integrating multiple omics, several thousands of cells, and training complex machine learning models requires specialized knowledge. Common mistakes, such as treating cells from the same sample as independent[184], double dipping[245], and data leakage[140], can be avoided by proper data science training, but are unfortunately still common. Comparing the quality of GRN inference methods requires standardized benchmarks with multiple datasets, preferably a mix of experimental data and simulated data[208,246,247]. Simulated data has the advantage that the ground truth is known which makes benchmarking straightforward, but has the clear disadvantage that the quality of simulated data

depends on its assumptions and may actually not be representative of real biological data. The DREAM challenges[230,231] and BEELINE platform[210] are great examples, with predefined datasets and quality metrics. Only by measuring network accuracy in equal settings will it be possible to properly compare methods. It is however important to note that the goal of GRN inference is to gain mechanistic insights, as opposed to getting an optimal benchmark score, which makes fair comparison between approaches hard.

Altogether, we expect the field of transcription regulation in development to move towards increasingly multimodal GRN inference techniques to identify causal relations between genes. Single cell sequencing adds a cell type-specific precision which bulk sequencing can not provide. Finally, we expect a more widespread adoption of artificial neural networks as the field matures in technology and formal training, as these methods are inherently more powerful than previously used techniques.

2.7 Perspectives

- Gene regulatory networks have served as powerful models to understand gene regulatory programs in development and disease. Amongst others, these networks have been applied to model developmental patterning, to identify relevant transcription factors for cell fate transitions and to characterize deregulated transcriptional programs in disease.
- We believe three relatively recent developments will impact the computational inference of GRNs. The combination of multiple data modalities, such as RNA expression and DNA accessibility, help to constrain GRN topology and to predict directed networks. Single cell sequencing will become the de facto standard, as it allows for cell type-specific models and is able to provide the high number of measurements that are needed. Finally, artificial neural networks have the capability to create flexible and powerful models of gene regulation, which will benefit efficient and accurate GRN inference.
- The developments outlined above have the potential to significantly improve GRN inference. To fully exploit these approaches we have to implement common data science practices, and develop community-driven benchmarks to consistently measure the performance of different techniques.

Chapter 3

Seq2science: an End-to-End Workflow for Functional Genomics Analysis

Maarten van der Sande*, Siebren Frölich*, Tilman Schäfers, Jos Smits, Rebecca R. Snabel, Sybren Rinzema, Simon J. van Heeringen

3.1 Abstract

Sequencing databases contain enormous amounts of functional genomics data, making them an extensive resource for genome-scale analysis. Reanalyzing publicly available data, and integrating it with new, project-specific data sets, can be invaluable. With current technologies, genomic experiments have become feasible for virtually any species of interest. However, using and integrating this data comes with its challenges, such as standardized and reproducible analysis. Seq2science is a multi-purpose workflow that covers preprocessing, quality control, visualization, and analysis of functional genomics sequencing data. It facilitates the downloading of sequencing data from all major databases, including NCBI SRA, EBI ENA, DDBJ, GSA, and ENCODE. Furthermore, it automates the retrieval of any genome assembly available from Ensembl, NCBI, and UCSC. It has been tested on a variety of species, and includes diverse workflows such as ATAC-, RNA-, and ChIP-seq. It consists of both generic as well as advanced steps, such as differential gene expression or peak accessibility analysis and differential motif analysis. Seq2science is built on the Snakemake workflow language and thus can be run on a range of computing infrastructures. It is available at https://github.com/vanheeringen-lab/seq2science.

3.2 Introduction

The Sequence Read Archive (SRA) at NCBI currently holds over 36 petabytes of sequencing data, and this volume is growing rapidly[248]. Due to the flexibility of using sequencing as a readout, a large variety of different assays are available, such as RNA-sequencing (RNA-seq)[33] for gene expression quantification, Chromatin Immunoprecipitation (ChIP) sequencing[249] to profile DNA-bound proteins and assay for transposase-accessible chromatin with sequencing (ATAC-seq)[31] to determine DNA accessibility. This wealth of public data enables researchers to verify results, re-analyze data with novel techniques, and to combine and integrate datasets from different studies. However, processing these large amounts of data is a challenging and time-consuming task, even for researchers that are already familiar with high-throughput sequencing data processing details. To address this issue, various workflow systems have been developed, roughly categorized into three approaches: community-oriented workflow collections, multi-purpose workflows, and single-purpose workflows.

Community-oriented workflow collections enable multiple users to contribute workflows, as long as they conform to the established style and language of the community. Examples of community-based workflow collections include Galaxy[250], Snakemake-Workflows[251], and nf-core[252]. These collections offer the advantage of supporting a wide range of workflows and assays, with an active community providing support. Multi-purpose workflows facilitate multiple highly consistent workflows and typically provide a single entry point for users. Examples include Snakepipes[253], ENCODE pipelines[254], and CellRanger[255]. These workflows are designed to maintain consistency across different analyses, making it easier for users to learn and analyze the supported workflows. Single-purpose workflows are tailored to address specific problems, such as ARMOR for RNA-seq[256] and PEPATAC to analyze ATAC-seq[257]. The advantage of these workflows lies in their high level of specialization, focusing on particular tasks or analyses.

Although there is a choice of publicly available, published workflows, we found that these did not address all of our requirements. First, apart from some exceptions such as Galaxy, most workflows have not been specifically designed with public data in mind. This requires users to download and prepare the data in advance. While this is doable for small studies, it quickly becomes prohibitive for more large-scale analyses combining data from different studies. Additionally, many workflows have been primarily developed for, and tested on, human and/or mouse data. This limits their applicability to non-model species. It can be cumbersome to add new genomes and supporting gene

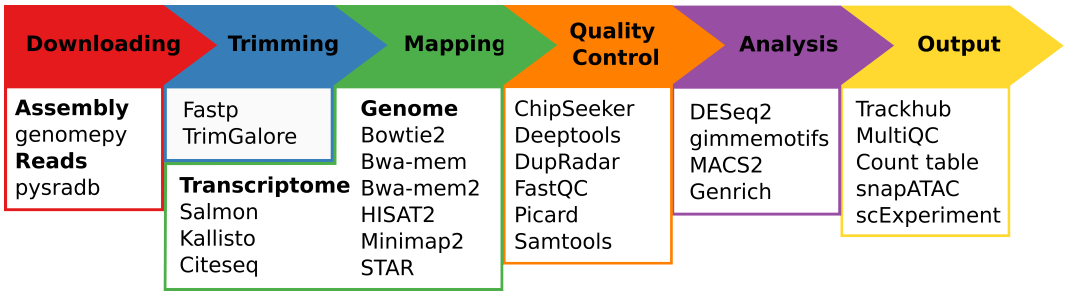


Figure 3.1: **Schematic overview of seq2science**. Seq2science can conceptually be split into six parts: downloading of samples and genome assembly, trimming of reads, transcriptome/genome mapping, quality control, initial analysis, and the final output. For each part the corresponding supported tools are listed.

and transcript annotation. Finally, workflows that do not actively encourage data exploration leave users susceptible to missing biases and failing to uncover concealed insights. Workflows should make sure to include a variety of quality control results and diagnostic plots, as it is essential to check the quality of the data. This includes, for instance, checking mapped data visually using a genome browser.

To address these limitations, we have developed seq2science, a multi-purpose workflow that supports virtually all public sequencing databases, multiple sequencing assays, and any species of interest. Seq2science is capable of automatically downloading genome assemblies and raw sequencing data from a range of sources. It supports multiple read trimmers, aligners, peak callers, and quantification methods, and generates an extensive quality report and a fully configured UCSC trackhub. It currently supports bulk ChIP-, ATAC- and RNA-seq, downloading of FASTQ files, and a generic genomic alignment workflow. Installation is easy through the Conda package manager, and extensive documentation is available online. Seq2science is designed to cater to both intermediate and advanced bioinformaticians. It serves as an accessible starting point for those with a basic grasp of bioinformatics concepts, thanks to its sensible default settings. Additionally, it offers a high degree of customization, making it appealing to advanced users who seek more tailored control over their analyses.

3.3 Methods

3.3.1 Implementation

Seq2science is built using Snakemake[258], a portable and open-source workflow system, which divides a workflow into independent modules called rules. Each rule includes a piece of code, its expected output, and optional input requirements. This design allows rules to be linked together, with the output of one rule serving as the input for another. Snakemake automatically determines the order in which rules need to be executed and distributes these tasks across available resources. To ensure reproducibility, most rules are assigned a specific virtual environment using the Conda ecosystem[259] that is automatically installed at the start of a run.

Seq2science requires two input files: a samples file and a configuration file. The samples file is a table containing a column of FASTQ files (or public identifiers for automatic downloading from any of the supported databases), a column with the assembly the FASTQ file must be mapped to, and optional additional metadata columns. Optional columns are the descriptive name of the samples, information about the relations between samples such as whether samples are technical and/or

biological replicates, and other such details. The configuration file is a YAML file with configurable parameters. These include whether to execute certain steps, the options to use when executing rules, and the directories where the output will be stored. Detailed explanations for the samples file and configuration file are available in the online documentation, and examples of these files are in the Supplemental Information and available at Zenodo (https://doi.org/10.5281/zenodo.8345208).

General overview

Regardless of the chosen workflow, when seq2science is executed, it checks for the local availability of the genome assembly and sequencing reads (in FASTQ format). If the genome assembly or the reads are not found locally, seq2science will download them. Once the raw data is obtained, the FASTQ files are prepared for alignment. This involves building a genome index and trimming the reads. The aligned reads are then duplicate-marked, filtered, and sorted according to the configuration settings. Nearly all workflows produce a set of indexed BAM (or optionally CRAM) files, an extensive quality control report, and a UCSC trackhub at this stage. The ATAC- and ChIP-seq workflows call peaks, and the output is stored and aggregated into a peak counts table. The RNA-seq workflow uses the specified quantifier to obtain raw gene counts and TPM tables. Optional differential gene expression, peak accessibility, and motif activity analyses are fully supported. Finally, a workflow explanation is generated with the parameters, version, and citation per tool which is also embedded in the QC report. For a schematic overview of the different steps and supported tools of seq2science see Figure 3.1.

Download-fastq

The download-fastq workflow can retrieve FASTQ files from various databases: the European Nucleotide Archive (ENA)[260], Gene Expression Omnibus (GEO)[261], the Sequence Read Archive (SRA)[262], the DNA Data Bank of Japan (DDBJ)[263], the Genome Sequence Archive (GSA)[264] and the ENCODE project[265], using their specific identifiers; ERX, ERR, GSM, SRX, SRR, DRX, DRR, CRX, ENCSR, or ENCFF numbers. The SRA, ENA, and DDBJ databases contain raw sequencing data that must be converted to FASTQ format, and they generally mirror each other in their content. EBI ENA, GSA, and ENCODE, however, store FASTQ files directly, so if a sample on the SRA or DDBJ is found to be mirrored on ENA by pysradb[266], seq2science will directly download FASTQ files from there, optionally using Aspera Connect (ascp), which is a high-speed transfer protocol developed by IBM. If the sample is not directly available in FASTQ format, seq2science uses the sra-toolkit[262] to download the raw data and parallel-fastq-dump, a parallelized version of fastq-dump, to convert the data to FASTQ files.

Alignment

The alignment workflow in seq2science processes FASTQ files that are either already present on the device or are automatically obtained using the download-fastq workflow. If necessary, the workflow will also download a genome assembly FASTA file and corresponding gene annotation file using genomepy[267]. The FASTQ files are trimmed for quality and adapters using either TrimGalore[268] or fastp[269], as specified by the user. The trimmed FASTQ files are then aligned to the genome using the selected mapper, such as bowtie2[270], bwa-mem[271], bwa-mem2[272], HISAT2[273], minimap2[274], or STAR[275]. The resulting BAM file is filtered based on criteria such as MAPQ value, duplicate status, or alignment in the ENCODE blacklist[276]. The filtered BAM file is then converted into a bigWig file and prepared for visualization in a UCSC trackhub[277] or, when the genome assembly is not hosted by UCSC, as an assembly hub. The filtered BAM file can optionally be stored as a CRAM file to save disk space. Quality checks are performed throughout the process using FastQC[278], samtools[279], Picard[280], and deepTools[281], and the results are summarized in a MultiQC report[282].

ATAC- & ChIP-seq

The ATAC- and ChIP-seq workflows are identical in implementation, except that they are initialized with different default settings. They internally use the same rules as the alignment and download-fastq workflows, which means that they either start by downloading FASTQ files, or analyze files that are already present. For the ATAC-seq workflow the aligned reads are by default Tn5 shifted[283] and only reads with a maximum template length of 150 base pairs are kept. Peak calling is done on the filtered BAM files with either MACS2[37] or genrich[284], with optionally specified control samples. Biological replicates can be combined either with the internal Fisher's method of either tool, or with IDR[285]. Peaks between conditions are combined when they fall within a certain range of each other with GimmeMotifs[139] and a count table is made for all samples based on the number of reads in peaks. Optional differential peak analysis with DESeq2[40], or differential motif analysis with GimmeMotifs[139] can be performed if selected. When doing a differential motif analysis, seq2science automatically converts the transcription factor gene names in the motif database into the (orthologous) gene names of the assembly used when a genome annotation is available. Additional QC is collected by deepTools[281], ChIPseeker[286], and Subread[287]. See file S14 for a directed acyclic graph of all the steps involved with the ATAC- and ChIP-seq workflows.

RNA-seq

The RNA-seq workflow begins with the acquisition and processing of FASTQ files as described in the alignment and download-fastq workflows. Gene expression quantification can be based on either genomic alignment or transcript quantification, depending on the settings. For genomic alignment, reads are aligned to the genome with a splice-aware aligner (STAR[275] or HISAT2[273]). The output BAM files are filtered and have their duplicate reads marked. Gene expression quantification is then performed by assigning reads to genes, using HTSeq[288] or featureCounts[287]. Ambiguous transcripts are minimized by providing the gene counting tools with the strandedness of each sample, which is inferred using RSeQC[289]. Additional gene-based TPM expression levels are generated using genomepy[267], based on longest transcript lengths. For the gene quantification approach, transcript abundances are quantified using Salmon[290] in mapping-based mode. To improve mapping accuracy, decoy sequences are generated as suggested by the Salmon documentation. The transcript abundances are aggregated to gene level using pytxi[291] or tximeta[292] and additionally converted to gene counts using genomepy[267].
Independent of the configured gene expression quantification approach, the workflow supports differential gene expression analysis with DESeq2[40] with batch effect correction to integrate (multiple) datasets, and can prepare an exon count table for downstream use with DEXSeq[293]. Strand-specific bigWig files are generated for visualization in a UCSC trackhub. The trackhub configuration file is updated with strand information for ease of use. Additional quality control metrics specific to RNA-seq are obtained from DESeq2[40] and dupRadar[294]. See Supplemental File S15 for a directed acyclic graph of all the steps involved with the RNA-seq workflow.

3.4 Use Cases

To briefly illustrate some of the capabilities of seq2science, we show how to use it to download publicly available FASTQ files, and finally show three example processing runs using public data.

3.4.1 Downloading FASTQ files

Downloading FASTQ files with seq2science has been made extremely easy. After installation of seq2science (see Supplemental File S1), all that needs to be provided is a tab-separated file with

Table 3.1: Example samples file for the download-fastq workflow.

sample
ERX000401
ERR022487
GSM2811115
SRX257149
SRR800037
DRX029591
DRR032791
CRX269079
ENCSR535GFO
ENCFF172MDS

database identifiers. Seq2science currently supports identifiers for ENA, GEO, SRA, DDBJ, GSA, and ENCODE. For this example, we will download one FASTQ file from each database. To get started we need to initialize seq2science in the current directory:

```
seq2science init download-fastq
```

After this initialization, we get a config and a samples file. The config file is practically empty as there is not much to configure for the download-fastq workflow. For this workflow, the relevant option is the directory where we want the samples to be downloaded.
We then edit the samples file and add the experiment (or run identifier) of each sample (Table 3.1). Here we show a mixture of paired-end and single-end samples to highlight seq2science's ability to work with both data types. To download these samples all we now have to do is run seq2science:

```
seq2science run download-fastq --cores 8
```

Seq2science will now start downloading our samples. Samples hosted on GSA, ENCODE, and ENA are downloaded directly as FASTQ file, whilst the samples not on ENA, ENCODE, or GSA first get downloaded as an intermediate SRA file which then gets converted into a FASTQ file. Even though we have specified the DRX and DRR samples by their DDBJ identifier, seq2science finds their ENA mirror if it exists and directly downloads those to save computational resources. At the end of this run, we end up with the corresponding FASTQ files for all ten samples.

3.4.2 A map of cis-regulatory elements in zebrafish

In this example, we reproduced part of the analysis of the study "A map of cis-regulatory elements and 3D genome structures in zebrafish"[295]. Yang et al. studied zebrafish chromosome conservation with a wide array of different functional genomics techniques. Here, we focused on the analysis of cis-regulatory elements and gene expression in different embryonic tissues. Using the default tools incorporated into seq2science, we downloaded the raw data for the different assays from the SRA, aligned these data to the zebrafish genome (Figure 3.2A) and performed a differential transcription factor motif enrichment analysis on the ATAC-seq data.
After running seq2science, we obtained a set of aligned BAM files, narrowPeak files, a count table, a trackhub, and a quality control report (see Supplemental Files S2-7 for the configuration, samples, and QC report). Figure 3.2A shows the alignment of reads visualized using the UCSC trackhub that was created by seq2science. The figure shows histone modification ChIP-seq data (H3K4me3, H3K27ac, H3K9me2 and H3K9me3), ATAC-seq, and RNA-seq for three different tissues

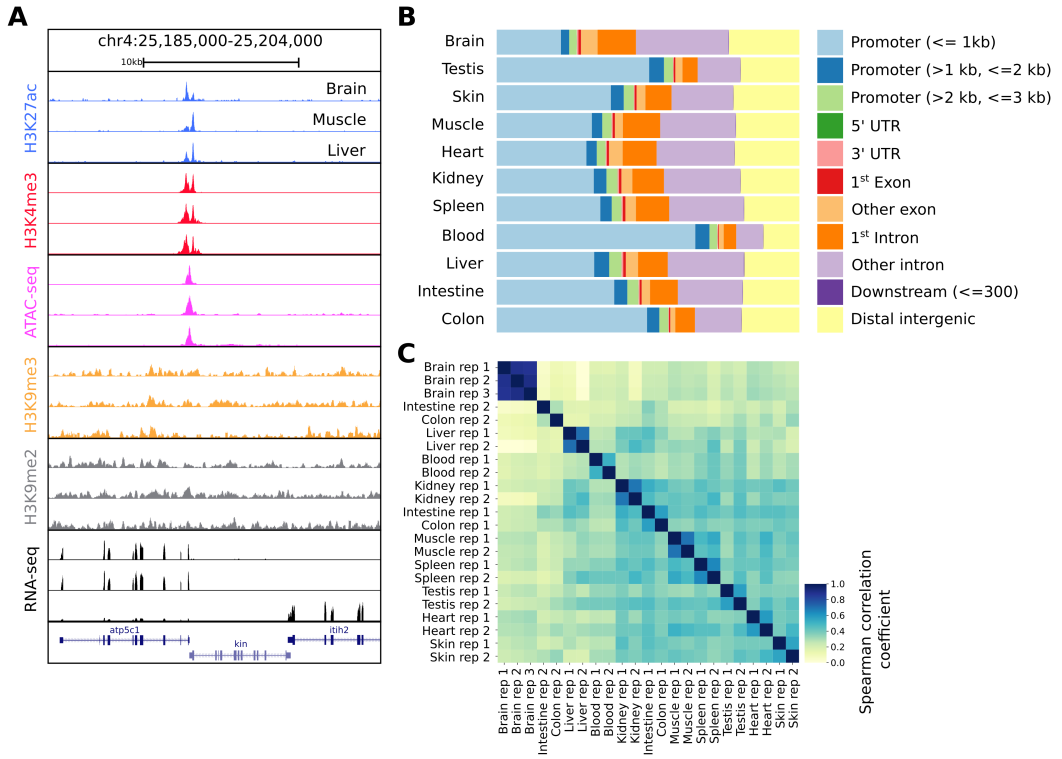
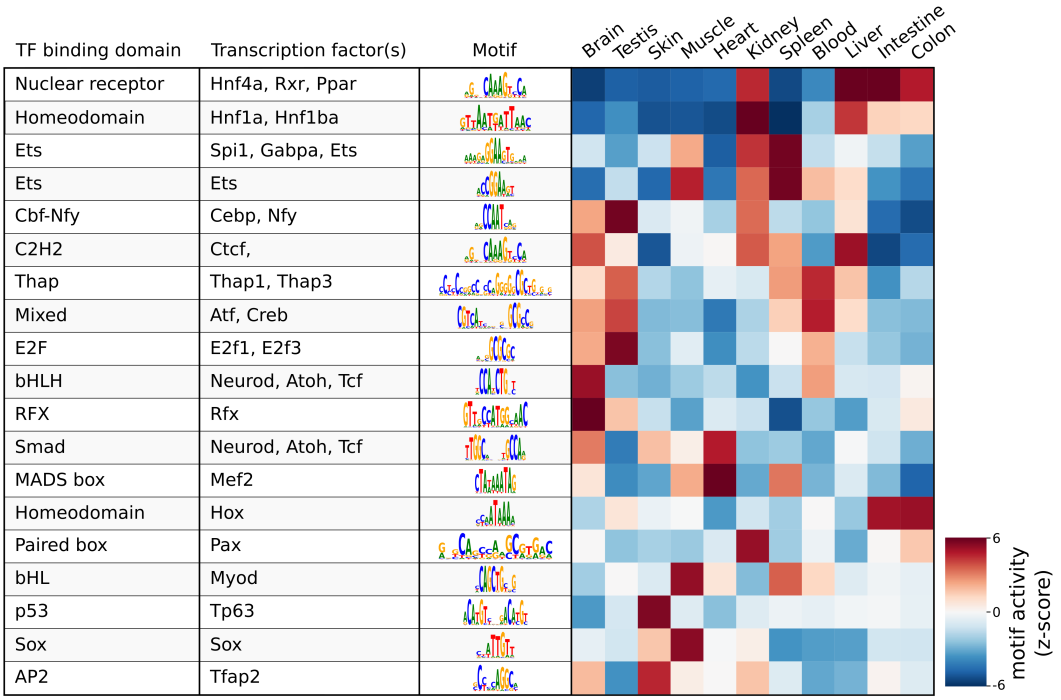


Figure 3.2: **Snapshot of the UCSC trackhub and quality figures of seq2science**. **(A)** Fully configured UCSC trackhub generated by seq2science which highlights some of the supported assays. **(B)** The fraction of ATAC-seq peaks predicted in each tissue and their genomic distribution, visualized by ChIPseeker. **(C)** Pairwise Spearman correlation coefficients of all the samples.



TF binding domain	Transcription factor(s)	Motif
Nuclear receptor	Hnf4a, Rxr, Ppar	
Homeodomain	Hnf1a, Hnf1ba	
Ets	Spi1, Gabpa, Ets	
Ets	Ets	
Cbf-Nfy	Cebp, Nfy	
C2H2	Ctcf,	
Thap	Thap1, Thap3	
Mixed	Atf, Creb	
E2F	E2f1, E2f3	
bHLH	Neurod, Atoh, Tcf	
RFX	Rfx	
Smad	Neurod, Atoh, Tcf	
MADS box	Mef2	
Homeodomain	Hox	
Paired box	Pax	
bHL	Myod	
p53	Tp63	
Sox	Sox	
AP2	Tfap2	

Figure 3.3: **Result of the differential motif analysis by seq2science between ATAC-seq peaks of different tissues of zebrafish**[295]. Per tissue, the top two most differentially enriched motifs have been selected, with automatically inferred orthologs. The transcription factor names were manually curated for clarity. The full table with the motif analysis results is included in the quality report (Supplemental File S7).

(brain, muscle and liver) on a region of chromosome 4. Figure 3.2B-C show a selection of diagnostic plots from the quality report. Figure 3.2B demonstrates the ratio of ATAC-seq peaks annotated to different genomic regions, created by ChIPseeker. Figure 3.2C shows the correlation of reads in ATAC-seq peaks between samples. This figure illustrates the value of extensive quality control, as it shows two samples where the replicates do not cluster together (colon and intestine). This would warrant further investigation, as it could indicate a potential sample swap. In our experience, this occurs frequently with samples downloaded from public databases, which may be due to a sample swap in the original analysis, or during submission to the repository.
As a demonstration of a more high-level analysis, Figure 3.3 shows the result of the differential motif analysis. Here, seq2science used GimmeMotifs to automatically convert the transcription factors in the motif database into the orthologous zebrafish genes. This automatic assignment means that motif analysis can also be used for non-model species that do not have a readily available motif annotation. Figure 3.3 shows the top motifs per tissue, based on the z-score. The complete table is part of the seq2science output report (see Supplemental File S7). In general, this unsupervised motif analysis recapitulates many of the findings of Yang et al., such as RFX and bHLH (Neurod, Atoh1) motifs enrichment in the brain and Hnf4a enrichment in the liver, colon, and intestine. Additionally, the GimmeMotifs analysis assigns Tp63 as a transcription factor enriched in the skin, as well as Tfap2, which are well-known regulators of epidermal development[296,297].

3.4.3 The regulatory landscape of whole-body regeneration in the three-banded panther worm *Hofstenia miasma*

The article "Acoel genome reveals the regulatory landscape of whole-body regeneration" by Gehrke et al. analyzed the gene expression and chromatin accessibility of the Acoel worm *Hofstenia miamia*[298] during regeneration. The paper contains ATAC-seq and RNA-seq time-series data of the response to amputation and during whole-body regeneration. These data, and the *Hofstenia miasma* genome are available from NCBI and consequently, seq2science can be applied to re-analyze this data with ease. The configuration files to reproduce the seq2science analysis and the complete output and quality control report are provided as Supplemental Information (Supplemental Files S8-13).
The seq2science ATAC-seq workflow generated a consensus peak set based on the union of peaks from all time points, together with a count table with the read quantification per time point. See Supplemental File S10 for the ATAC-seq QC report. We clustered the count table to visualize the temporal accessibility patterns using log-transformed and z-score normalized read counts (Figure 3.4A). The figure nicely recapitulates the original work, with most ATAC-seq peaks showing strong signal at either 0 hours or 48 hours post-amputation (hpa).
To demonstrate the RNA-seq workflow, we performed a differential expression analysis of 6 hpa versus 0 hpa using DESeq2 (see Supplemental File S13 for the complete output). The results are visualized as a volcano plot in Figure 3.4B. As reported, the *Hofstenia miamia* EGR ortholog is the most significant differentially expressed gene, followed by the RUNX homolog *runt*, which is involved in wound healing and regeneration. The role of *egr* is further confirmed by the differential motif accessibility analysis, performed on the ATAC-seq peaks in the seq2science ATAC-seq workflow (Figure 3.4C). As stated earlier, GimmeMotifs automatically assigns the transcription factors in the database to the orthologous genes in our assembly. As expected, the top enriched motif is EGR, which shows a high z-score during the post-amputation time-series, from 3 to 24 hpa. After the knockdown of EGR using RNAi, this is the most depleted motif with a strong negative z-score. In conclusion, this use case reproduced the main findings from the the original paper, using the default tools incorporated in the RNA-seq and ATAC-seq workflows. Additionally, it demonstrates that seq2science is also easily applicable to non-model species.

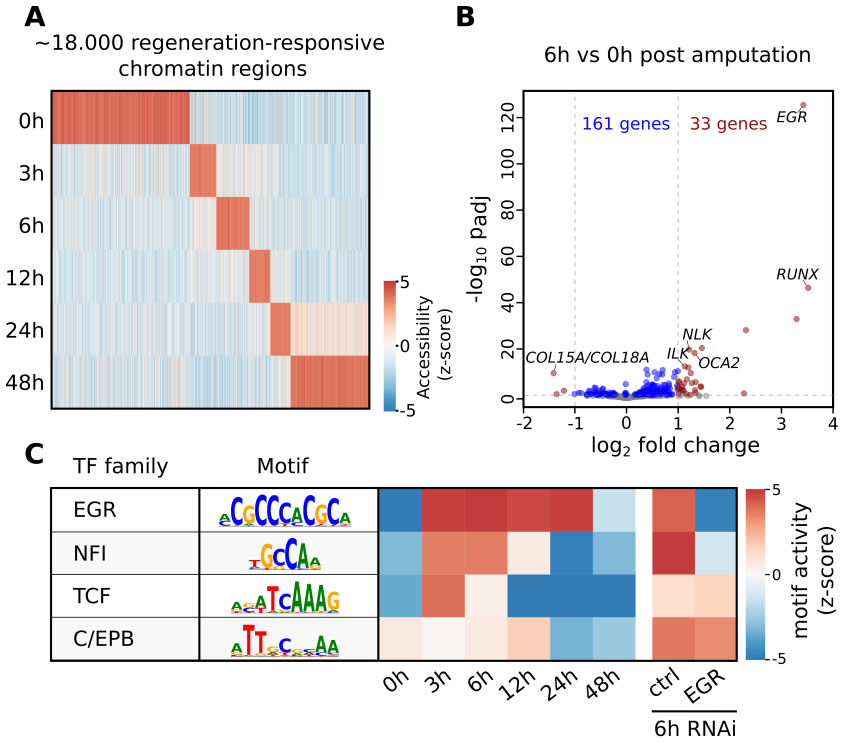


Figure 3.4: **Summary of selected results from the re-analysis of the RNA-seq and ATAC-seq regeneration time series from Gehrke et al**. **(A)** Heatmap of *Hofstenia miamia* chromatin accessibility during tail regeneration post amputation. ATAC-seq read counts provided by seq2science were log-transformed, and columns were normalized using the z-score. **(B)** Differential gene expression during tail regeneration. The X axis shows the log2 fold change of 6hpa vs 0hpa; the Y axis shows the -log10 transformed p-value. Significantly changed genes (padj ≤ 0.05; log2 fold change ≥ 1) are marked in red. Top results were labeled with human ortholog gene families. **(C)** Motif activity prediction during tail regeneration. The top four motifs, as identified by GimmeMotifs, are shown with the activity z-score predicted by gimme maelstrom. Both Gehrke et al. and our seq2science analysis identified the EGR and NFI motifs as the top differentially active motifs in the knockdown experiment.

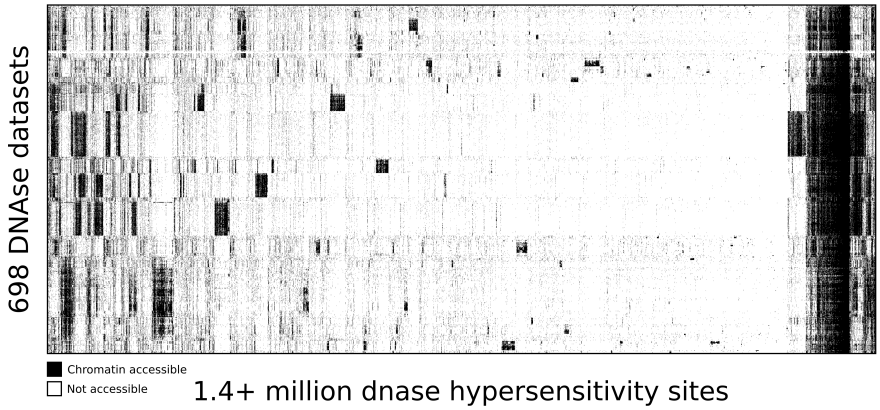


Figure 3.5: **DNA accessibility at 1.4 million consensus DHSs assayed across 698 samples encapsulated in a visually compressed DHS-by-biosample matrix**. Recurring accessibility patterns indicate extensive sharing across cell contexts. Chromatin accessibility is defined as more than 1 count per million reads.

3.4.4 Map of human DNase I hypersensitive sites

To highlight the ability of seq2science to scale with large data sets we performed a re-analysis of "Index and biological spectrum of human DNase I hypersensitive sites"[299]. This paper analyzed 733 human DNAse I samples across different tissues. Of the 733 samples they reported, we were able to find 698 in the SRA database. To accommodate the DNAse I assay, which is not supported by default, we used the high degree of customizability of seq2science and adapted the default ATAC-seq workflow by turning the tn5_shift flag off. The final output consists of over 1.7TB of sorted BAM files, with 1,404,721 peaks in the consensus peak set between these samples. Figure 3.5 shows the clustered output of the final count table, where the distinction between accessible and not accessible is made based on whether there is more than 1 count per million reads or not.

3.5 Conclusion

Seq2science facilitates reproducible preprocessing of high-throughput sequencing data of different assays through a unified setup. The tool is integrated with all major public sequence and assembly databases, and outputs extensive quality control and processed results to speed-up analysis. Seq2science requires minimal user input to get started, but offers a high degree of customizability. Each workflow and its configurable options are fully explained in the online documentation. The output from seq2science is reproducible and directly ready for analysis.

3.6 Acknowledgements

We are grateful to the open source bioinformatics community for their support in the development of seq2science. Special thanks to the bioconda and conda-forge teams for their help with packaging. More specifically we want to thank Saket Choudhary, Johannes Köster, Tao Liu, Devon Ryan, and Phil Ewels for their assistance with Pysradb, Snakemake, MACS2, Bioconda, and MultiQC respectively.

Chapter 4

Quantitative Genomic Comparisons of Embryogenesis Between Species do not Support a Generalized Phylotypic Stage

Maarten van der Sande, Gert Jan C. Veenstra, Simon J. van Heeringen

4.1 Abstract

The phylotypic stage is the period during embryonic development that is most conserved between species of the same phylum. Whereas evolutionary conservation in the phylotypic stage has historically been defined by qualitative morphological descriptions, more recently it has been explored and analyzed in the context of quantitative molecular similarity. Here we explore the concept of the molecular phylotypic stage in the context of comparative analyses, by focusing on the predictions of the different evolutionary developmental models. We argue that these models are not explicitly defined, and because of this ambiguity are unfalsifiable. As such we advocate for more explicit definitions and expectations of evolutionary developmental models and recommend the use of within-species, within-phylum, and between-phylum controls. We apply these controls to four recent molecular studies of the phylotypic stage and show that these analyses do not support the conclusions that are drawn from them. In our re-analyses, most between-species patterns seem to be caused by within-species effects. A notable exception is the inverse hourglass model, also known as the mid-developmental transition, which is caused by a statistical artifact of normalization. To our knowledge, there is no comparative molecular study that includes all three controls, and as such we see no molecular support for a well-defined phylotypic stage or its related hourglass and inverse hourglass models.

4.2 Introduction

Embryonic development is a complex and highly orchestrated process that begins with a single fertilized egg and culminates in the formation of a multicellular organism with a defined body plan and specialized organs. Even though both the eggs and adults of related species can be morphologically quite different, a seemingly remarkable period of similarity occurs in early development. This period of similarity is most easily observed for related species and is therefore known as the phylotypic stage (or period[300]). Over the years different models and explanations for its existence have been proposed[301–303].

The idea of a morphologically similar embryonic stage between related species dates back to Aristotle[304], but was formalized and popularized by Karl Ernst von Baer and Ernst Haeckel[49,53]. In the early 1800s, von Baer formulated his four laws of embryology based on post-gastrulation embryos. His first law states that *the more general characters of a large group appear earlier in the embryo than the more special characters.* This means that as an embryo develops, it first develops its oldest phylum-specific features, to then respectively develop its class, order, family, and species-specific features. Simply put, embryos of related species become increasingly diverse as development proceeds. Haeckel promoted a more radical view. Expanding on the work of Etienne Serres and Johann Friedrich Meckel, Haeckel related evolutionary history to developmental conservation in the recapitulation theory, popularized with the phrase *ontogeny recapitulates phylogeny.* The recapitulation theory suggests that embryonic development is a replay of evolutionary history, where an embryo consecutively develops from adult stages of ancestral species to adult stages of descendants. In its strongest form, the recapitulation theory has been discredited, as development is not a repetition of evolution[305]. Nonetheless, Haeckel's observations of similarities between embryos of different vertebrate classes, have, together with von Baers' laws, formed the basis for the current models of evolutionary development (evo-devo).

The notion of similar embryonic stages between related species has led to two competing models of evolutionary development. The "funnel" or early-conservation model is closely linked to von Baer's first law, predicting the highest morphological similarity early in development. The hourglass model instead is based on the longstanding notion of a similar stage during mid-embryogenesis[306], but the first to describe it as an hourglass was Paul Medawar in 1954[54]. Medawar argues that some-

where mid-embryogenesis is the most morphologically conserved stage for vertebrates. This stage corresponds to Haeckel's phylotypic stage, but different from Haeckel's recapitulation theory, different species are thought to be more diverse both before and after the mid-embryogenesis state. More recently, these ideas have been examined with molecular genomic data, leading to a generalization of the hourglass model across phyla and kingdoms (insects[62], plants[307] and fungi[308]), where each phylum now is expected to have its separate stage of maximum similarity during mid-embryogenesis. The phylotypic stage, initially called the *phyletic stage*, refers to the point of maximum similarity in these models[309,310]. More recently an inverse hourglass model has been proposed for comparisons between phyla, where specifically the beginning and end of embryonic development seem conserved, and the least molecular similarity is seen at the phylotypic stage[56]. Whereas the phylotypic stage has originally been defined based purely on qualitative morphological descriptions alone, the definition has recently been interpreted in terms of conserved patterns of gene expression. A popular addition is the idea that HOX genes are the master regulators of the phylotypic stage for vertebrate development[55]. Moreover, similarity in specific morphological features has been generalized to mean similarity across all genes or genomic features such as regulatory elements. These quantitative comparisons can roughly be divided into two distinct methodological approaches; (i) calculating a conservation metric for a single time series for each time point, where conservation metrics include the average evolutionary age of transcripts[60], the gene mutation rate index[63,307], embryonic lethality[311], the relationship between timing of DNA accessibility and its evolutionary age[64], and the variance between replicates[312,313]. These results are usually visualized as a line graph where the x-axis represents embryonic development, and the y-axis the conservation metric. The second approach (ii) compares orthologous features between time points of two different species directly. Orthologous features that have been compared are cell type proportions[58], gene expression similarity[56,57,61,62,314,315], and regulatory DNA accessibility similarity[59,316]. The comparison between two time-series is usually visualized as a heat map, where each axis represents the embryonic development of a species, and the color represents (dis)similarity (Fig. 4.1). Both of these quantitative approaches have been used to study the phylotypic stage, but the second approach, comparing two time-series directly, is the focus of this study.

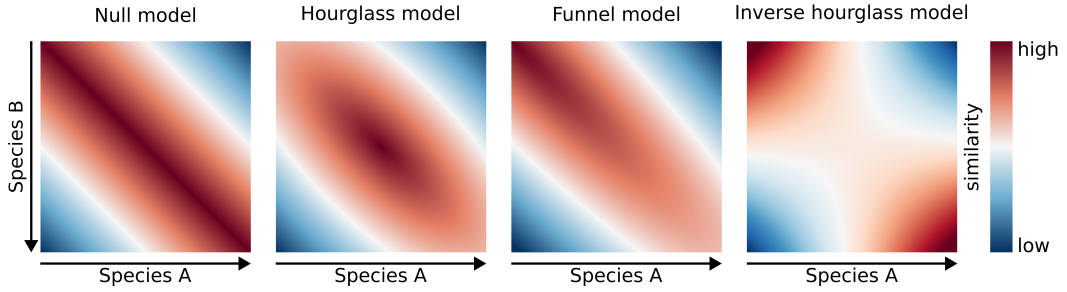


Figure 4.1: **Examples of developmental similarity along embryonic development for different comparative models**. These figures are usually ordered from top-left to bottom-right, where the x-axis represents embryonic development for species A, and the y-axis represents embryonic development for species B. The null model represents the case where there is no developmental stage of higher or lower similarity between the two species. The hourglass model predicts a point of maximum similarity somewhere mid-embryogenesis for species of the same phylum, and the funnel model instead predicts the highest similarity at the start of embryonic development. Finally, the inverse hourglass model predicts that for comparisons between-phyla, a point exists in mid-development that is the least conserved.

While quantitative studies are seen as offering an unbiased approach to studying the phylotypic stage, their results can be influenced by experimental design, including the analysis of the data. This, in turn, highlights the importance of incorporating appropriate controls. To give some examples; based on morphological timings, both an hourglass[317] and an inverse hourglass[318] have been found. Based on RNA-seq data, Chan *et al.* found an hourglass pattern, but with an identical analysis on microarray data, they found no temporal conservation pattern[319]. Piasecka *et al.* find that the observed conservation pattern is highly dependent on the metric used. In their work, genes are split into groups based on their maximum expression. There is no difference between the groups based on the d_n/d_s-ratio. Gene duplications and new gene introductions are least occurring in early conservation, but genes expressed in mid-development have the most highly conserved non-coding elements. However, all gene-based properties coherently show the least conservation for the latest stages[63]. Moreover, Piasecka *et al.* show that a popular similarity metric, the transcriptomic age index[60], expresses completely different conservation patterns based on whether or not the data has been log-transformed[63]. Finally, Levin *et al.* describe a universal between-phyla inverse hourglass[56], whilst Perez-Posada *et al.* with a similar methodology instead report an hourglass for their comparison between deuterostomes and the chordate amphioxus[314]. Moreover, in an independent re-analysis of the work of Levin *et al.*, their inverse hourglass was found not to be statistically significant[320].

The dependence on experimental design when studying the phylotypic stage is problematic and highlights the need for elaborate controls and clearly defined expectations. The phylotypic stage, the hourglass model, and the funnel model, however, do not pose explicit expectations on the similarity of embryos from different phyla. Yet, as the word phylotypic is a compound word of phylum and typical, it is suggestive of features conserved within a phylum but not between phyla. This implies that the point of maximum (molecular) similarity between phyla does not coincide with the phylotypic stages of the phyla involved. Levin *et al.*[56] use this implication in their inverse hourglass model as a new definition to distinguish phyla, where they note that embryonic dissimilarity is largest between species from different phyla at their respective phylotypic stages. Is it thus safe to assume that the point of maximum similarity between species from different phyla does not occur at their respective phylotypic stages? Similarly, what is expected if we were to compare the embryonic development of a species against itself? Transcriptomic variance between replicates, for instance, is lowest mid-development (Fig. S4.1), something that is used as an argument for the hourglass model[312,313]. The reasoning here is that gene regulation is most tightly regulated mid-development and that this translates to low transcriptomic variance between replicates. However, this changing variance over time can affect the similarity for direct comparisons between species. This begs the question of whether high similarity between species is an effect of a high similarity within species. Or would we expect that the phylotypic stage has a higher similarity between species of the same phylum even when corrected for within-species variance?

To address these ambiguities in the description of the phylotypic stage we have re-analyzed four recent molecular studies. By explicitly testing the within-species, within-phylum, and between-phyla assumptions about the phylotypic stage in these studies, we expose crucial flaws in the interpretation of the results from these analyses.

4.3 Results

4.3.1 The phylotypic stage between zebrafish and frog is a superimposition of within-species effects

In the paper *Amphioxus functional genomics and the origins of vertebrate gene regulation*[57] Marlétaz *et al.* investigated the similarity of orthologous gene expression in several chordates (*Branchiostoma*

lanceolatum, *Danio rerio*, *Gallus gallus*, *Oryzias latipes*, and *Xenopus tropicalis*), and consistently find a point of maximum similarity during mid-embryogenesis, in support of the hourglass model. All comparisons are within the same chordate phylum, but between-phyla or within-species comparisons are missing. In our re-analysis, we focus specifically on the comparison between *D. rerio* and *X. tropicalis*. By analyzing the original data with a similar methodology, we reproduce their result of a point of maximum similarity mid-embryogenesis. Furthermore, we incorporate within-species controls by comparing the *D. rerio* time series from our study with a similar time series of *D. rerio* from a different study[321], and apply the same approach for *X. tropicalis* using data from[316]. We show that the point of maximum similarity between *D. rerio* and *X. tropicalis* corresponds to the point of maximum similarity within each species. Finally, the current analysis cannot distinguish within-species effects from between-species effects, and as such no hard claims can be made about temporal conservation between these species.

Figure 4.2B shows the pairwise similarity between all sampled stages of *Danio rerio* and *Xenopus tropicalis*, where similarity is based on the Jensen-Shannon distance (JSD) of the gene counts (TPMs). The JSD is a distance metric where a high value means low similarity between distributions and vice versa. The JSD follows an hourglass pattern with the point of maximum similarity at 20 hours post-fertilization for *Danio rerio* and at 30-32 hours post-fertilization for *Xenopus tropicalis*, marked by a black star. Our result is visually similar to the comparison by Marlétaz *et al.* and the actual point of maximum similarity is adjacent to theirs, marked by a black pentagon. Figure 4.2A and C show the related within-species comparisons, where the time-series of *X. tropicalis* and *D. rerio* are compared to the time-series of independent studies[316,321]. Both these within-species comparisons show an hourglass-like pattern, with their point of maximum similarity during mid-embryogenesis. Note how the points of maximum similarity within-species closely match with the points of maximum similarity between-species. Similarly, figure S4.2 shows all the within-species comparisons against itself, where we find that there is a high self-similarity at the point of maximum similarity between species.

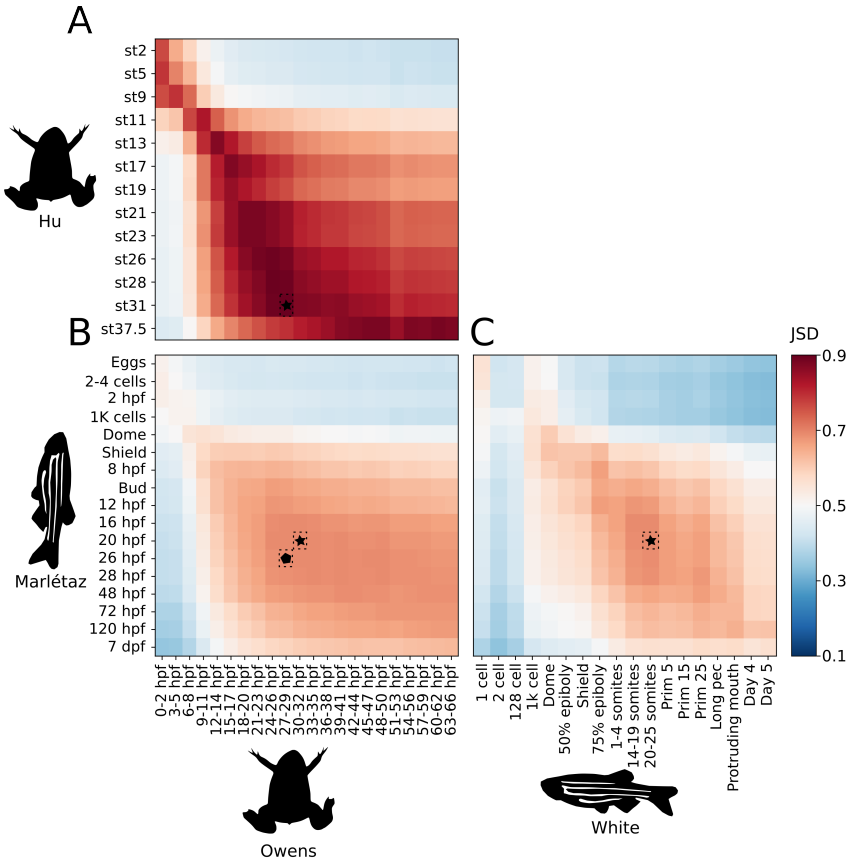


Figure 4.2: **Within-species transcriptomic similarities translate to between-species transcriptomic similarities**. The heatmap of pairwise Jensen-Shannon distances between **(A)** two *Xenopus tropicalis* time series, **(B)** *Xenopus tropicalis* and *Danio rerio*, and **(C)** two *Danio rerio* time series. The point of lowest Jensen-Shannon distance is marked in each comparison by a black star, and the black pentagon represents the point of maximum similarity between *Xenopus tropicalis* and *Danio rerio* in the original study by Marlétaz et al.[57]. The between-species comparison appears to be a superimposition of the within-species effects.

Our re-analysis reveals that an hourglass pattern is already present when comparing the transcriptomes of *D. rerio* and *X. tropicalis* from different studies within each species independently. Consequently, any additional comparison made between such time series is susceptible to this effect. Based on our re-analysis, the between-species comparison can be explained as a combination of the two within-species patterns alone. It is still possible that even when correcting for within-species effects, the similarity between *D. rerio* and *X. tropicalis* is highest at the phylotypic stage. However, without explicitly correcting for the within-species effects we cannot definitively determine whether certain embryonic stages are more or less conserved between species

4.3.2 The mid-developmental transition is a statistical artifact of gene standardization

In the paper *The mid-developmental transition and the evolution of animal body plans*[56] Levin *et al.* compared the correlation coefficient of the expression of orthologous genes over time between ten

species from different phyla. They found that most species-species comparisons display an inverse hourglass, with a high similarity early and late during development, but a period of low similarity mid-development. They refer to this period as the *mid-developmental transition*. They note that the mid-developmental transition between phyla seems to correspond with each species' phylotypic stage. They then suggest that this pattern could be used to distinguish different phyla. In this re-analysis, we reproduce the finding of a between-phyla inverse hourglass. As suggested by Hejnol *et al.*[322] we include a within-phylum control and show that this comparison also produces an inverse hourglass. Finally, we show that the inverse hourglass is a statistical artifact of standardization, and can not be used to infer temporal conservational patterns.

Figure 4.3A shows the pairwise Pearson correlation coefficient of one-to-one orthologs between each developmental stage of *Drosophila melanogaster* and *Danio rerio*. With the same methodology as the original study, we get a dual-phase pattern where both the early and the late stages between the two species seem conserved, but with a period of low conservation in the middle. If we now apply the same methodology to the chordates *D. rerio* and *X. tropicalis* of the previous section, we get a similar biphasic pattern. Note that figure 4.3B is based on the same sequencing data as figure 4.2B. The main difference in processing between those is the inclusion of gene standardization (z-score).

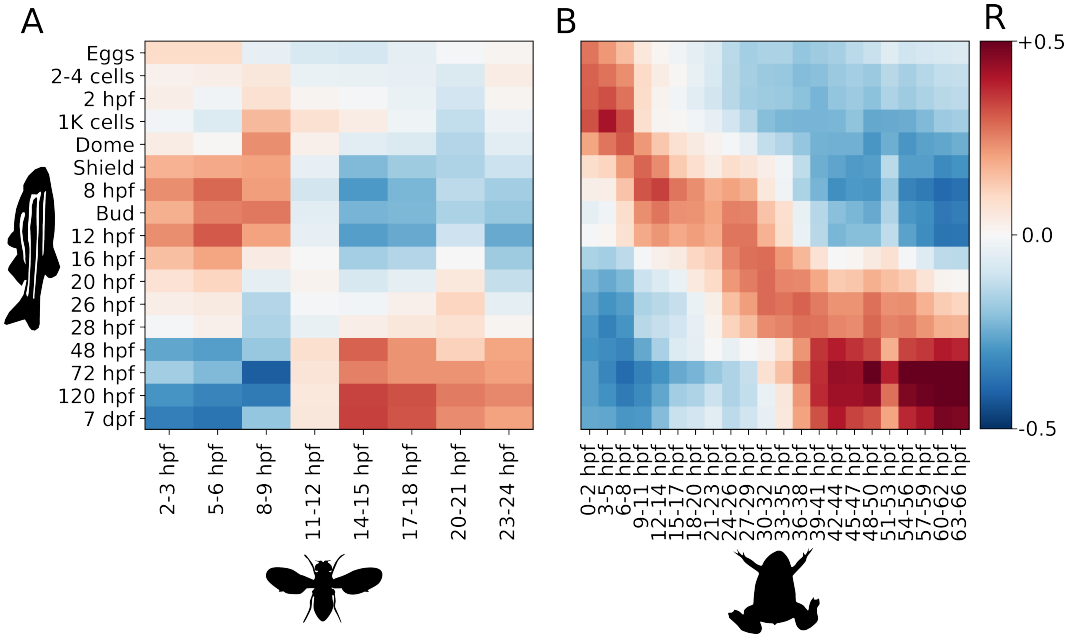


Figure 4.3: **The inverse hourglass is not exclusive to between-phyla comparisons**. Heatmaps of pairwise Pearson correlation coefficients **(A)** for a between phylum comparison of *Danio rerio* and *Drosophila melanogaster*, and **(B)** a within phylum comparison of *Danio rerio* and *Xenopus tropicalis*. A mid-developmental transition is visible for both comparisons.

Levin *et al.* apply gene standardization, which involves subtracting the mean and dividing by the standard deviation over time for each gene. Standardization is generally a good practice for parametric methods like the Pearson correlation coefficient. In the case of gene expression data, standardization effectively scales each gene to have equal weight in the correlation coefficient. The use of standardization here leads to an unexpected side effect. In the absence of standardization, the data exhibits an hourglass-like pattern (see Fig. 4.4A). However, after applying standardization, the

pattern transforms into an inverse-hourglass shape. Even more surprising, we can cut each time series into four equal parts, and after standardization, three out of four comparisons still display a mid-developmental transition (Fig. 4.4B).

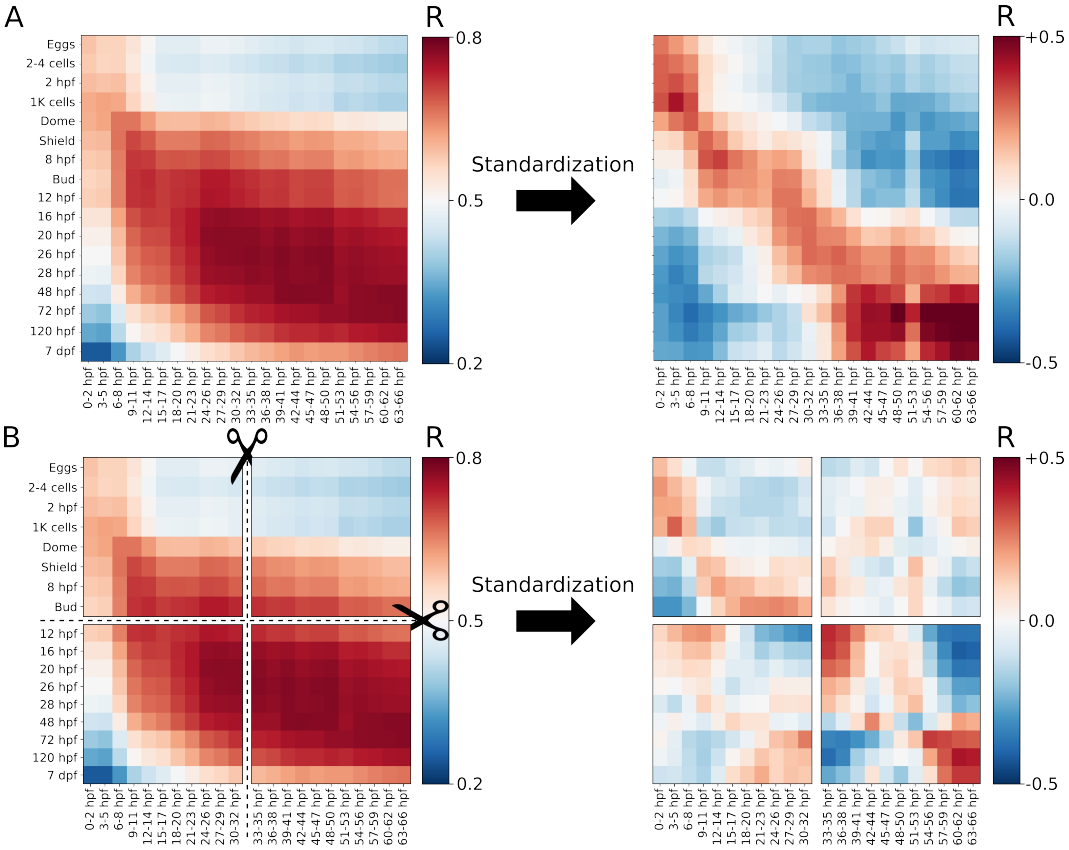


Figure 4.4: **The inverse hourglass is a statistical artefact of gene standardization**. **(A)** shows the effect of standardization. Before standardization, an hourglass-like pattern is visible, whereas after standardization an inverse hourglass is visible. **(B)** shows the effect of standardization after dividing the time series into four equal parts. After standardizing each part, three out of four subsets now display an inverse hourglass.

To get an understanding as to why this happens we need to study the patterns of gene expression during development. Levin *et al.* introduced the concept of a gene landscape, which offers a way to visualize gene expression patterns. The gene landscape is a histogram of the Pearson correlation coefficient for each gene with a linearly increasing line. A coefficient of 1 means that a gene is linearly going up over time, a coefficient of -1 means that a gene is linearly going down over time, and a coefficient of 0 means that a gene shows no linear temporal pattern. In figure 4.5 we show the gene landscape for *D. rerio*, *D. melanogaster*, and *X. tropicalis*. For each gene landscape, we observe a bimodal distribution, with an enrichment for genes that are either going up or down in expression over time and relatively few genes having no (linear) temporal pattern. The scatterplot that Levin *et al.* report (Extended Data Figure 5[56]), unfortunately, hides this pattern, and a 2D histogram would have been a better choice. As embryos grow, one would expect practically all genes to increase in expression over time. But as RNA sequencing (without spike-ins) is inherently

relative, genes are split into either one of two expression groups; a group where gene expression goes up or increases faster than the average gene (right side of the gene landscape) or the group where gene expression goes down or increases slower than the average gene (left side of the gene landscape). See figure S4.3 for the difference between per-embryo (spike-in) normalization and transcript per million (TPM) normalization for *X. tropicalis* embryos.

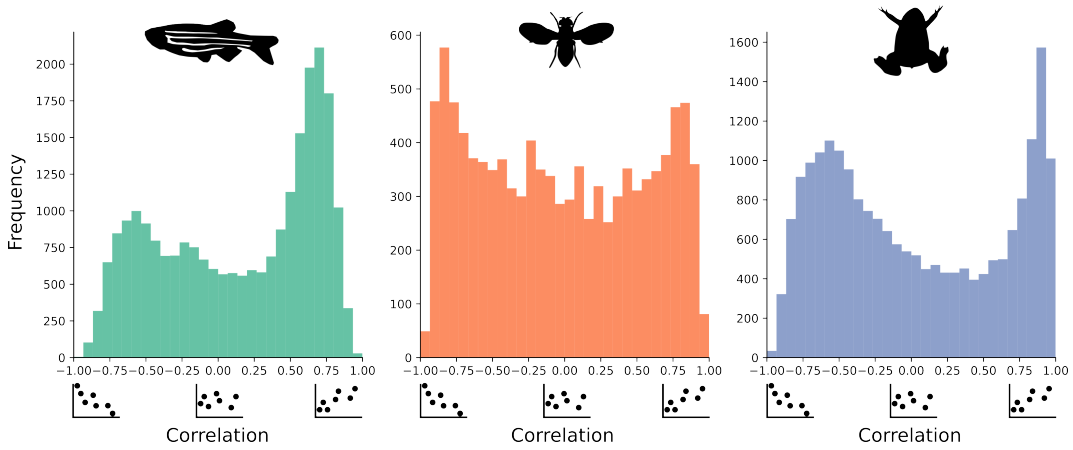


Figure 4.5: **The gene landscape of *D. rerio, D. melanogaster*, and *X. tropicalis* embryos**. The gene landscape is a histogram of the correlation coefficient of the expression of each gene with a linearly increasing line. Note how generally speaking there are two groups of genes; a group of genes that is upregulated, and a group that is downregulated during embryonic development.

Now that we understand that our genes can roughly be classified as either going up or going down over time, we can think about how this influences our analyses. To analyze the effects of the gene expression landscape on hourglass analyses, we simulated the time series of a single gene that has a binary expression profile, where at the start of our time series the gene is *off*, and at a random time point switches *on* and stays *on* until the end of the series. We now imagine the expression profile of this gene in a related species and assume again that it starts *off* and at a random time point switches *on*. We can express the probability that these two imaginary time series are equal (section 4.7.1), and if we visualize these probabilities it is clear these odds display a mid-developmental transition (Fig. S4.4). This theoretical derivation is however an oversimplification of what happens both biologically and methodologically and considers only a single gene. For this reason, we simulated two time series with continuous expression profiles. In these series half of the genes start in an *off* state where expression is zero, and half in an *on* state where expression is one. Similarly to the single-gene thought experiment, these genes, at a random time-point, gradually switch from *off* to *on*, or vice versa. We can now calculate the Pearson correlation coefficient between these simulated series, and we get a clear mid-developmental transition (Fig. 4.6). Similar to the biological data, if we cut the simulated data into halves, and apply standardization afterwards, we get a mid-developmental transition per subset of the data (Fig. S4.5).
By incorporating a within-phylum comparison it becomes clear that the mid-developmental transition is a statistical artifact of gene standardization. It can be considered a special case of Simpson's paradox, where by standardization gene expression gets put into two groups (high vs low expression). And even though there is no particular correlation within each group, there is a clear correlation when comparing the data set as a whole[323]. This pattern is not caused by lowly expressed genes, as we applied the same criteria as Levin *et al.* to only include dynamic genes (minimum expression of 10 TPM and at least a log2 fold change). We conclude that gene expression standard-

ization in combination with the landscape of gene expression dynamics observed, produces the appearance of an inverse hourglass irrespective of the stages selected for analysis.

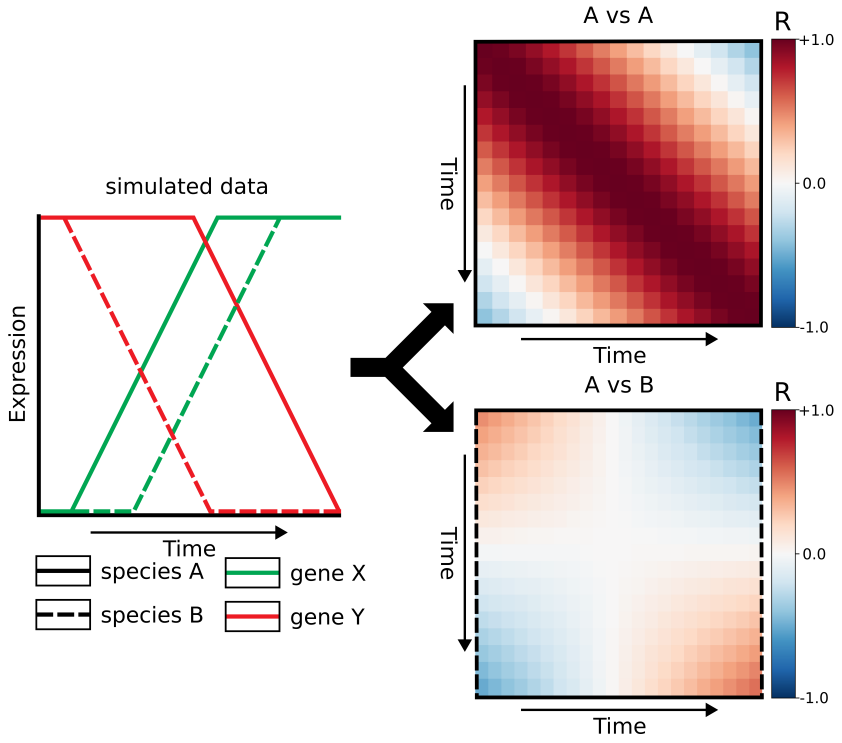


Figure 4.6: **The inverse hourglass is present in simulated data with no temporal conservation.** A schematic representation of the simulation of gene expression of two time series (A and B) on a continuous scale. Two groups of genes exist and are identical between the time series; upregulated (green) and downregulated (red) genes. The timing of up- and downregulation is completely random. When comparing such time series with themselves (A vs A) there is no temporal pattern (null model), however when comparing two such time series (A vs B) an inverse hourglass appears.

4.3.3 The mouse and rabbit cell type proportion bottleneck is a within-species effect

In the paper *Time-aligned hourglass gastrulation models in rabbit and mouse*[58] Mayshar *et al.* analyzed the similarity between developing rabbit and mouse embryos on a single-cell level. One of the comparisons they make is how the correlation of cell type proportions between rabbits and mice changes over time. They observe an "hourglass-like" bottleneck of cell type proportions pre-gastrulation. In this re-analysis, we reproduce the between-species cell type proportion bottleneck. Additionally, we compared the mouse and rabbit time series against themselves and discovered similar bottlenecks. This bottleneck is caused by the combination of a statistical artifact of new cell types appearing and an inappropriate temporal scale. It is unlikely that the bottleneck signifies an evolutionary-developmental effect between *M. musculus* and *O. cuniculus*.
Figure 4.7 shows the pairwise Pearson correlation coefficient between cell type proportions of developing *M. musculus* and *O. cuniculus*. Figure 4.7B is the between-species comparison and shows the same result as the original paper. Mayshar *et al.* describe this as a stereotypical pattern, where

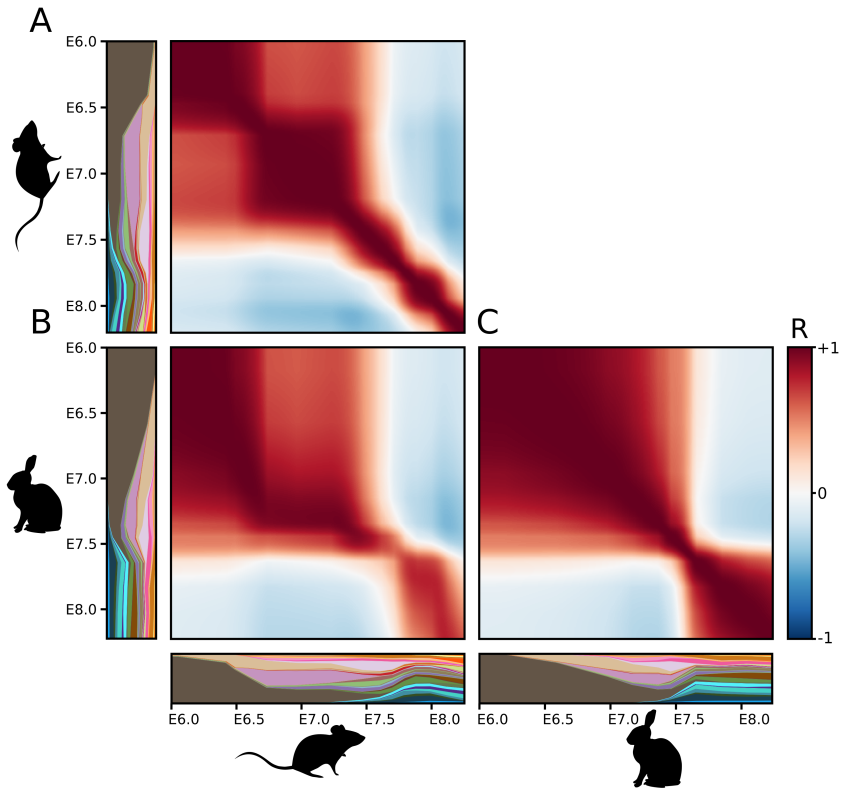


Figure 4.7: **Within-species cell type proportion similarities translate to between-species cell type proportion similarities**. Heatmap of pairwise Pearson correlation coefficients between **(A)** *M. musculus* with itself, **(B)** *M. musculus* and *O. cuniculus*, and **(C)** *O. cuniculus* with itself. The heatmaps are accompanied by cell-type proportion charts with similar coloring as the original paper. The between-species pattern seems to be a superimposition of the within-species effects.

the beginning of development is aligned but not synchronized. This then leads to a bottleneck at approximately E7.5 followed by a more synchronized gastrulation process marked by cellular diversification. They conclude that around E7.5-E7.7, the narrowest point of the bottleneck, mouse and rabbit gastrulation are aligned with maximum specificity. However, when we compare the time series of *O. cuniculus* against itself (a within-species comparison) we see a similar bottleneck (Fig. 4.7C). Before E7.5 the rabbit embryo consists of a small number of cell types, forcing the cell type distributions into one of two groups; a group of cell types that do not occur, and a group of cell types that do occur. From E7.7 on, practically all cell types are present. All comparisons within *O. cuniculus* before E7.5 give high correlation values because in this case, the Pearson correlation represents whether identical cell types are present. It says little about the correlation coefficient between cell types (known as Simpson's paradox[323]). The bottleneck at E7.5 is caused by the rapid appearance of new cell types. Similar problems exist for the *M. musculus* time series but are less pronounced. These within-species effects then get carried over to the comparison between *M. musculus* and *O. cuniculus*, leading to a deceptive between-species pattern.
What is surprising about this analysis is that the pattern that Mayshar *et al.* describe as stereotypical pattern and hourglass-like, is neither stereotypical nor fits the hourglass model. For species of the same phylum, no such sudden changes in mid-development are expected. Instead, the oppo-

site is predicted by the hourglass model. Moreover, the comparison in its current form displays a funnel pattern, with the highest similarity between *M. musculus* and *O. cuniculus* at the start of the time series which is gradually decreasing, although this is caused by global differences in cell proportions and does not represent a meaningful biological signal (Simpson's paradox). Furthermore, Mayshar *et al.* conclude that their analysis shows that one can use absolute (linear) time for pairwise comparisons between gastrulating embryos. But considering the within-species comparisons, we can clearly see that the temporal axis is not representative of change. There is a higher rate of change happening between E7.5-E7.7 (5 hours) than between E6.0-E7.0 (24 hours). Altogether, the hourglass-like shape of temporal similarity between *M. musculus* and *O. cuniculus* seems to be caused by within-species dynamics of cell type composition, which in turn is partially caused by unrepresentative temporal sampling and the fact that subgroups exist (Simpson's paradox).

4.3.4 Temporal enhancer conservation between *Drosophila* is confounded by the number of enhancers at each respective stage

In the paper *The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection*[59] Liu *et al.* compare the similarity of accessible regions over five matched embryonic developmental time points between two *Drosophila* species (*D. melanogaster* and *D. virilis*). Liu *et al.* find that the middle time point (TP3, 8-10 hours after egg laying) has the highest number of enhancers accessible, and that at TP3 *D. virilis* and *D. melanogaster* have the highest fraction of shared enhancers. TP3 coincides with the *Drosophila* phylotypic stage, and thus this result is seen as supportive for the hourglass model. In this reanalysis, we show that this high conservation of enhancers at the phylotypic stage is explained by a different number of enhancers found per time point, and is not an evolutionary-developmental pattern.

In this study, conservation is estimated by the similarity between time point-specific enhancers between *D. melanogaster* and *D. virilis*, where time point-specific enhancers are defined as enhancers that are accessible in only one time point (TP). Enhancers are defined as accessible regions farther than 500 bp from a transcription start site. Finally, the similarity between the two species is calculated by dividing the number of TP-specific enhancers overlapping between both species by the total number of TP-specific enhancers for both species (Jaccard index). For all *D. virilis* TP-specific enhancers, we inferred their corresponding orthologous regions in the *D. melanogaster* with pslMap, restricting to one-to-one orthologs. Figure 4.8A shows the conservation over time between *D. melanogaster* and *D. virilis*, with the highest conservation at TP3, closely matching the results of Liu *et al.* Figure 4.8B shows the number of TP-specific enhancers per time point. The high amount of TP-specific enhancers at TP1 and TP5 can be explained by the fact that they are respectively the start and end of the time series. The high number of TP-specific enhancers at TP3, however, is not easily explained and in turn, raises the question of whether this can explain the high similarity between *D. melanogaster* and *D. virilis*.

To test whether a relationship between the number of TP-specific enhancers and the Jaccard index exists, we made a consensus set of enhancers for *D. melanogaster* and *D. virilis* over all time points. Each time point gets assigned a set of randomly picked enhancers, whilst keeping the original number of enhancers per time point. This removes all biological meaning from the data, thus if the analysis is unbiased it should generate equal Jaccard indexes for all time points. Yet TP3 clearly shows an enriched Jaccard index (Fig. 4.9A). This indicates that the number of enhancers per time point influences the analysis. The obvious way to control for this would be to subsample all time points to the same number of enhancers (Fig. 4.9B). After subsampling, we find that TP2 (6-8 hours after egg laying), a period that precedes the *Drosophila* phylotypic stage[62,312], shows the highest conservation between enhancers between *D. melanogaster* and *D. virilis*. The dependence of the

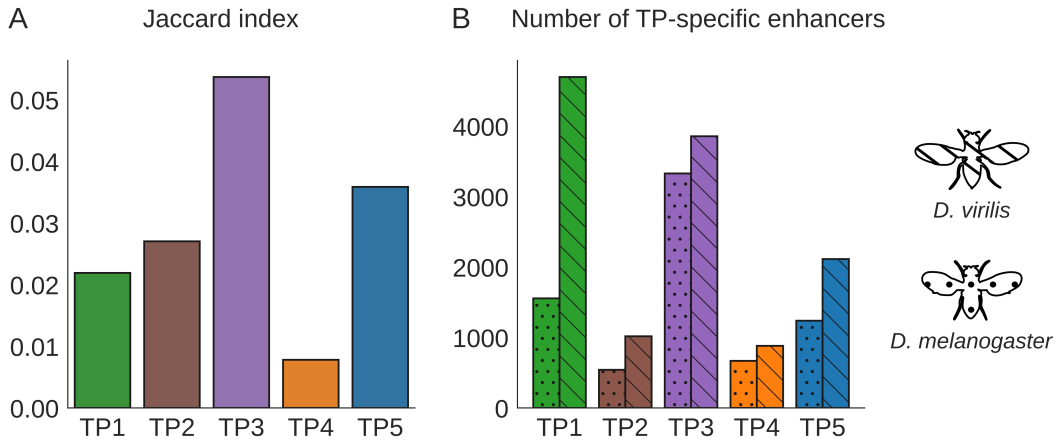


Figure 4.8: **Temporal similarity between *D. virilis* and *D. melanogaster*. (A)** The proportion of conserved stage-specific enhancers at each development stage between *D. melanogaster* and *D. virilis*. TP1 corresponds to 2-3 HAEL (hours after egg laying), TP2: 6-8 HAEL, TP3: 10-12 HAEL, TP4: 14-16 HAEL, TP5: 18-20 HAEL. **(B)** The number of time point specific enhacers for *D. melanogaster* and *D. virilis* over time.

Jaccard index is also present in a within-species comparison between *D. virilis* replicates. The dependence of the Jaccard index on the number of enhancers can also be formally established, see section 4.7.2.

Apart from a biological interpretation for the different number of TP-specific enhancers, there are also two methodological interpretations. First, the focus on TP-specific enhancers creates artificially separated sets of enhancers. TP-specific enhancers, as defined by Liu *et al.*, are enhancers occurring at one time point only. This means that if two time points are sampled more closely in (developmental) time, these two time points would share most of their enhancers, resulting in a low number of TP-specific enhancers. This is something the authors themselves already note. Similarly, enhancers at the beginning and the end of the time series have a higher chance of being TP-specific purely because there is only one adjacent time point, whilst the rest of the time series are compared against two. A similar approach that doesn't suffer from the artificial separation, would be to visualize the percentage of reads in the consensus peak set in enhancers vs promoters per time point. Applied to this data we find no clear enrichment for enhancers for a specific time-point (Fig. S4.8). Secondly, arbitrary thresholds are used during peak calling to decide whether a region is *enriched*. This threshold depends, for instance, on the signal-to-noise ratio, which is expected to change for developing embryos, and the sequencing depth[324]. Neither of these confounders has been corrected for in the original analysis.

Moreover, Liu *et al.* proceed to train a computational model on the sequences of TP-specific enhancers of *Drosophila*. This model is then used to determine per enhancer whether it has been subjective to positive or non-positive selection. They then find that the ratio of enhancers subjected to positive selection vs enhancers subjected to non-positive selection is high for TP1, TP3, and TP5, and thus conclude that the molecular basis for the phylotypic stage can partially be explained by positive selection of gene regulation conservation. Alternatively, there are two other likely (methodological) explanations for this pattern that have not been tested. The first explanation is, again, that as there has been no correction for signal-to-noise ratio and sequencing depth, there is a difference in the type(s) of peaks the pre-processing picked up, which in turn translates to different

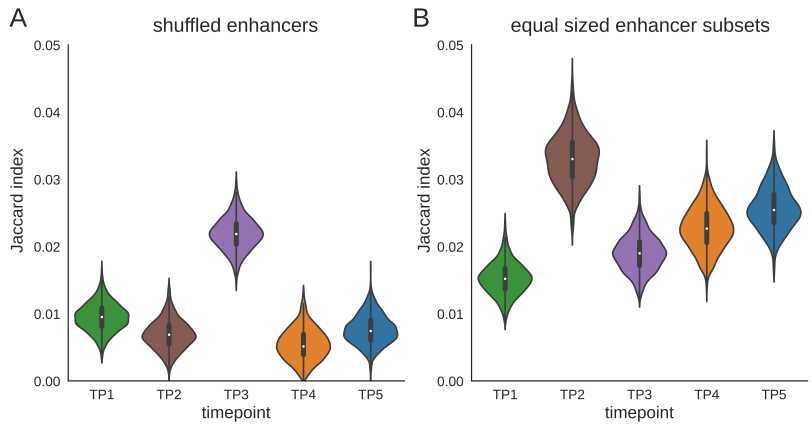


Figure 4.9: **The Jaccard index depends on the number of enhancers. (A)** Distribution of Jaccard scores after randomly distributing enhancers to time points, but with an identical number of enhancers per time point as originally found (1,000 permutations). If the number of enhancers per time point does not influence the result an equal distribution is expected. The enrichment of TP3, however, indicates that this methodology is sensitive to the number of enhancers per time point ((*dmel*, *dvir*) TP1: (6,768, 12,150) TP2: (6,093, 8,334) TP3: (12,052, 13,404) TP4: (5,597, 5,941) TP5: (6,352, 8,390)). **(B)** Subsampling to 2,000 enhancers per time point (1,000 permutations). This removes the dependency of the Jaccard index on the number of enhancers. TP3 is not enriched after subsampling.

predicted positive selection ratios[324,325]. The second explanation is that, as is true for practically all machine learning, the model generalizes and performs better with more training data. Looking at their reported model performance (Fig. 3[59]) we can see that the model's performance for TP1, TP3, and TP5 is notably higher than for TP2 and TP4. In figure S4.6B we see that the high model performance closely follows the amount of training data (number of TP-specific peaks). Simply speaking, a model that has been trained with more data generalizes better and appears to predict a higher amount of positive selection. Without a proper test for these potential confounders, it is hard to reach a definitive conclusion about positive selection concerning the hourglass model and the phylotypic stage.
In summary, in our re-analysis we reproduce the result that TP3 has the highest overlap between enhancers between *D. melanogaster* and *D. virilis*. However, the Jaccard index is dependent on the combination of biological similarity and the number of enhancers per time point. This dependence is visible when comparing *D. virilis* replicates among themselves and when shuffling the data. Moreover, it is likely that the number of TP-specific enhancers also influences the downstream training of the gkm-SVM and in turn the pattern of enhancer positive selection.

4.4 Discussion

Our study highlights the importance of including within-species, within-phylum, and between-phylum comparisons in the quantitative analysis of the phylotypic stage. We demonstrate that the high transcriptome similarity between *D. rerio* and *X. tropicalis* at the phylotypic stage[57] can be explained by within-species effects alone. Similarly, we show that the cell type proportion "hourglass-like" bottleneck between *M. musculus* and *O. cuniculus*[58] is also caused by within-species effects.

Furthermore, we identify the mid-developmental transition[56] and *Drosophila* enhancer conservation at the phylotypic stage[59] as methodological artifacts.

When comparing time series directly, it is important to control for statistical biases in the data, such as the correspondence between replicates, but also the temporal sampling strategy, data distribution, and gene sets used. A key bias that is often overlooked is whether the sampling schema follows developmental time[326]. It is not uncommon for studies to report a discontinuous blocked pattern of within-species correlation, where the difference between stages is not equal. This difference usually gets assigned a biological explanation; for instance as embryonic genome activation[327] or developmental milestones[328]. However, this blocked within-species pattern proceeds to affect any downstream between-species comparison, where in turn the same blocks are visible. The only way to make a fair comparison between-species, however, is when the within-species similarity between all adjacent time points is equal (which corresponds to the null model within a time series of Fig. 4.1), or if one explicitly corrects this bias statistically. When sampling stages closely in time, one increases the probability of stages matching closely in a comparison with another time series. This in turn increases the expected similarity. The reverse is also true, where stages sampled in large intervals reduce the probability of matching stages between time series, which then reduces the expected similarity. Under the assumption that embryonic development has no point of higher selective pressure, a comparison between two time series will have the highest expected similarity at the point where sampling has happened at the shortest interval. For this reason, we argue for a re-definition of molecular embryonic time, where not morphological features or absolute time post-fertilization is used, but where the (molecular) difference between adjacent stages is equal.

A different potential bias is the changing gene expression distributions, which are a consequence of developmental maturity. It has been demonstrated that transcriptional entropy, a measure of disorder in a system, is decreasing over development[329]. Pluripotent tissues have relatively many genes activated, whilst more differentiated cells in mature tissues make use of a smaller, more specialized set of genes. These changing gene expression distributions during development can introduce unexpected biases and artifacts when calculating similarity. For instance, the JSD and Pearson correlation coefficient between two count tables from different distributions can never be maximal (respectively zero and one). As an effect, directly comparing similarity scores between time points can be misleading as the range of possible similarity scores is different per comparison. To circumvent this problem one can resort to non-parametric distance metrics, such as the rank-based Spearman correlation coefficient[61], or quantile normalizing the data before calculating the distance metric[57]. Then, one can compare the similarity scores directly, but these approaches ignore the biologically relevant gene expression distribution changes. As long as there is no consensus on what type of molecular similarity is expected at the phylotypic stage, it can both be correct as well as incorrect to ignore the distributional changes.

Another point to consider is the changing proportions of cell types. An embryo is formed by all its cells together, and their combined signal is measured in the study of the molecular phylotypic stage. But not every cell type expresses the same number of transcripts[330,331], which consequently gives cells with a high number of transcripts a higher importance in the (dis)similarity calculation. Moreover, molecular changes between embryos could be caused by changes in gene expression, but also by changes in cell type proportions, and most likely a combination of them. For example, during the vertebrate phylotypic stage the neural tissue is already highly developed and neural cells are relatively abundant. Could the high similarity at the vertebrate phylotypic stage be an effect of the over-representation of these neural cell types? Or, as another example, a practically universal feature of embryos is that they grow during development. But because of this growth, the surface (skin) to volume (organs) ratio changes, which in turn could explain the molecular phylotypic stage. We currently have no understanding of whether the molecular similarity at the phylotypic stage is based on cell type proportion similarity or whether it is based on gene expres-

sion similarity (or the combination of both). Only a single-cell approach can distinguish the effects of cell type proportion and gene expression, which is essential for better understanding the basis of the molecular phylotypic stage.
Finally, calculating a single similarity score between species seems like a gross oversimplification of the complexity of evolutionary development. When focusing on one-to-one orthologues features, all newly derived and lost features are ignored. For instance, we started our analysis with 21.154 genes for *X. tropicalis* and 24,417 genes for *D. rerio* respectively. However, if we focus only on one-to-one orthologs this gets reduced to 5,444 genes. This means that for this comparison specifically, we discard approximately three-quarters of our biological signal. To our knowledge, the only method that considers these relationships is the transcriptomic derivedness index[315]. The transcriptome derivedness makes use of the average expression of orthogroups, where empty groups get assigned a gene expression of zero. But in turn this makes this approach vulnerable to Simpson's paradox, where the correspondence between lost and common genes is dominating the similarity instead of the similarity between common genes[323]. Moreover, it is still vulnerable to changing levels of similarity between replicates over development, and requires a within-species control. Nevertheless, the transcriptomic derivedness index is currently the only approach to model all evolutionary relationships. Besides ortholog pairings, gene expression is regulated by the complex interplay of multiple gene regulatory mechanisms. For instance, a single differentially expressed transcription factor can affect thousands of downstream genes. Biological data is notoriously not independently and identically distributed. Whole-transcriptome comparisons are thus biased towards the largest groups of co-regulated genes. In addition, the fact that comparisons subsetted on different GO terms result in different conservational patterns is a clear indication that a single metric for whole-transcriptome similarity conceals the different layers of conservation at the phylotypic stage[332–334].
In the past decades, there has been extensive research into the (molecular) basis of the phylotypic stage. Yet, as a field, we have neglected to explicitly define the various evolutionary-developmental models, their predictions, and their limitations. Currently, similarity is often defined on an arbitrary basis depending on the available data. However, due to differences in methodology, vastly different results are obtained[63,319,320]. Furthermore, this unstructured approach leaves us unaware of nonconforming results as science tends to predominantly report positive findings, leaving us unaware of negative or inconclusive results. Regardless of the similarity metric used, does the hourglass model predict that maximal developmental similarity is a between-species effect? Or does it already exist when comparing replicates of the same species? There is an implicit expectation that a point of maximum similarity exists between species of the same phylum, and not between species of different phyla. Yet phyla are an artificial framework, and to date, no evolutionary process has been described that is exclusive along the 35 phyla stems[322]. In summary, the phylotypic stage lacks an explicit definition and there is conflicting evidence supporting its existence. The field must shift focus from generating more data and comparative analyses, and instead reflect on past observations while establishing a clear and falsifiable definition for the phylotypic stage. Without such a definition, our efforts remain unproductive.

4.5 Material and Methods

4.5.1 Overview of public data

Full sample tables can be obtained from https://zenodo.org/doi/10.5281/zenodo.10457767.

Comparative analyses

Xenopus tropicalis transcriptomic time series data was obtained from DDBJ:PRJDB3785[316], and NCBI:PRJNA275011[335], and mapped against assembly UCB_Xtro_10.0. *Danio rerio* transcriptomic time series data was obtained from NCBI:PRJNA416866[57] and EBI-ENA:PRJEB7244[321] and mapped against assembly GRCz11. *Drosophila melanogaster* transcriptomic time series data was obtained from NCBI:PRJNA527284[59] and mapped against assembly BDGP6.32. The *Drosophila melanogaster* DNAse I data was obtained from EBI-ENA:PRJEB10089[312] and mapped against the dm6 assembly. The *Drosophila virilis* DNAse I data was obtained from EBI-ENA:PRJEB10089[312] and mapped against the droVir3 assembly.

Within-species temporal variance

The within-species transcriptomic temporal variance is based on the data of EBI-ENA:PRJEB7244[321] (GRCz11), NCBI:PRJNA527284[59] (BDGP6.32), and NCBI:PRJNA345017[336] (ce11).

Per-embryo vs TPM normalisation

Per-embryo vs TPM normalisation comparisons are based on the count tables directly provided by NCBI:PRJNA275011[335].

4.5.2 Transcriptome analyses

Preprocessing of RNA-seq was done automatically by seq2science v0.9.8[337] using the rna-seq workflow. Public samples were downloaded from the Sequence Read Archive with the help of the NCBI e-utilities and pysradb[266]. Genome assemblies UCB_Xtro_10.0 (*X. tropicalis*), GRCz11 (*D. rerio*), BDGP6.32 (*D. melanogaster*), ce11 (*C. elegans*) and GRCm38.p6 (*M. musculus*) were downloaded with genomepy 0.13.0[267]. Reads were trimmed with fastp v0.20.1[269] with default options. Reads were aligned with STAR v2.7.6a[338] with default options. Subsequently, duplicate reads were marked with Picard MarkDuplicates v2.23.8[280]. General alignment statistics were collected by samtools stats v1.14[339]. Read counting and summarizing to gene level was performed on filtered bam using HTSeq-count v0.12.4[288]. TPM normalized gene counts were generated using genomepy based on longest transcript lengths. Quality control metrics were aggregated by MultiQC v1.14[282].
The within-species permutation test (Fig. S4.1) was performed by randomly choosing two replicates of the same time point and calculating their Spearman correlation coefficient, with 250 random pairs per time point.
Orthologs between species were derived by the gimmemotifs motif2factors command (v0.18.0). This command downloaded the genome assemblies of cattle (*ARS-UCD1.2*), fruit fly (*BDGP6.32*), lancetfish (*BraLan2*), human (*GRCh38.p13*), mouse (*GRCm38.p6*), zebrafish (*GRCz11*), frog (*UCB_Xtro_10.0*), maize (*Zm-B73-REFERENCE-NAM-5.0*), nematode (*ce11*), chicken (*galGal6*), tardigrade (*nHd_3.1*), koala (*phaCin_unsw_v4.1*), and turtle (*rCheMyd1*) through genomepy[267], and converted all transcripts to peptides with gffread 0.12.7[340]. Only the longest peptide per gene was kept, which was then provided to orthofinder 2.5.4[341].
For the re-analyses we then took the average TPM per time point, kept only the one-to-one orthologs per species-species comparison, and did similar processing as the original studies. Specifically for the re-analysis of Marlétaz *et al.* we quantile normalized[342] the TPMs and calculated the Jensen-Shannon distance on a log2 scale with TPMs divided by a million. The *absolute* JSD values are vastly different between our and the original analysis. This is because we opted to represent TPMs as probabilities (TPM divided by a million) and calculate JSD using a log base of 2. This causes the JSD to be bound between 0 and 1, which makes comparisons between different data sets easier as they are on the same scale. For the re-analysis of Levin *et al.* we kept all genes with a

minimum TPM of 10 or higher, and with at least a 2-fold change. Then we log10 transformed the remaining data and calculated the Pearson correlation coefficient.

4.5.3 Enhancer conservation

We had trouble closely reproducing the original results of Liu *et al.*, so we have opted to use the original processed data for the between-species comparison, but our own processed data for the within-species comparisons as this data is missing. By closely reproducing the original results with their data it shows that our ortholog inference works similarly to theirs.
Preprocessing of the samples for the within-species comparisons was done automatically by seq2science v0.9.8[337] using the atac-seq workflow. Public samples were downloaded from the Sequence Read Archive with the help of the NCBI e-utilities and pysradb[266]. Genome assemblies *dm6* and *droVir3* were downloaded with genomepy 0.13.0[267]. Paired-end reads were trimmed with fastp v0.20.1[269] with default options. Reads were aligned with bwa-mem2 v2.2.1[272] with options '-M'. Afterwards, duplicate reads were marked with Picard MarkDuplicates v2.23.8[280]. Before peak calling, paired-end info from reads was removed with seq2science so that both mates in a pair get used. The peak-calling effective genome size was estimated by khmer v2.0[343] by calculating the number of unique k-mers with k being the average read length per sample. Peaks were called with macs2 v2.2.7[37] with options '–shift -100 –extsize 200 –nomodel –buffer-size 10000' in BAM mode. A consensus set of all summits was made with gimmemotifs 0.17.2[139]. We then removed all summits that fall within a TSS with pyranges v0.0.120[38] to only keep enhancers.
For the between-species comparison, we used pslmap to map all *droVir3* enhancers to the *dm6* assembly and only kept one-to-one orthologous regions. This is not necessary for the within-species comparisons. Time-point-specific enhancers are defined as enhancers of which the summits are more than 200 base pairs removed from the enhancers of other time points. Overlap over time is calculated as the Jaccard index.
Permutation tests were performed on a table where the rows are the enhancers and the columns time points, and where each value is whether or not for that enhancer-time point combination the enhancer is called as a peak. We then randomly shuffled each column separately, so that the total number of enhancers stays the same, but the biological structure is removed from the data. On these tables, we then calculated the Jaccard index, and we repeated the process 1,000 times to get a distribution. Similarly, for the subsetted data, we took a random subset of 2,000 rows (enhancers) and calculated the Jaccard index on this. We similarly repeated this 1,000 times to get a distribution.

4.5.4 Cell proportion re-analysis

The table containing the inferred cell type proportions, including the color per cell type, were shared with us directly over email. Pairwise Pearson correlation coefficients were calculated and the corresponding area graphs were added to each time series.

4.5.5 Code availability

The final analysis can be found at https://github.com/vanheeringen-lab/phylotypic_hourglass.

4.6 Acknowledgments

We would like to thank Jialin Liu for his quick response to queries about the *Drosophila* transcriptome dataset, Eileen Furlong for her help with processing the DNAse *Drosophila* samples, Michal Levin and Itai Yanai for sharing their original orthology dataset, Ofir Raz for sharing the mouse and

rabbit data, David Emms for help with questions about orthofinder, Mike Keesey for developing Phylopic and finally Eivind Fonn for his help with formalizing the mathematical derivations.

4.7 Supplementals

4.7.1 Mid-developmental transition derivation

A basic proof that the mid-developmental transition is a methodological artifact follows relatively easily from the thought experiment where a gene starts the time series in an *off* state, and at random switches to an *on* state somewhere along this time series:

$$G(t) = \begin{cases} \text{off}, & t < t_{\text{activate}} \\ \text{on}, & t \geq t_{\text{activate}} \end{cases} \quad \text{and} \quad t_{\text{activate}} = \text{Uniform}(0, 1) \quad \text{and} \quad 0 \leq t \leq 1$$

then:

$$\begin{aligned} P(G(t) = \text{on}) &= t \\ P(G(t) = \text{off}) &= 1 - t \end{aligned}$$

If we now assume we have two of these time series (x and y), we can define the probabilities that both genes are in the same state for t_x and t_y:

$$\begin{aligned} P(G(t_x) = G(t_y)) &= P(G(t_x) = \text{on}) \cdot P(G(t_y) = \text{on}) + P(G(t_x) = \text{off}) \cdot P(G(t_y) = \text{off}) \\ &= t_x \cdot t_y + (1 - t_x) \cdot (1 - t_y) \end{aligned}$$

When visualizing the probability for equality over all x and y a mid-developmental transition becomes clear (fig. S4.4).

4.7.2 Jaccard index

We assume that all enhancers are conserved between species x and y. Moreover, we assume that our pre-processing randomly finds enhancers for each time-point with chance α_{ts}, where t represents the time-point and s represents the species. We can calculate the number of time-point specific peaks for U_{ts} in this case as:

$$\begin{gathered} E \text{ is the collection of all conserved enhancers between species } x \text{ and } y \\ U_{ts} \text{ represents the time-point specific enhancers for species } s \text{ at time point } t \\ |U_{ts}| = \alpha_{ts}\beta_{ts}|E|, \text{ where } \beta_{ts} = \prod_{i \in \{1,2,3,4,5\} \setminus \{t\}} (1 - \alpha_{is}) \end{gathered}$$

Where β_{ts} represents the fraction of enhancers that are still eligible to be time-point specific. We can now calculate the expected overlap (union) between two time points of species x and y:

$$\begin{aligned}\mathbb{E}(|U_{tx} \cap U_{ty}|) &= |U_{tx}| * \frac{|U_{ty}|}{|E|} \\ &= \alpha_{tx}\beta_{tx}|E| * \frac{\alpha_{ty}\beta_{ty}|E|}{|E|} \\ &= \alpha_{tx}\beta_{tx}\alpha_{ty}\beta_{ty}|E|\end{aligned}$$

From this we can calculate the expected Jaccard index:

$$\begin{aligned}Jaccard(U_{tx}, U_{ty}) &= \frac{|U_{tx} \cap U_{ty}|}{|U_{tx} \cup U_{ty}|} \\ &= \frac{|U_{tx} \cap U_{ty}|}{|U_{tx}| + |U_{ty}| - |U_{tx} \cap U_{ty}|} \\ &= \frac{\alpha_{tx}\beta_{tx}\alpha_{ty}\beta_{ty}|E|}{(\alpha_{tx}\beta_{tx} + \alpha_{ty}\beta_{ty} - \alpha_{tx}\beta_{tx}\alpha_{ty}\beta_{ty})|E|} \\ &= \frac{\alpha_{tx}\beta_{tx}\alpha_{ty}\beta_{ty}}{\alpha_{tx}\beta_{tx} + \alpha_{ty}\beta_{ty} - \alpha_{tx}\beta_{tx}\alpha_{ty}\beta_{ty}}\end{aligned}$$

For simplicity we assume that all found enhancers are time-point specific ($\beta_{ts} = 1$), simplifying the formula to:

$$Jaccard(U_{tx}, U_{ty}) = \frac{\alpha_{tx}\alpha_{ty}}{\alpha_{tx} + \alpha_{ty} - \alpha_{tx}\alpha_{ty}}$$

We can now visualize the Jaccard index for different α_{ts} values and can see a clear dependence on the fraction of enhancers found and the Jaccard index (Fig. S4.7). From this, it is clear that we need to correct for the number of enhancers. β_{ts} only influences the height of the Jaccard index, but not the pattern.

4.7.3 Supplemental figures

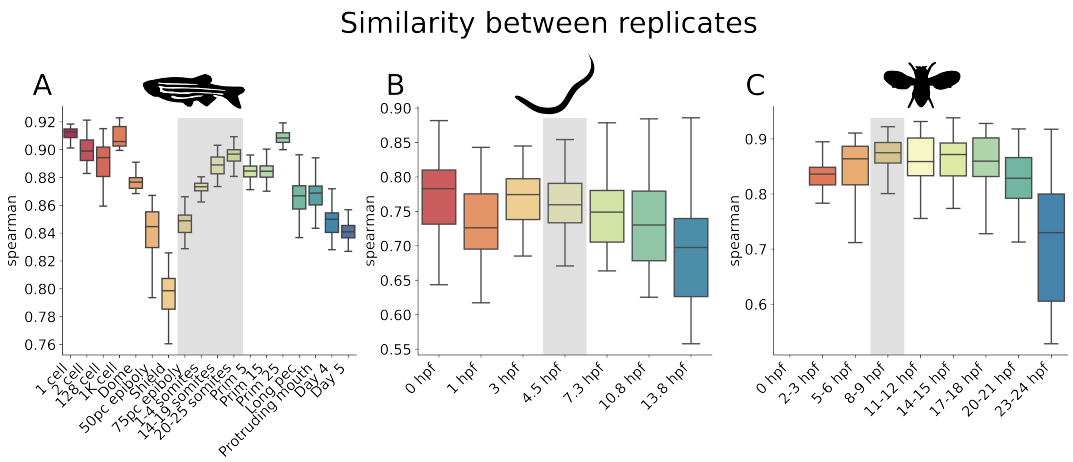


Figure S4.1: **Changing levels of similarity between replicates is a potential cause for bias.** Boxplots of gene expression Spearman correlation coefficients between replicates belonging to the same stage for (A) single embryo *D. rerio* samples, (B) pools of 10 *C. elegans* embryos, and (C) single-embryo *D. melanogaster* samples. The data is based on 250 randomly sampled pairs of replicates. The shaded area indicates the phylotypic stage. Note that all three species display a seemingly similar pattern, with high similarity between replicates at the start and mid-development. And a drop in early and late development. There are many potential causes for the changing levels of similarity over time; more biological variation between replicates at certain stages, a stage spanning too much time, or inversely a rapid development during certain stages, lab protocols that are optimized for certain stages and not others, etc.

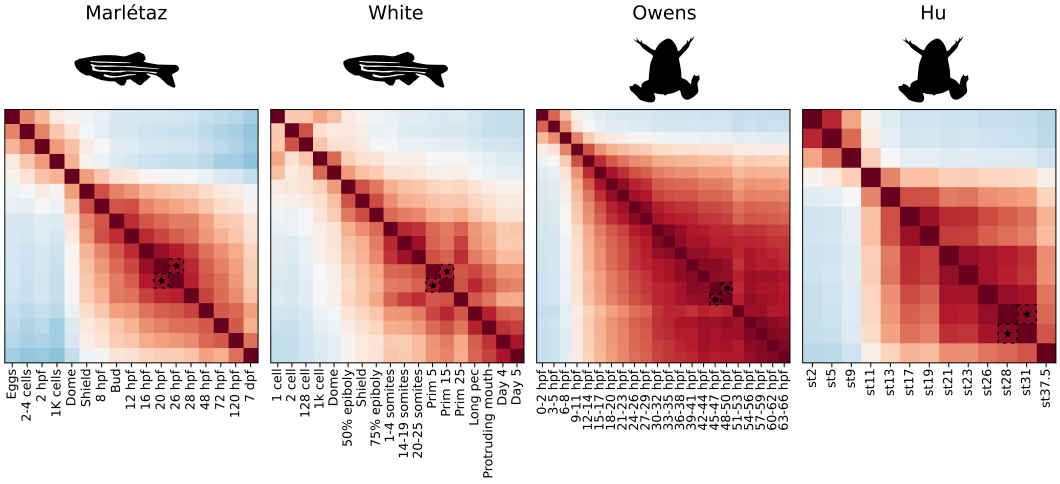


Figure S4.2: **High self-correlations at the phylotypic stage.** Heatmaps of pairwise Jensen Shannon Distances. Note how some series already display an hourglass-like pattern. Also note that the highest similarity between *X. tropicalis* and *D. rerio* in figure 4.2 corresponds to the highest self-similarity for *D. rerio* (Marlétaz).

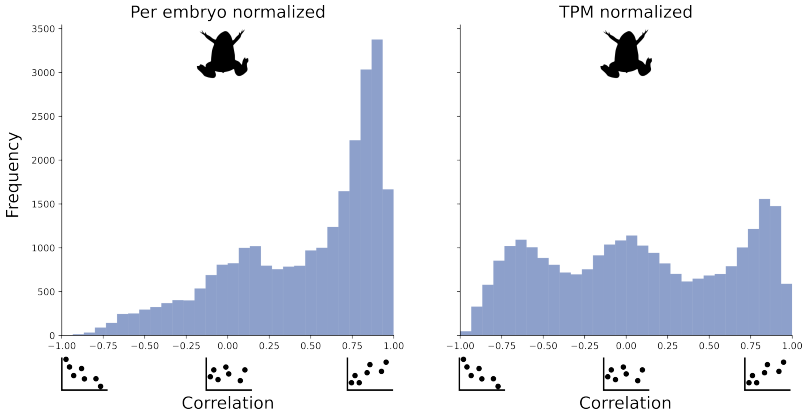


Figure S4.3: **The global per-embryo and TPM normalized gene expression patterns are vastly different for developing *X. tropicalis* embryos.** Histograms of gene landscapes between per-embryo normalization and TPM normalization of *X. tropicalis*. The left panel shows the gene landscape of per-embryo normalized gene expression. Because the *X. tropicalis* embryo grows in size practically all genes are upregulated on a per-embryo level. In the right panel, the gene landscape of TPM normalized gene expression is visualized. Here we see that there are roughly three groups of gene expression, a down-regulated, a constant, and an up-regulated group.

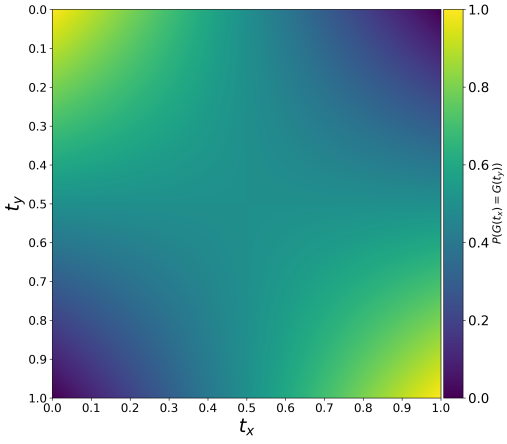


Figure S4.4: **The probability of two genes being equal in a simple system displays an inverse hourglass.** This assumes a simple system of two genes (x and y) that both start in the same *off* state and at a random time point switch to an *on* state.

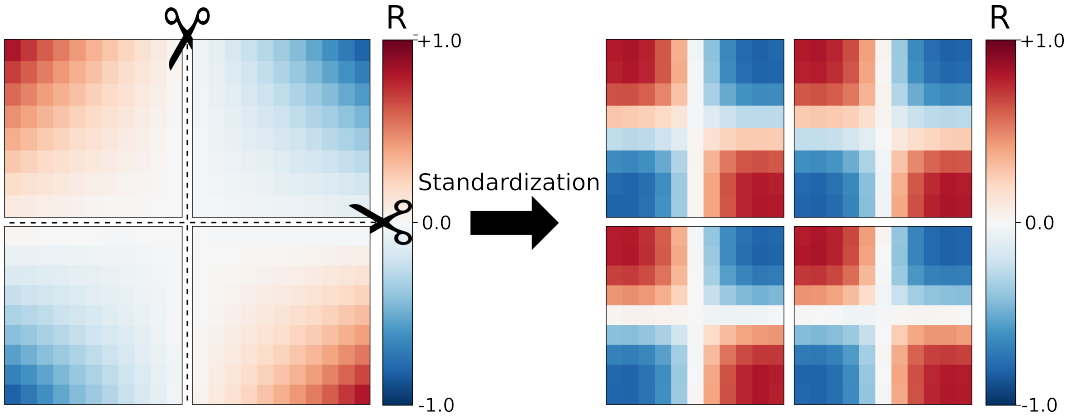


Figure S4.5: **Simulated data after being divided into four equal parts still shows an inverse hourglass after standardization.** This is a clear indication that the simulated suffers from the same statistical artifact as the real biological data. This means that two data sets that have no temporal relation can still show a temporal pattern.

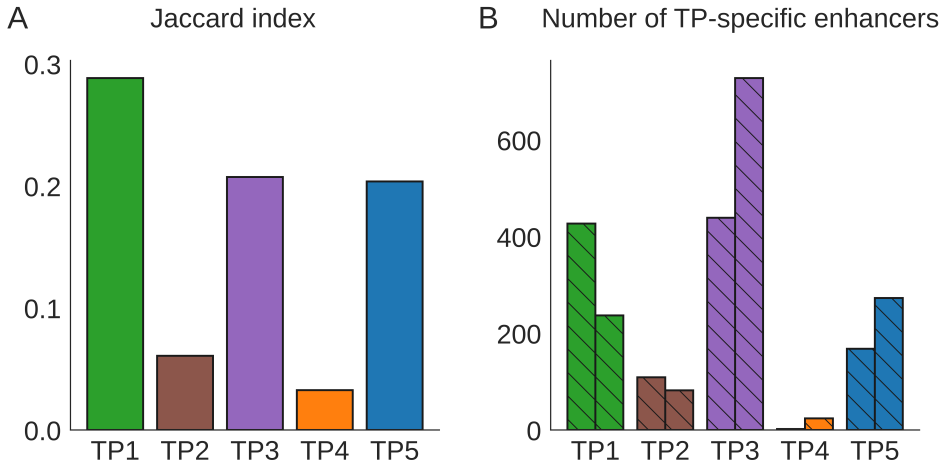


Figure S4.6: **Within species dependence of the Jaccard index on the number of enhancers**. **(A)** The proportion of conserved stage-specific enhancers at each development stage *D. virilis* replicates. **(B)** The number of time-point specific enhancers for *D. virilis* replicates over time.

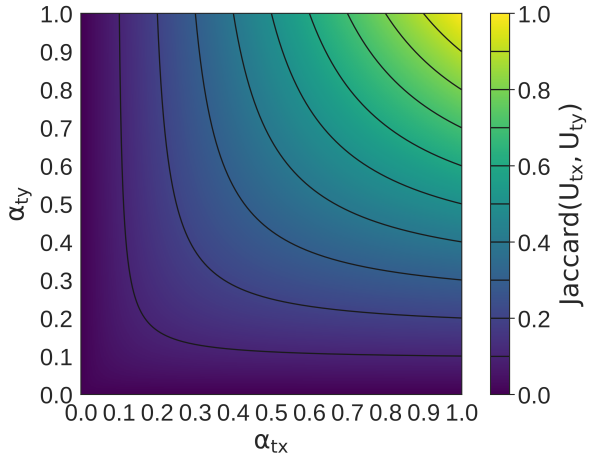


Figure S4.7: **The Jaccard index landscape for a different ratio of time-point specific enhancers.** This assumes that all enhancers are shared between two species, and α_{tx} is the ratio of total enhancers found at timepoint t for species x. The Jaccard index depends on the number of enhancers found and is especially sensitive to the lowest number of enhancers found in the two time series. This effect is present in the real data of figure 4.8 and figure S4.6.

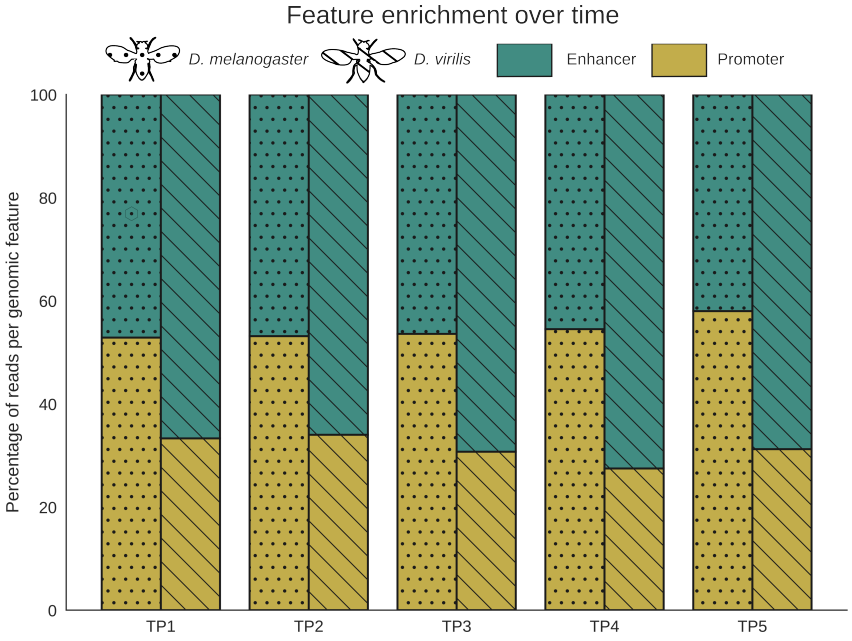


Figure S4.8: **There are no major sudden changes in the accessibility of enhancers and promoters during *Drosophila* embryonic development.** Bar plots of the percentage of reads in the consensus peak set belonging to either enhancers or promoters. Only looking at the number of enhancers per time-point can give a false indication of chance, as the total amount of peaks called can change as well as the signal-to-noise ratio.

Chapter 5

Edge Effects in the Temporal Analysis of Morphological Characteristics

Maarten van der Sande, Simon J. van Heeringen

In 2003 Olaf R. P. Bininda-Emonds, Jonathan E. Jeffery, and and Michael K. Richardson wrote a highly critical research paper about the hourglass model and the phylotypic stage, with the title "*Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development*"[318]. Their article starts with a critique of the lack of a definition of the hourglass model and its related phylotypic stage. They then compare the timings of morphological features across different vertebrates and find an inverse hourglass-like pattern with the most variation during mid-development. Based on their analysis they argue for the existence of an inverse hourglass of conservation and conclude that there is no evidence for the hourglass model. In **Chapter 4** we discuss the need for appropriate controls and careful interpretation of comparative temporal analyses. By simulating data, we show that the methodology that Bininda-Emonds *et al.* use is vulnerable to edge effects, and already produces an inverse hourglass with time-series of constant temporal variation.

Bininda-Emonds *et al.* compare the rank of 116 morphological timings of developmental characteristics of 14 mammals and two amniotes with each other. This results in a table of 16 columns representing the species, and 116 rows, with each row consisting of integers in the range 1-116. These numbers represent the ranks of the timing of events within each species, which allows for the comparison of the relative timing between species. They then visualize two patterns based on these tables to estimate temporal conservation. For the first pattern, they visualize the mean rank of each morphological feature on the x-axis, and the standard error of each feature on the y-axis (Fig. 5.1A). This pattern exhibits a clear trend with a low standard error for features appearing early or late in development, and a high standard error for features that appear in mid-development. The authors note that this pattern could be caused by edge effects, and alternatively propose a second method that should not be vulnerable to edge effects. In this visualization, they calculate the phenotypic divergence of a group of features (PD) and visualize that against the mean rank of the group (Fig. 5.1B). The PD is an estimation of the diversity among species. Their visualization shows a clear curve with the highest PD during mid-development and the lowest divergence early and late during development. Both these results are in direct contradiction with the hourglass model, which caused Bininda-Emonds *et al.* to argue in favor of the inverse hourglass model between mammals. To test the claim of Bininda-Emonds *et al.* that the methodology is not vulnerable to edge effects, we generated 16 identical time series consisting of the numbers 1 to 116 in order. We added Gaussian noise to these numbers with a standard deviation of 14.5 (or $1/8^{th}$ of the time series). We then calculated the ranks per time series and ordered the features based on the average rank across species. The result is that we have 16 time series of 116 ranked features with no temporal pattern of stronger conservation at specific time points. We then apply the same methodology as Bininda-Emonds *et al.* and get nearly identical results (Fig. 5.1C-D). Despite the authors' claim that the PD visualization is not sensitive to edge effects, both the PD, and mean-standard error patterns can be fully explained by the edge effects of simulated data with no specific temporal conservation. This is caused by the effect that the most extreme timings (earliest and latest characteristics) have the smallest odds of changing their rank because an even more extreme timing does not change the ranking. As such, any set of conserved characteristics between species, with a constant level of temporal conservation, will show an inverse hourglass with these methods. Even though Bininda-Emonds *et al.* are rightfully frustrated by the lack of a definition of the phylotypic stage and quantitative proof for it, we see no evidence for their claim of an inverse hourglass based on morphological characteristics. Consequently, this analysis is consistent with our recommendations in **Chapter 4** for implementing proper controls in comparative research.

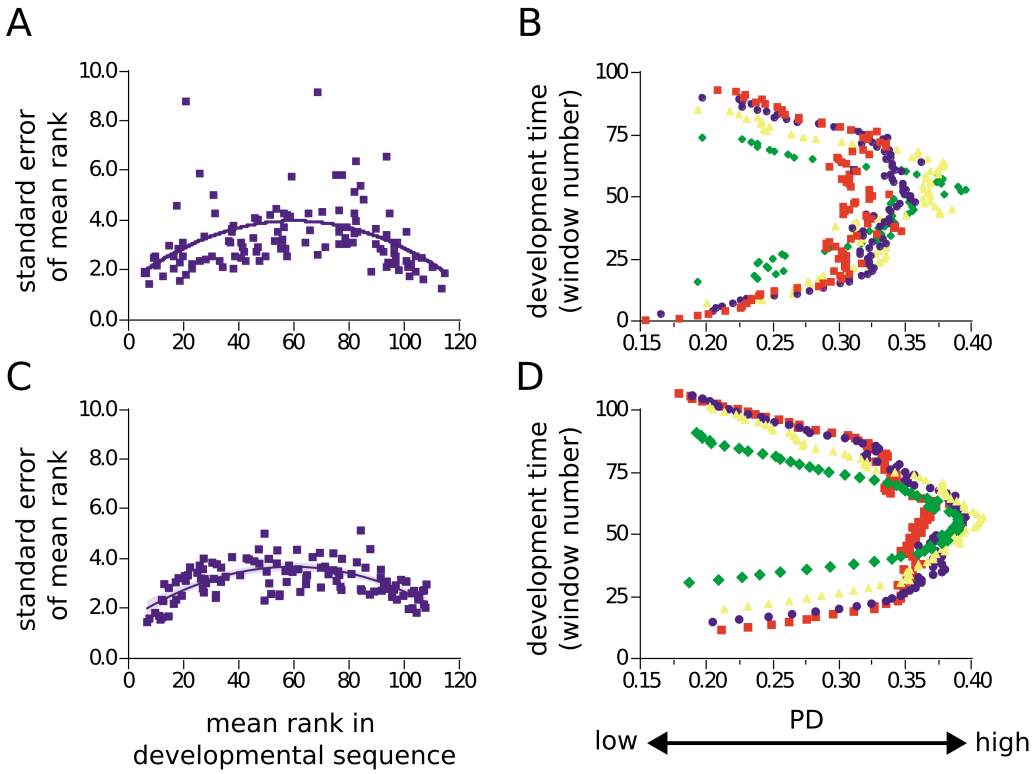


Figure 5.1: **Simulations without any temporal conservation pattern capture the same dynamics as data based on real timings of morphological characteristics. (A)** Figure taken directly from Bininda-Emonds *et al.*[318] which shows the standard error of the mean rank of a developmental event versus the mean rank based on real data. **(B)** Figure taken directly from Bininda-Emonds *et al.*[318] which shows overlapping windows (*i.e.* composed of events of ranks 1–5, 2–6, 3–7, ...) with durations of one-fifth (red squares), one-quarter (blue circles), one-third (yellow triangles) or one-half (green diamonds) of the length of the developmental time span examined based on real data. **(C-D)** are based on simulated data with no specific temporal conservation, and show the same patterns as the real data of panels A and B.

5.1 Methods

5.1.1 Phenotypic divergence

The PD is calculated identical to Bininda-Emonds *et al.*, in sliding windows of 1/5, 1/4, 1/3, and 1/2 of the total time series by the original authors to avoid artefacts arising from a single subjective placement of an event.

$$\text{events} = \{1^{st} \text{ somite pair}, 4^{th} \text{ somite pair}, \ldots, \text{ neural fold fusing}\}$$

$$PD_{\text{event}} = 1 - \frac{|n_{\text{present}} - n_{\text{absent}}|}{n_{\text{present}} + n_{\text{absent}}}$$

$$PD = \frac{1}{|\text{events}|} \sum_{\text{event}}^{\text{events}} PD_{\text{event}}$$

5.1.2 Simulated data

The simulation of the data was conducted using Python 3.10, leveraging libraries such as numpy and pandas for data manipulation, and matplotlib and seaborn for visualization.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

fig, axs = plt.subplots(ncols=2)

def make_random_data(n=10, t=100):
    # make n identical time series of t long
    xs = pd.DataFrame([np.arange(0, t) for _ in range(n)])

    # add gaussian noise
    xs +=  np.random.normal(0, t/8, xs.shape)

    # convert to ranks per time series
    xs = xs.rank(axis=1)

    # sort the features based on average occurrence
    xs = xs[xs.mean().sort_values().index]

    return xs

xs = pd.DataFrame(make_random_data(n=16, t=116))
sns.regplot(x=xs.mean(axis=0),
            y=xs.sem(axis=0),
            order=2,
            color="#4f247fff",
            marker="s",
            scatter_kws={"s": 60, "alpha":1},
            ax=axs[0]
)
axs[0].axis(xmin=0.0,xmax=120, ymin=0, ymax=10)

def phenotypic_diversity(xs):
    PD = []
    for i in range(1, 116):
```

```
        n_present = (i == xs).sum().sum()
        n_absent = xs.shape[0] - n_present

        PD_e = 1 - np.abs((n_present - n_absent) / (n_present + n_absent))
        PD.append(PD_e)
    return np.mean(PD)

colors = ["#e33d2a", "#4f247f", "#f1f178", "#009e3c"]
markers = ["s", "o", "^", "D"]
stepsizes = [xs.shape[1]//5, xs.shape[1]//4, xs.shape[1]//3, xs.shape[1]//2]

for stepsize, color, marker in zip(stepsizes, colors, markers):
    PDs = [phenotypic_diversity(xs.iloc[:, i:i+stepsize]) for i in range(0, xs.
    shape[1] - stepsize)]
    windows = [x + (xs.shape[1] - len(PDs)) // 2 for x in range(len(PDs))]
    axs[1].scatter(PDs, windows, c=color, s=60, marker=marker)
axs[1].axis(xmin=0.15,xmax=0.45, ymin=0, ymax=100)
```

Chapter 6

Unveiling Transcription Factor Dynamics in Cardiac Cells Using Single-Cell RNA-Seq and Epigenomic Data Integration

Rebecca R. Snabel*, Maarten van der Sande*, Simon J. van Heeringen

6.1 Abstract

Transcription Factors (TFs) bind to specific DNA motifs in cis-regulatory regions and regulate the gene expression of the involved genes. Certain epigenomic assays, such as DNA accessibility and H3K27ac ChIP-seq, coincide with cis-regulatory regions and are used to infer key transcription factor motifs between cell types. Conversely, RNA-seq assesses gene expression, allowing for the analysis of co-expressed regulons and variations in TF gene expression. Integrating epigenomic and transcriptomic data significantly enhances the identification of regulatory TFs. Here we present SCEPIA, a computational tool that predicts transcription factor motif activity based on single-cell RNA-seq. SCEPIA works by matching cells to a reference epigenomic database and infers motif activities based on their differential cis-regulatory regions. By correlating these activities to the gene expression of the corresponding TFs, SCEPIA is able to infer differential regulation of TFs between cells. We confirm the accuracy of the inferred motif activities by SCEPIA through a multiomic re-analysis of the Human Heart Cell Atlas. In conclusion, SCEPIA is a computational tool that improves motif activity inference based on single-cell RNA-seq, and is freely available at https://github.com/vanheeringen-lab/scepia.

6.2 Introduction

The gene activity of a cell is shaped by a complex network of interacting genes, with a special role reserved for transcription factors (TFs). TFs selectively bind to specific DNA sequences (motifs) in accessible cis-regulatory DNA and interact with other nearby transcription factors or recruit co-activators[9]. These co-activators chemically modify histone tails, where different modifications are associated with specific regulatory roles. The H3K27ac modification loosens the histone-DNA association, which in turn makes DNA more accessible[13]. More accessible DNA exposes new motifs, which allows for new TFs to bind to regulate gene expression[344]. Thus, to understand transcription regulation, it is crucial to consider the interaction between cis-regulatory regions and transcription factors[29,30,190,345–347]. The interactions between TFs, cis-regulatory regions, and their target genes are generally represented and studied as a gene regulatory network (GRN).

Single-cell sequencing has emerged as a rapidly expanding field, allowing researchers to explore the gene expression and epigenomic profiles of mixtures of cell types. One of the earliest methods to infer GRNs based on single-cell transcriptomics is the method SCENIC[188]. SCENIC matches TFs and target genes based on co-expression, and filters interactions based on TF motifs in cis-regulatory regions in proximity of the TSS of the target gene. Whilst co-expression based gene networks can be insightful, they suffer from spurious interactions. Spurious interactions arise because these methods cannot differentiate between co-expressed and co-regulated genes[348]. Furthermore, the correlation between transcript abundance and protein count is often weak[349,350], and transcriptomic methods are incapable of measuring post-translational modifications or considering chromatin context. As a consequence, single-cell multiomics GRN inference methods such as Pando[345], SCENIC+[190], CellOracle[346], and FigR[347] have been introduced. These multiomic-based methods leverage the chromatin context by determining TF and target gene interactions based on the motifs present in accessible regions and use the level of DNA accessibility as a proxy for regulatory activity. Although multiomic sequencing assays are immensely valuable, they remain costly and generate sparser data compared to their single-cell transcriptomics and bulk epigenomic counterparts[351]. Given the fact that specific epigenomic assays, such as ATAC-seq and H3K27ac ChIP-seq, demonstrate enrichment near actively expressed genes[352,353] and, conversely, show depletion near silent genes, there exists a potential for computationally aligning gene expression profiles with epigenomic assays from similar cell types.

Here we explore the question of whether transcription factor motif activity inference from single-

cell gene expression data can be improved by linking to an independently obtained epigenomic reference. This approach has two advantages. First, single-cell RNA-seq is cheaper and more straightforward to perform in many cell types than single-cell epigenomic or multiomic assays (*e.g.* scATAC, scCUT&Tag, 10X Multiome). Second, by linking these assays one effectively obtains multiomic data. This, in turn, allows for improved motif activity inference, by linking motif activities with gene expression and ultimately removing spurious interactions. We call this approach Single-cell EPigenome-based Inference of Activity (SCEPIA). Based on a re-analysis of the human Heart Cell Atlas (hHCA)[354], we demonstrate SCEPIA's ability to identify key transcriptional regulators of cardiac cell types (*i.e. GATA4*, *TBX5*, *FLI1* and *RUNX1*). Furthermore, this approach was previously shown instrumental in uncovering *Onecut2* as a regulator of M-cells in intestinal organoids[81]. With SCEPIA, we present a new methodological approach to uncover transcription regulators from single-cell transcriptomic data.

6.3 Results

6.3.1 Accurate motif activities by matching gene counts to an epigenomic reference database

We hypothesized that incorporating a measure of cis-regulatory element activity, such as ATAC-seq or H3K27ac ChIP-seq, would be beneficial to identify transcription factor motif activity from single-cell transcriptomic data, even when this activity is not experimentally measured in the same experiment. Here, we used the regulatory potential to match single-cell RNA-seq profiles to a collection of predetermined H3K27ac ChIP-seq profiles. The regulatory potential is defined as the weighted average H3K27ac signal per gene[352]. To determine how well regulatory potential specifically matches RNA-seq data for identical cell types, we obtained data from 96 human RNA-seq cell types and 121 human H3K27ac cell types from ENCODE[355]. We calculated the regulatory potential for all H3K27ac samples, and computed the correlation coefficient for all combinations of regulatory potential and RNA-seq values (Fig. 6.1A). In general, the regulatory potential shows only a high positive correlation for a subset of the transcriptomic data, indicating that the measure captures cell type-specific signals. The mean correlation coefficient for identical cell types between regulatory potential and transcripts per million (TPM) was found to be 0.53 with a standard deviation of 0.14. In 64% of the H3K27ac experiments, the regulatory potential of a given cell type exhibited the highest correlation with the TPM of the same cell type. Furthermore, in 77% of the cases, the highest correlation was observed with a tissue sharing the same ontology term (UBERON ontology for tissues, Cell Line ontology for primary cells, and cell lines are not mapped to an ontology term). As noted previously[352], the specific parameters used for calculating regulatory potential do not significantly impact the results. For this specific analysis, the H3K27ac signal in the promoter region (2kb up- and downstream of the TSS) alone is sufficient to predict the TPM (Table S6.1). In summary, our findings suggest that a collection of H3K27ac regulatory potential can effectively serve as a reliable classifier for characterizing cell states based on transcriptomic data.

In general, the H3K27ac signal in regulatory sequences can be used to determine motif or transcription factor activity[164,356,357]. This motif activity, which quantifies the contribution of each motif to H3K27ac peak strength, is a powerful measure of the relevance of a transcription factor for a specific cell state. The relationship between RNA-seq and regulatory potential led us to assume that transcription factor motif activity can be inferred from transcriptomic data, through an intermediate collection of matched H3K27ac data. As the reference collection contains a limited number of cell types, an exact matching cell type may not always be available for a specific cell type measured by RNA-seq. Therefore, we decided to calculate a composite measure of motif activity, based on a

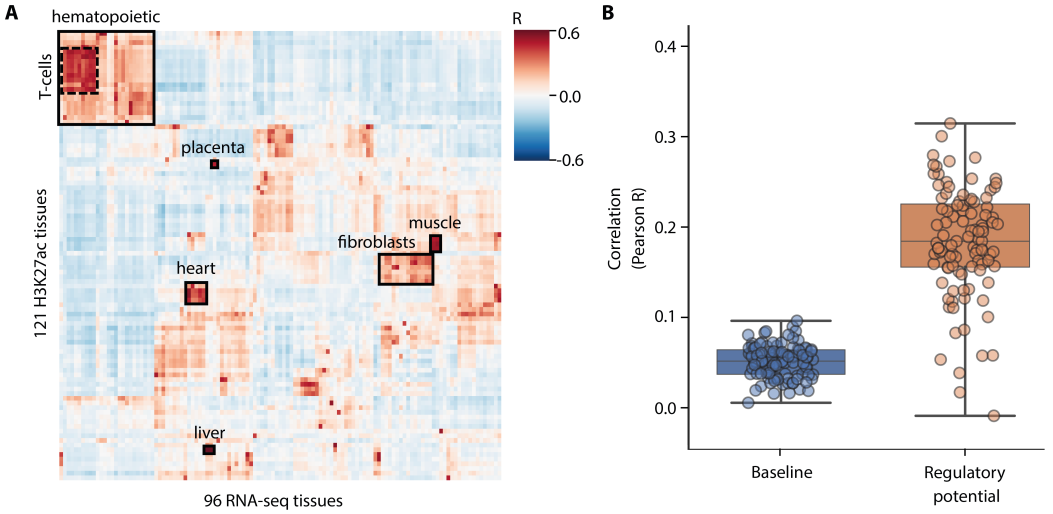


Figure 6.1: **Automatically matching epigenomic information to transcriptomic data improves motif activity inference. (A)** The Pearson correlation coefficients between all combinations of regulatory potential of 121 H3K27ac cell types and TPM values of 96 RNA cell types. Rows and columns are hierarchically sorted, which clusters similar cell types in proximity, for example, the hematopoietic, heart, and muscle cells. **(B)** Regulatory potential-based motif activities are superior to the baseline based on transcriptomic data alone. The motif activities of 5 random cell types are inferred based on the H3K27ac signal in the top 25,000 differential enhancers (100 permutations). The baseline approach to motif activity inference with only transcriptomic data is to assume that the TPMs of transcription factors are identical to their motif activity (mean correlation of 0.04). The regulatory potential-based approach automatically matches transcriptomic samples to H3K27ac samples in a hold-out set. It infers the motif activity scores on their top 10,000 differential enhancers (mean correlation of 0.17).

combination of reference cell types (see Methods). To calculate this measure, we regressed the TPM values of the 2,000 most variable genes of a collection of cell types against our regulatory potential database. From there, we selected the top 50 cell types based on the absolute regression coefficients and identified the top 10,000 most differentially enriched cis-regulatory regions within these 50 cell types. Subsequently, we calculated the motif activity using these enhancers and obtained the final motif scores by taking the dot product of motif scores and regression coefficients.

To test the accuracy of this method, we made one hundred random subsets of five tissues common to both our H3K27ac reference database and transcriptomic dataset. Our first step involved establishing a ground truth for motif activities by conducting a motif scan on the top 25,000 most differentially enriched cis-regulatory regions within each of these five tissues. Here, these regions were defined based on a union of all public transcription factor binding peaks, as obtained from the ReMap database[358] (see Methods). As a baseline approach to estimating motif activity based solely on transcriptomic data, we assumed a direct one-to-one translation between the transcripts per million (TPM) values of a transcription factor and its motif score. However, as Figure 6.1B illustrates, this baseline approach displays a particularly poor correlation with our ground truth (mean correlation of 0.04). Instead, our approach based on regulatory potential demonstrates a significantly improved correlation (mean correlation of 0.17). It is important to note that the cell types used for the ground truth are excluded from the reference database. Taken together, these

analyses show that RNA-seq can be reliably matched to H3K27ac in a cell type-specific manner and that this data can be used to estimate transcription factor activity in a more reliable manner than gene expression alone. See supplemental figure S6.2 for a one-factor-at-a-time parameter sweep, which shows that the results of SCEPIA are robust across a wide range of settings.

6.3.2 SCEPIA accurately annotates single-cell populations and identifies known cardiac regulators

We showed the possibility of matching cell type-specific H3K27ac-inferred regulatory potential to an RNA-seq dataset. SCEPIA was set up to incorporate this principle into single-cell transcriptomic analysis. Figure 6.2) shows a schematic overview of SCEPIA and individual steps are explained in more detail in the Methods. In short, the gene expression levels per cell type are inferred from the promoter accessibilities of the reference data and this inferred expression is used to match with the provided single-cell query to annotate the cells (steps 1 & 2). From the top 50 highest-ranking cell types in the annotation, the 10,000 most variable peaks are used to perform motif scanning (step 3). The cell type-annotation weights are used to determine single-cell motif activities (step 4), and by correlating motif activity and TF expression, the best TF-motif match is established (TF activity; step 5). Finally, in this last step, the significance of the hits is determined by permutation testing on randomized data. To evaluate performance, we conducted a comparative analysis on a multimodal dataset from the human Heart Cell Atlas v2 (hHCA;[354],) which encompasses over 700,000 cells profiled by scRNA-seq and spans the different anatomical regions and cellular subtypes of the human heart. We ran SCEPIA on the single cell transcriptomic assay of this atlas and used various strategies to evaluate the different steps incorporated in SCEPIA (Fig. 6.2). First, we ran SCEPIA on all cells of the atlas to asses its overall performance. We checked the intermediate outcomes of the run, such as the ability to annotate the cell type of the single cells correctly with the H3K27ac reference (outcomes of steps 1 & 2), as well as the inference of regulatory factors of importance (outcome of steps 3-5).

Accurate annotation of cell types by SCEPIA is important (Fig. 6.2: Step 2), as all subsequent analyses rely on this initial step. We demonstrated that this matching is accurate for bulk RNA-seq data (Fig. 6.1A). To validate if matching of the reference also works on single-cell data, we compared the inferred annotation of SCEPIA (Figure 6.3) with the cluster annotation as provided by the hHCA authors (Figure S6.1). SCEPIA correctly assigned the atrial and ventricular cardiomyocyte clusters to heart ventricle and cardiac atrium, respectively. The fibroblasts, neural cells and mural cells (*i.e.* smooth muscle cells or pericytes of the vasculature), all matched to tissues or cell types similar to their identity, albeit originating from a different organ or tissue (*i.e.* lung or tibia) (Figure 6.3A). Similarly, the myeloid and lymphoid clusters were matched with cell types from the correct major immune lineage branches, namely CD14+ monocytes and CD8+ T cells, respectively. SCEPIA had more difficulty annotating the endothelial cells, which showed the best match with heart left ventricle and a comparatively weaker match to umbilical vein endothelial cells (Fig. 6.4). This may be attributed to the high prevalence of endothelial cells in the heart myocardium, as also evident from the cluster size of the endothelial cells in the atlas. Ultimately, in the analysis by SCEPIA, the endothelial cluster was represented by a combined signature of heart left ventricle and an endothelial cell type. Likewise, the adipocytes are represented by a mixed signature of right cardiac atrium and adipose tissue (Figure 6.4). For the mesothelial cluster, there was no matching cell or tissue type available in the ENCODE database, which could explain the lack of any strong matches for this cluster. Overall, these findings highlight that SCEPIA matched the majority of the clusters correctly to the ENCODE reference.

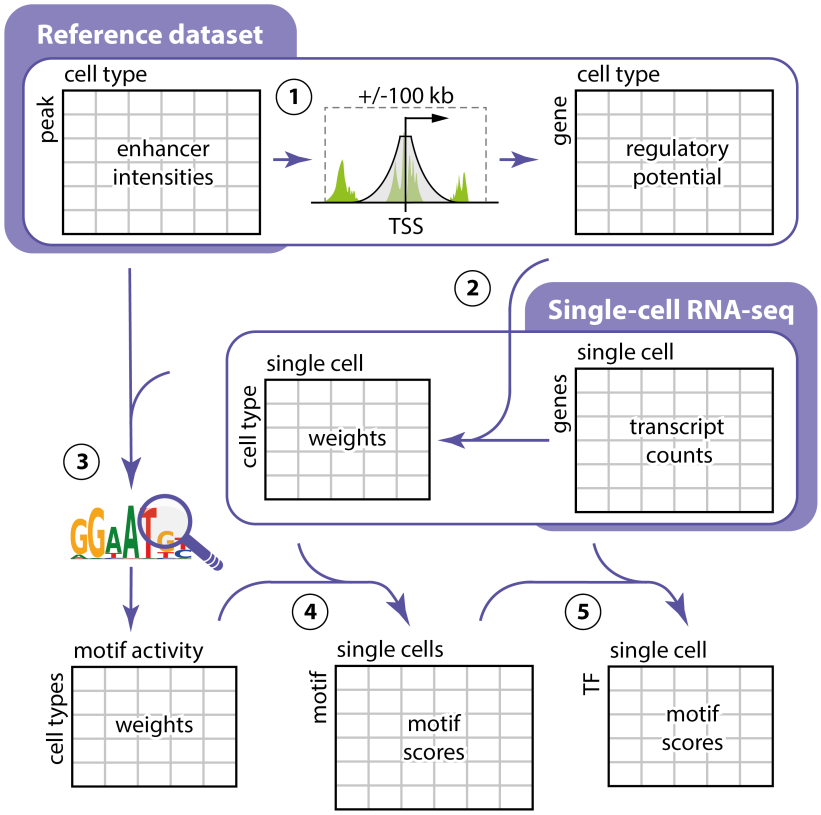


Figure 6.2: **Overview of Single-Cell EPigenome-based Inference of Activity (SCEPIA).** See the methods for an extensive explanation per step. **Step 1**: Calculate the regulatory potential from distance-weighted chromatin assay intensities around each gene for each cell type in the reference collection. **Step 2**: Regress the single-cell gene expression on the regulatory potential. Keep the top 50 cell types with the absolute highest regression coefficients to use as the cell type weights for each cell. **Step 3**: Select the top 10,000 most differential enhancers between the top 50 cell types, and use them for motif activity inference. **Step 4**: Take the dot product between the motif activities and the cell type weights, resulting in motif activities for each motif per cell. Motif activities are scaled to a motif score between 0-1. **Step 5**: Calculate the correlation coefficient between the expression level (log2 TPM) and the motif activity for TF-motif pairs. A permutation test between gene expression and motif activities allows for the estimation of the statistical significance of their relation.

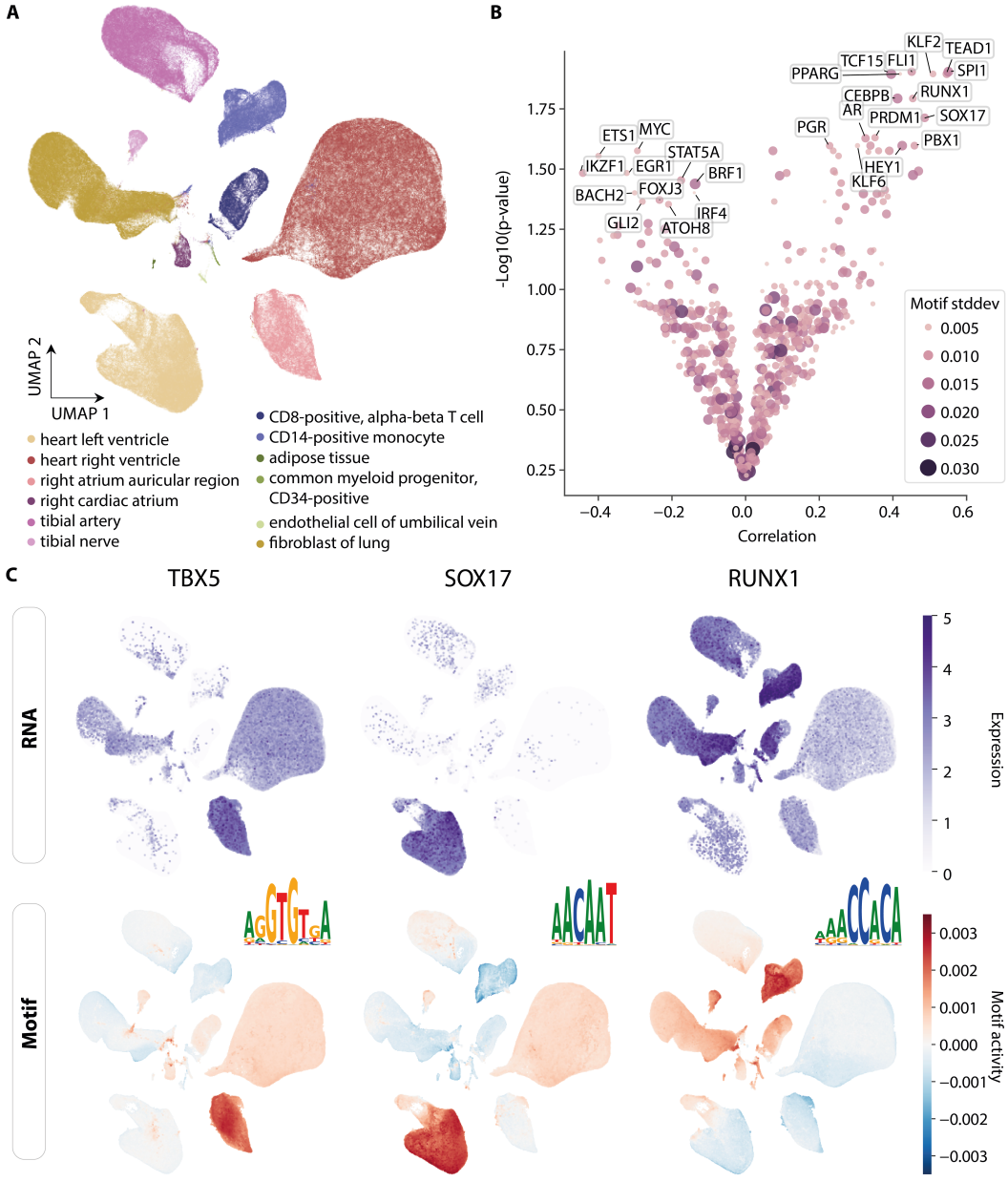

A
UMAP 2
UMAP 1
heart left ventricle
heart right ventricle
right atrium auricular region
right cardiac atrium
tibial artery
tibial nerve
CD8-positive, alpha-beta T cell
CD14-positive monocyte
adipose tissue
common myeloid progenitor, CD34-positive
endothelial cell of umbilical vein
fibroblast of lung
B
-Log10(p-value)
Correlation
Motif stddev
0.005
0.010
0.015
0.020
0.025
0.030
KLF2
TEAD1
TCF15
FLI1
SPI1
PPARG
CEBPB
RUNX1
AR
PRDM1
SOX17
PGR
PBX1
ETS1
MYC
STAT5A
IKZF1
EGR1
BRF1
HEY1
KLF6
BACH2
FOXJ3
IRF4
GLI2
ATOH8
C
TBX5
SOX17
RUNX1
RNA
Motif
Expression
Motif activity

Figure 6.3: **Transcription factor activity in Human Heart Cell Atlas predicted by SCEPIA. (A)** UMAP representation of all cells in the hHCA based on coordinates provided by the original study[354], with cluster annotation labels as predicted by SCEPIA. **(B)** Volcano plot of SCEPIA hits, where the x-axis is the correlation between motif and TF expression, and the y-axis is the log-likelihood of this correlation. The top 15 transcription factors are labeled, for putative positive regulators as well as putative repressors. Dot size and color labels are based on the motif score standard deviations. **(C)** Gene expression (top) and the corresponding predicted motif activity by SCEPIA (bottom), of three well-known markers: *TBX5* (cardiomyocytes), *SOX17* (endothelial cells), *RUNX1* (blood cells). Sequence logos of the binding motifs are indicated, with the GimmeMotifs database identifiers (left to right): GM.5.0.T-box.0005, GM.5.0.Sox.0021, GM.5.0.Runt.0003.

Following the initial annotation of cell types, we investigated the transcriptional regulators considered important by SCEPIA in these clusters. Among the top 15 hits, as ranked by their correlation coefficients, were several known markers for various cell types present in the atlas (Figure 6.3B). Examples include *FLI1*, *ETS1* and *KLF2* for immune and endothelial cells[359–361], *TBX5* for cardiomyocytes[362,363], *HEY1* for atrial cardiomyocytes[364], and *PPARG* for adipocytes[365]. SCEPIA correctly predicted the motif activity of *TBX5* in the cardiomyocyte clusters, which also aligned with its expression pattern (Figure 6.3C). Predicted motifs for *SOX17* and *RUNX1* exhibited variable motif activities, with expected higher activities in endothelial and lymphoid clusters, respectively (Figure 6.3C)[366–368]. *BACH2* is a known repressor and immune-regulating transcription factor and is also implicated in neural differentiation[369,370]. *BACH2* is most highly expressed in neural and lymphoid cells, where the motif activity (bZIP.0003) was decreased, and the reason for SCEPIA to identify this TF as a repressor (Figure S6.5A). However, the inferred motif activity of *BACH2* was increased in the myeloid and endothelial cluster, contradicting its established role as a repressor in myeloid cells. *BACH2* is known for its repressive regulatory functions in myeloid lineages[371], and was recently described for its activator role in B-lymphocytes[372]. This duality, seemed to pose a challenge for correct inference of the activity for this factor in these cell types. This could be attributed to the absence of particular cellular subtypes in the H3K27ac reference, restricting the full capture of *BACH2*'s complex transcriptional regulation in immune populations. To further confirm the regulatory role of *BACH2* in this dataset, we checked the expression level for putative target genes, based on ChIP-seq (Fig. S6.5B). The compound expression level of 57 predicted *BACH2* targets, showed patterns of expression similar to the motif activity as predicted by SCEPIA, and thereby for many clusters, the reversed of *BACH2* expression itself. In summary, the *BACH2* motif enrichment showed a similar pattern to the expression of its target genes, with a decreased motif activity in clusters exhibiting increased *BACH2* expression. These findings underscore the difficulty of distinguishing some of the more complex regulatory roles of TFs. However, altogether SCEPIA mostly identified relevant regulators, as well as a well-known repressive regulator.

6.3.3 Prioritization of cardiac regulators by SCEPIA improves with directed subsampling of abundant cell types

For a subset of the hHCA, chromatin accessibility data (scATAC-seq) is available. For this data set, chromatin accessibility was measured in the same nucleus as the transcriptomic data. This offered an ideal use case for benchmarking SCEPIA. The multimodality allows for a comprehensive evaluation of transcriptional regulators identified by SCEPIA, as the predicted motif activity can be compared to the motif activity based on the scATAC-seq data. First, we ran motif analysis on the pseudobulk of the scATAC-seq clusters using GimmeMotifs Maelstrom[139] (hereafter referred to as maelstrom analysis), to infer the cell type-specific motif activities. As mentioned, SCEPIA is

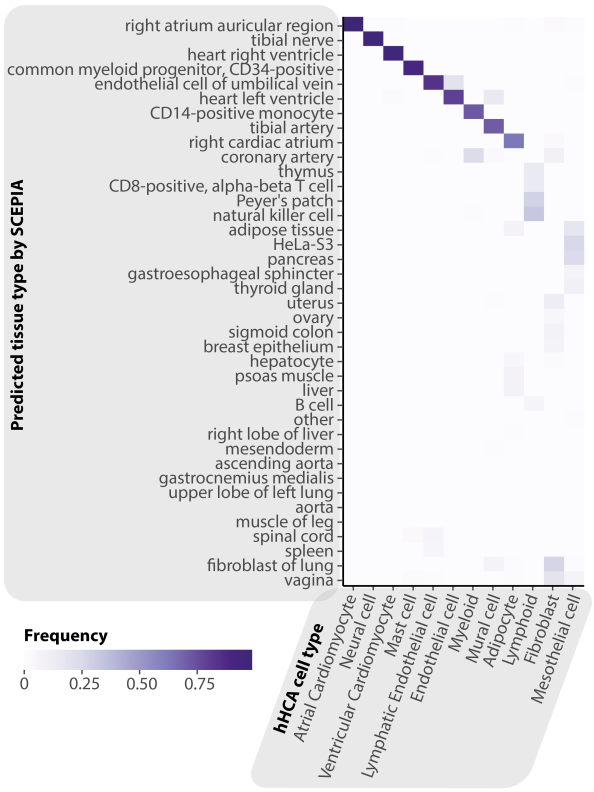


Figure 6.4: **Cell type annotations by SCEPIA.** The columns are the clusters as annotated by the hHCA, and the rows the cell type annotations by SCEPIA. The frequency is the number of cells per cluster annotated for each cell type. The cell annotations are used as weights by SCEPIA (See section 6.5.7, step 2 and 4).

tailored to infer enhancer activity from any epigenomic reference data, in this case, we used the ENCODE H3K27ac reference. Although the information on chromatin context between H3K27ac and ATAC-seq differs (active enhancers versus accessibility of chromatin, respectively), the motif analysis on the pseudobulk ATAC-seq served in this case as our closest approximation to a ground truth.

To evaluate SCEPIA's performance in which gene expression is used as a proxy to retrieve information on chromatin accessibility, we set out to acquire a baseline of correlation coefficients for this type of analyses. Since we have the hHCA with transcriptomic information, as well as the chromatin accessibility of the same cell, we correlated TF expression with the binding motif activities of its clusters. The comparison is solely based on predicted motif-TF links from the GimmeMotifs database without any further prioritization and will thereby illustrate a baseline of coefficients that can be achieved comparing this type of multiomic data. To enable multiple runs in our comparisons and evaluate reproducibility, we divided the scRNA-seq data into seven random subsets, each consisting of 100K cells. Next, we correlated the motif activities computed with the maelstrom analysis with the expression levels of each binding TF across the cell types, and for each subset. This comparison yielded coefficients ranging from a r of -0.02 to 0.26 (Fig. 6.5 blue) and reflected the degree of similarity between TF expression levels and their motif activities measured in a single-cell multimodal setting.

To benchmark SCEPIA against these results, we compared the SCEPIA runs on each of these seven subsets to the maelstrom analysis. For all the significant SCEPIA hits (p-adj < 0.05), we averaged their inferred motif activities per cell type cluster. These averages were correlated to the activities for the same motifs from the maelstrom analysis. Interestingly, the mean coefficients per cell type in these correlations were consistently higher than those found in the baseline comparison, except for the mesothelial and mural cell cluster (Fig. 6.5 orange, Table S6.4). However, these scores exhibited substantial differences across the cell types, with a median r ranging from 0.14 for the mesothelial to 0.58 for the myeloid cluster (Table S6.4).

As we observed variations in the performance across the cell types, with some scoring lower (*e.g.* cardiomyocyte clusters, fibroblasts and endothelial cells, with a median r < 0.4), than others (myeloid and mast cell clusters, median r > 0.4), we implemented a step of geometric sketching. Geometric sketching is a method that subsamples the cells in the dataset while preserving the inherent heterogeneity, or geometry, of the dataset[373]. Since a geometric sketch will lead to a more equal representation of the different cell types across the dataset, including rare cell types, we included this subsampling step during the pre-processing of the data and preceding the SCEPIA run on each of the seven subsets (see Methods). This significantly improved the correlation coefficients with the maelstrom results for both cardiomyocyte and endothelial clusters, as well as the myeloid cluster (Fig. 6.5 green, Table S6.4). The neural and adipocyte, mural and mesothelial cell-clusters displayed more variability in their coefficients, with some obtaining equal or worse coefficients after the addition of the geosketching. While the geometric sketch could not fully alleviate the relatively low performance of SCEPIA for certain cell types, incorporating this step significantly improved the performance for five out of the twelve cell clusters in this specific benchmark.

6.3.4 SCEPIA identifies additional cardiac regulators compared to motif analysis in ATAC-seq

To validate our approach in selecting regulatory factors based on motif activity and TF expression patterns, we compared the outcomes of SCEPIA with a selection based on motif activity alone. Given the improved performance of SCEPIA across various cell types with the inclusion of a step of geosketching, we further examined its efficacy by downsizing the full dataset to a geosketch of 20K cells and reran SCEPIA (Fig. S6.3). We then explored biological differences between the

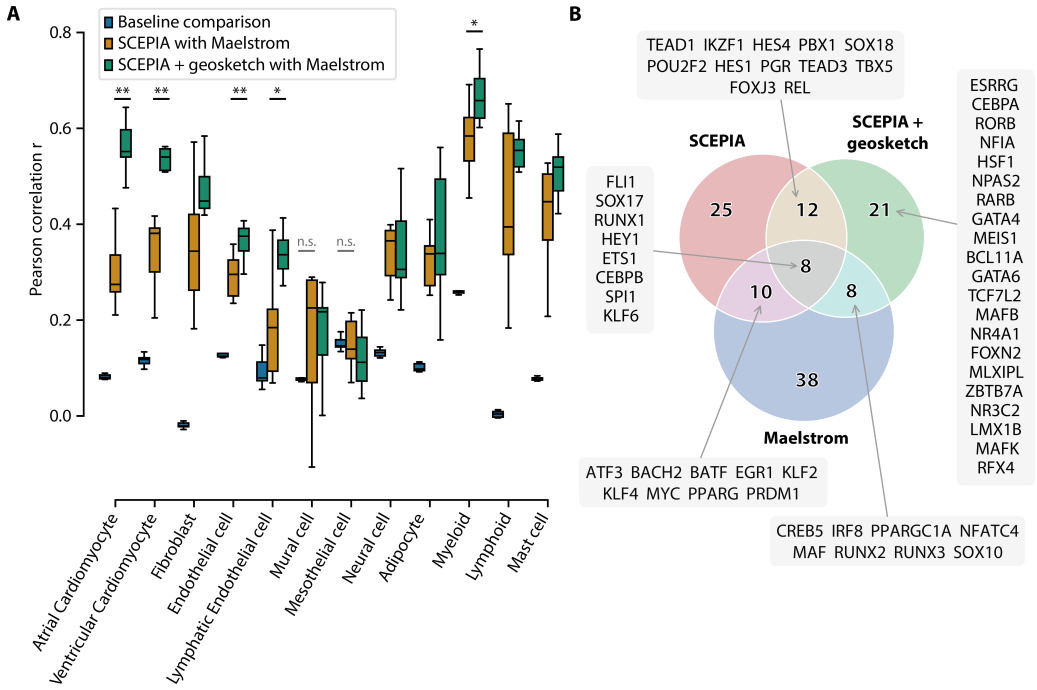


Figure 6.5: **Benchmarking SCEPIA results between repeated runs and with a geometric sketch of the data** (**A**) Correlations between motif activity computed in pseudobulk ATAC-seq fraction (maelstrom analysis) and expression values of the corresponding TFs from each of the seven 100K cells scRNA-seq subsets (blue; Baseline correlations). In orange the motif activities of significant (p-adjusted < 0.05) SCEPIA hits from each of the seven 100K subset analyses, were correlated with the matched motif activities computed with the maelstrom analysis (SCEPIA with Maelstrom). Green represents the correlation coefficients of the comparisons with maelstrom motif activities, for the seven 100K subset runs with SCEPIA with prior pre-processing with geometrical sketch subsampling of each of the seven 100K cells subsets (SCEPIA+geosketch with Maelstrom). Significance of differences between conditions was calculated with the one-sided Wilcoxon signed-rank test. Only non-significant results are indicated for the baseline versus SCEPIA comparison in grey, and only significant results indicated for the SCEPIA versus SCEPIA+geosketch comparisons (all p-values can be found in Table S6.4). **: p-value <= 0.01, *: p-value <= 0.05, n.s.: non-significant. (**B**) Venn diagram presenting overlaps and differences between the SCEPIA runs on the full dataset (700K cells of the hHCA), the SCEPIA run with geosketch subsampling of the full dataset (subsampling 700K to 20K cells) and the maelstrom analysis on the pseudobulk of the scATAC-seq data of the hHCA. hHCA = human Heart Cell Atlas.

methods.
Of the 49 total hits, the geosketched SCEPIA run yielded 20 overlapping hits with the full dataset run and a partial overlap with TFs linked to maelstrom motif hits (16 hits; Figure 6.5B). Common hits, such as *FLI1* and *SOX17*, implicated in cardiac cells, as well as *HEY1* (cardiac progenitor marker) and *ETS1* (lymphoid marker), reaffirmed SCEPIAs consistency. The maelstrom analysis pointed to variable motif activity for a BACH2-binding motif and indicated higher activity in the fibroblast, endothelial and mesothelial clusters. The first SCEPIA run inferred increased motif activity for *BACH2* in these same clusters, with additional high levels in the myeloid and mural cells (Figure S6.5A). However, differences also emerged between the runs. *GATA4* and *GATA6* for example, vital for cardiac (mesoderm) development and inferred with increased motif activities in the cardiomyocyte clusters, were exclusively found in the run on the data preprocessed with geosketch (Figure S6.5C)[374,375]. Notably, *GATA1* was only identified in the non-geosketched run, showing specific expression and increased motif activity in mast cells (Figure S6.5D), a cell type-specificity confirmed in literature[376,377]. *RUNX1* was consistent, while *RUNX2* and *RUNX3* were additionally identified in the geosketch run, all linked by SCEPIA to the same motif (Runt.0003). The motif activity changed to be more restricted to immune cells in the geosketch run, whilst in the run without geosketch this motif activity was additionally inferred across, for instance, the mural, neural and fibroblast cells (Figure S6.5E, 6.3C).
The overlapping hits between geosketch SCEPIA and the maelstrom analysis revealed specific points of interest. Of the *SOX* gene family, *SOX17* and *SOX18* emerged in the initial SCEPIA run, associated with motifs that exhibited similar activities across the clusters. The geosketch SCEPIA analysis additionally revealed *SOX10* in the neural and, to a lesser extent, endothelial cells (Figure S6.1F). Similarily, maelstrom identified two *SOX10* motifs ranking among the top three in the neural cluster (specifically Sox.0018 and Sox.007; Figure S6.4). This cluster was highlighted in the original paper for its pan-glial marker expression. Therefore these *SOX10* results are in line with its significance in neural crest development and pivotal role in differentiation to neuronal or glial cells[378]. Furthermore, SCEPIA not only confirmed a similar motif specificity for *SOX10* in this dataset, but also identified a motif sequence which corresponds to a subset of the motifs inferred by maelstrom (Figure S6.5G). Differences included *IKZF1*, identified by SCEPIA with an inferred activity in almost all clusters, with only a specific depletion of signal in the immune clusters (Fig. S6.6B, D), aligning with its dual repressor and activator role in immune cells[379]. Other examples were *TBX5* and *MEIS1*, both well-known cardiac mesoderm markers, alongside *ESRRG*. Binding motifs for these factors were inferred with an activity across multiple, if not all, clusters in the dataset. Only by additionally considering the expression levels of the binding TFs, a specificity became apparent (example of *ESRRG* in figure S6.6A, C). Conversely, there were also hits in Maelstrom, that were not identified in any of the SCEPIA runs. A notable discrepancy involved the *MEF2* transcription factor family. Maelstrom analysis showed motif enrichment in cardiomyocytes and mural cell clusters, however, the expression levels of *MEF2A-D* were consistent across the majority of cell types (Fig. S6.6E) and showed low correlations with their inferred motif activities (Tables S6.2 and S6.3), which could explain the absence in SCEPIAs hits. These findings suggest that SCEPIA is better suited in identifying factors with ubiquitous motif activities, because of the inclusion of expression level information in selecting the factors. Conversely, if the expression of the factor is not highly variable or not well (anti-)correlated with the predicted motif activities, SCEPIA will overlook these factors, even if the binding motifs were inferred with variable activities over the cell types.
Several mesothelial regulators mentioned by the authors of the hHCA, *WT1* and *BCN1*, or other epicardial marker genes (*TBX18* and *TCF21*) were not found in our maelstrom analysis of the hHCA scATAC-seq data[380]. This could explain the difficulty in predicting the appropriate regulators for this and similarly small clusters, which are lost by performing the analysis on this whole dataset at once. This could also explain why some of the cell types are not performing well compared to

more abundant clusters in the dataset. Looking at the cluster sizes in the data: less than 2000 cells were sampled for both the mural and mesothelial cluster, whereas the two smallest clusters after these, the neural and adipocyte clusters, are three times as big.

6.4 Discussion

In this study, we present SCEPIA, a novel computational method for inferring transcription factor motif activities using single-cell transcriptomic data in combination with a reference H3K27ac collection. SCEPIA successfully identifies established key regulators in various cardiac cell types, including endothelial cells (*FLI1*), cardiomyocytes (*TBX5*, *HEY1*), adipocytes (*PPARG*), and immune cells (*RUNX1*), in addition to recognizing the suppressor *BACH2*. This method stands out by leveraging pre-existing bulk epigenomic data, thus augmenting single-cell transcriptomics datasets and creating a multimodal resource for the identification of transcription regulators.
SCEPIA matches cells to a combination of known H3K27ac signatures, thereby establishing new H3K27ac identities that are not explicitly defined in the reference. In the case of cardiac endothelial cells and adipocytes, which were both missing from the reference, this approach has demonstrated that it can identify appropriate key TFs (*e.g. FLI1* and *PPARG*). However, we faced challenges with the identification of key TFs for other cell types that were not in the reference, such as mesothelial or mural cells. This difficulty likely arises from two main issues: the lack of closely related cell types in the reference dataset (for example, epicardial cells) and the low number of these cells in the cardiac tissue samples we studied. One solution for this could potentially be found in expanding on the current ChIP-seq reference in SCEPIA. Expanding the provided references with other epigenomic collections, either generated by the researcher themselves or using readily available and vast collections (*e.g.* SRA) could greatly enrich SCEPIA's reference database and potentially improve its accuracy and applicability, also for more rare subtypes. Alternatively, one could subset the data even further, to only include a subset of specific cell types.
The transcription factor *MEF2C* plays a crucial role in chromatin remodeling which increases DNA accessibility in cardiomyocytes[381–383] and during neural development[384]. Despite its important role, SCEPIA did not identify *MEF2C* or other members of the *MEF2* family, as key cardiac regulators. However, in our comparison based on single-cell DNA accessibility alone, their motif was flagged for its differential inferred activity. Even though we observed a variable gene expression of *MEF2C* across different cell types, there was a poor match between the gene expression levels and predicted motif activity. This exemplifies the complex nature of the role of *MEF2C* in regulating genes during heart development. *MEF2C* is known to produce several isoforms that regulate different genes[385]. Additionally, *MEF2* factors often form hetero- or homodimers to more precisely regulate their target genes[383,386]. The complexity of the *MEF2* gene regulatory function demonstrates the drawback of transcriptomic-based regulator inference methods, including SCEPIA. These methods cannot account for regulatory events that occur following transcription, and as such cannot fully capture the regulatory roles of transcription factors like for instance, *MEF2C*.
Establishing an accurate baseline and ground truth for GRN inference presents a significant challenge, which extends to benchmarking the results of SCEPIA. For the single-cell comparisons, our ground truth is calculated on regression coefficients based on a measure of DNA accessibility. However, SCEPIA uses a different reference, based on H3K27ac, a marker for active cis-regulatory regions. Whilst DNA accessibility and H3K27ac generally coincide, these two assays measure distinct aspects of chromatin context and gene regulation. For instance, our SCEPIA analysis identified nuclear receptor motifs, such as *TBX5* and *GATA4*, using the H3K27ac reference. These motifs were not detected based on our analysis of scATAC-seq data. This discrepancy can be attributed to the nature of these transcription factors: *TBX5* and *GATA4* bind in accessible chromatin, and this process is positively or negatively influenced by the presence of the other factor as well as the chro-

matin context, as shown during cardiac reprogramming[381]. Therefore, ideally, both the reference database and the ground truth are based on the same assay. Additionally, the motif activities based on DNA accessibility in SCEPIA are inferred by the ensemble computational method GimmeMotifs Maelstrom[139]. As a consequence, the ground truth of this study is based on a computational analysis rather than an experimentally validated ground truth. Lastly, by only correlating the activities for the exact same motifs between the maelstrom analysis and the SCEPIA outcomes, we underestimate SCEPIAs performance. The activity of a TF will be shared over all of its potential binding motifs in the regressions performed, and only the motif activity that correlates best with the TF's expression, will in the end be linked to the TF. Meaning, there is no guarantee, or need, for these motifs to be exactly the same in both analyses.

We foresee several ways to improve the current implementation of SCEPIA. First,the appropriate selection of highly variable genes and relevant cells can be improved. Currently, SCEPIA selects the top 2,000 genes based on dispersion-normalized gene expression, which may inadvertently bias the analysis towards overrepresented cell types. Potentially more sophisticated approaches could be used, such as comparing the highly variable genes between clusters, performing geometry-preserving sampling (geosketch)[373], or selecting genes based on selectively detected or undetected expression in a particular cell neighbourhood (SEMITONES)[387]. Cell sampling methods that represent all distinct cell types, such as geosketch, have as an added benefit that the TF-motif significance calculation will be improved. Currently, TF-motif significance is calculated with a permutation test of the correlation between TF and motif. The presence of large groups of similar cells influence this calculation, therefore a more representative subsampling of the data could be beneficial. Second, SCEPIA could be improved by incorporating additional genomic assays. Given the similarities between H3K27ac and ATAC-seq data, integrating an extensive ATAC-seq collection seems promising. SCEPIA already includes pre-built references from diverse sources, such as scATAC-seq datasets of fetal human cell types[388] and various adult mouse tissues[389]. Third,the permutation test on the correlation coefficient between TF and motif activity can be refined. The relationship between TF gene expression and its motif activity is often unclear. Instead, a more informative relationship can be found between motif activity and the gene expression of downstream genes. Given that SCEPIA includes a comprehensive reference of cis-regulatory regions, these regions can be associated to their nearest gene, although this approach has its limitations[176]. By establishing this association, we can link motif activities directly with specific genes, thereby basing the significance test on a more direct and relevant relationship. Moreover, combining these relationships with the expression of the TF itself, provides a clearer understanding of the functional impact of TFs within regulatory networks.

Single-cell omics datasets provide a detailed view of transcription regulation unique to each cell type, enabling the exploration of lesser-known and more challenging-to-obtain populations within tissues. Given that heart and vascular diseases are among the leading causes of death worldwide[390], it is crucial to better understand epigenomic (dis)regulation at a cellular level. Our study introduces SCEPIA, a powerful computational tool designed to enrich single-cell transcriptomic data with transcription factor motif activity. This enhancement facilitates the precise identification of regulators specific to different cell types. The capability of SCEPIA to integrate detailed transcription factor information into single-cell transcriptomics paves the way for a more comprehensive understanding of how transcription is regulated in various cell types.

6.5 Methods

6.5.1 Overview of public data

Sequencing data

Full sample tables can be obtained from https://zenodo.org/doi/10.5281/zenodo.10457767.

Cis-regulatory regions

We used a collection of cis-regulatory regions, based on all human transcription factor ChIP-seq peaks from ReMap 2018[358] (http://remap.univ-amu.fr/storage/remap2018/hg38/MACS/remap2018_all_macs2_hg38_v1_2.bed.gz), as described previously[29]. This collection of cis-regulatory regions is available at Zenodo (https://zenodo.org/records/4066424).

Human heart cell atlas

The human heart cell atlas transcriptomes covered 704,296 cells and nuclei and was obtained as a matrix with log-normalized counts[391]. Part of the atlas was sampled with 10X Multiome, and thereby consisting of scATAC-seq assay covering 144,762 nuclei, which was downloaded as a raw peak count matrix[392].

6.5.2 RNA-seq processing

Preprocessing of RNA-seq was done automatically by seq2science v1.0.3[337] using the RNA-seq workflow. Public samples were downloaded from the Sequence Read Archive[262] with the help of the NCBI e-utilities and pysradb[266]. Genome assembly GRCh38.p13 was downloaded with genomepy 0.16.1[267]. Paired-end reads were trimmed with fastp v0.23.2[269] with default options. Reads were aligned with STAR v2.7.10b[338] with default options. Afterwards, duplicate reads were marked with Picard MarkDuplicates v3.0.0[280]. BAM files were converted to CRAM format with samtools v1.16[339]. Read counting and summarizing to gene level was performed on filtered BAM files using HTSeq-count v2.0.2[288]. TPM-normalized gene counts were generated using genomepy based on longest transcript lengths.

6.5.3 H3K27ac ChIP-seq processing

We downloaded 327 H3K27ac aligned samples from the ENCODE data portal[355], spread over 121 cell types. For each sample, the coverage over the remap peaks was computed by the GimmeMotifs coverage_table command[139]. The values between samples was quantile normalized[342], and consequently log2(x+1) normalized.

6.5.4 Regulatory potential

The cell type-specific regulatory potential P of gene g is calculated similar to[352]:

$$P_g = \sum_k w_k s_{k,g}$$

where w_k is the weight at position k, and $s_{k,g}$ is the h3k27ac signal at position k for gene g.
The weight w_k is calculated identically to the method ANANSE[29]:

$$w_k = \begin{cases} 1, & \text{if } k \in (0\,\text{kb},\ 5\,\text{kb}] \\ \frac{2e^{-\mu|k-t_g|}}{1+e^{-\mu|k-t_g|}}, & \text{if } k \in (5\,\text{kb},\ 100\,\text{kb}] \end{cases}$$

where parameter t_g is the genomic position of the TSS of gene g, and μ determines the decay rate as a function of distance from the TSS, set such that an enhancer 10 kb from the TSS contributes one-half of an enhancer within 5 kb from TSS. t_g is the distance from the TSS.

6.5.5 Regulatory motif analysis and motif and transcription factor activity

In the regulatory potential benchmark (Fig. 6.1B) and in SCEPIA we use the motif activity as a measure of motif and transcription factor importance. The motif activity[356,357] is calculated using Bayesian ridge regression implemented in gimmemotifs[139], with the motif log-odds scores as features and the H3K27ac signal as predictor variable. In short, for each cis-regulatory region in the input we compute the motif log-odds score for each motif in the gimmemotifs database (gimme.vertebrate.v5.0). This motif databases contains a non-redundant collection of vertebrate transcription factor motifs[139]. We assume that the H3K27ac signal in each enhancer, expressed as log2(nr of reads + 1), is the sum of all motif log-odds scores multiplied by their respective motif weights. We then estimated these motif weights using Bayesian ridge regression, where the regularization parameter (λ) was determined through 5-fold cross-validation. These motif weights, the feature coefficients from the fitted regression model, are used as a measure of motif importance, the motif activity. This measure can then be used as transcription factor activity based on the TFs that are predicted to bind to the motif.

6.5.6 Benchmark of regulatory potential-based cell type and motif assignment.

To assess the performance of the regulatory potential-based approach compared to the actual H3K27ac signal, we conducted a systematic analysis. Our approach involved data subsampling and calculating correlation coefficients to evaluate the relationship between the ground truth and the regulatory potential-based approach. Specifically, this analysis used two data sources: the bulk H3K27ac reference database encompassing 121 tissues from ENCODE and 1,268,775 REMAP peaks, as well as the RNA-seq database featuring 96 tissues from ENCODE.

To establish a ground truth for comparison, we calculated the motif activity using the top 25,000 enhancers with the highest coefficient of variation with GimmeMotifs version 0.18.0[139]. We selected random subsets of five tissues from tissues shared between the RNA-seq database and the H3K27ac reference database, and calculated the Pearson correlation coefficients between the estimated motif activity and the ground truth motif activity for each tissue in each comparison.

As our baseline approach to estimating motif scores, we consider the transcripts per million (TPM) of a transcription factor directly as the motif score. When multiple transcription factors are associated with a single motif, we use their average expression.

In the regulatory potential-based approach, we exclude the five ground truth tissues from the reference database. We then convert the H3K27ac signal from the reference database into regulatory potential and subsequently perform regression analysis against the TPM values. This results in a (5 x nr cell types) table of regression weights. We select the top 50 tissues with the highest absolute regression weights and identify the top 10,000 differential enhancers between them. We then conduct a motif scan similar to the one performed for the ground truth, regressing the H3K27ac signal with the motif scores in these enhancers. Finally, the motif scores for our original five tissues are calculated by taking the dot product of the motif scores in the top 50 tissues and the tissue weights. This entire process was repeated one hundred times to generate a distribution of correlation coefficients, providing an estimate of the performance of each approach.

6.5.7 Detailed description of the implementation of SCEPIA

The single-cell regulatory-based approach (SCEPIA) is similar to the bulk approach. However, due to the increase in data, some steps need to be altered from a computational resource point of view. In addition, the fact that a single-cell dataset usually consists of multiple related cell types, with a more fluent gradient in gene expression and thus motif scores, makes it so that this information can be used to infer the significance of differential transcription factors based on motif scores and gene expression data.
Required input:

- Reference database matrix of peak intensities (D) with dimensions (peaks x cell types). SCEPIA comes with multiple extensive reference databases, and the user does not need to provide these themselves. These reference databases can be extended, however.
  - The default human reference is based on REMAP 2018[358], and consists of 1,268,775 putative enhancers. It contains 121 ENCODE[355] cell types. The number of reads in these enhancers per cell line was log2(x + 1) transformed, and quantile normalized[342] to enforce the same distribution.
  - The default reference database is based on H3K27ac signal, but this can be any chromatin mark associated with regulatory activity, for example, ATAC-seq.
- Single-cell RNA-seq dataset S with dimensions (cells x genes).

The single-cell epigenome-based inference of activity can be divided into five main steps (see Fig. 6.2):

1. **Conversion of Reference Database Matrix into Regulatory Potential**

   Convert the reference database matrix of peak intensities D into a database matrix of regulatory potential[352] per gene P with dimensions (genes $\times$ cell types). The reference database includes all REMAP peaks, including promoters, and is prepared beforehand. See section 6.5.4 for a detailed explanation of the calculation of regulatory potential.

2. **Cell Annotation from Single-Cell Dataset**

   Match cells in the single-cell dataset S with regulatory potential P, resulting in an annotation matrix A with dimensions (nr cells x nr cell types). The transcript counts of each cell are regressed against the regulatory potential database. The annotation matrix represents the regression coefficients, and cells receive a tissue or cell type annotation based on the highest regression coefficient.

   - The first step involves selecting a subset of relevant cell types to speed up the cell annotation. It assumes that the user has already performed Louvain or Leiden clustering, and averages the counts per cluster. The top 2,000 most variable genes (dispersion normalized) are chosen. For each cluster c, the regression coefficients are calculated by lasso regression:

     $$\underset{A_c}{\operatorname{argmin}} |S_c - PA_c| + \lambda |A_c|$$

     Where the regularization parameter (λ) is estimated through 5-fold cross-validation.

     Absolute regression weights are summed per tissue/cell type, and the top 50 cell types/tissues are retained in the regulatory potential database.

- Mean center the single-cell counts and set each cell as the mean expression value of its neighbors. For each cell i, the regression coefficients are calculated using Bayesian ridge regression with the top 50 cluster weights:

$$\underset{A_i}{\operatorname{argmin}} |S_i - P \cdot A_i|^2 + \lambda |A_i|^2$$

- Cells are initially assigned the cell type or tissue with the highest weight, and clusters are annotated based on the most common cell type in that cluster. Cell types are further refined by taking the dot product of cell type weights with neighborhood weights. At least 50 cells need to be assigned a specific cell type; otherwise, they are labeled as "other".

3. **Motif Activity Inference over Differential Enhancers**

   Select the to 10,000 enhancers with the highest variance between the annotated cell types. The resulting vector is denoted as E and has dimensions (10,000 x nr cell types). Infer the motif activities over these enhancers.

   - Only known motifs are considered. By default, the GimmeMotifs[139] vertebrate v5.0 motif database is used.
   - Scan and keep the maximal motif score in each enhancer. The result is a vector O with dimensions (10,000 x nr of motifs).
   - Motif activities (M) are calculated by Bayesian ridge regression:

$$\underset{M}{\operatorname{argmin}} |E - O \cdot M|^2 + \lambda |M|^2$$

     where the M has dimensions (nr of motifs x nr of cell types).

4. **Calculation Per Cell Motif Activities**

   Calculate motif activities for the cells based on the motif activities of the reference top tissues. The motif activities per cell are calculated as a dot product of the cell type annotation weight and motif scores per reference cell type.

$$F = M \cdot A^T$$

5. **Correlation Analysis of Motif and Transcript Scores**:

   Determine significant combinations of motif activities and transcription factor expression by correlating motif scores (F) and transcript scores (S) between cells:

   - Calculate the correlation coefficient between motif score and transcript counts.
   - Randomly shuffle motif scores and calculate their correlation with transcript counts. Repeat this 100,000 times to obtain a distribution of correlation coefficients.
   - Estimate two different p-values per TF-motif combination from this analysis: one based on the correlation coefficient relative to the total permuted set and another using only the permuted set of motif correlations. Combine these p-values using Fisher's method.
   - Calculate motif activity by fitting a Gaussian mixture model with two components over the motif activity scores. These components represent "high" and "low" motif scores. Motif activity is computed as the probability that a motif score belongs to the "high" expressed group, and is thus constrained to the range of 0 to 1.

6.5.8 Human Heart Cell atlas Single-cell comparison

Human Heart Cell Atlas single-cell comparison

We have used the full global log-normalised counts-containing h5ad object of the human Heart Cell Atlas v2 (hHCA) project as provided by the authors ([354]), and pre-processed the data with Scanpy (v1.9.2). Cell type assignment and UMAP-embeddings were used as provided by the authors. The dataset was subsetted to contain only highly variable genes (selected using the default settings within Scanpy; "min_mean = 0.0125, max_mean = 3, min_disp = 0.5") and the normalized counts were scaled. Both the neighborhood connectivities of the single cells, as well as PCA were rerun, as necessary input for SCEPIA. The infer_motifs method of SCEPIA was run on the preprocessed data with the following settings: using the top 2000 highly variable genes, the top 10,000 most variable enhancers and a maximum of 50 cell types from the reference and ridge regression was used for motif activity analysis. The ENCODE H3K27ac human reference dataset and gimme.vertebrate.v5.0.pfm motif file from the GimmeMotifs were used as input. The resulting data was visualized with seaborn v0.12.2, matplotlib v3.8.0 and gimme logo from GimmeMotifs (v0.18.0). *BACH2* target genes were taken from the ChIP-Atlas resource dataset[393–395], and selected for a distance to TSS of ±1kb and an average MACS2 binding score of > 100. The target genes selected this way, a compound gene score was calculated with the scanpy implementation of the *score_genes* functionality, which calculates the average expression of *BACH2*'s 57 target genes in a cell, and substracts this with the average of a reference set of genes randomly sampled from each binned expression level (577 genes were selected as reference).

Motif scanning on the scATAC-seq fraction of the Heart Cell Atlas

The ATAC peak matrix object as provided by the authors, was normalized per cell for sequencing depth (multiplied by 10,000) and log1p transformed, the resulting dataset was averaged per cell type. Only peaks outside of promoter regions (2,000 bp up- and downstream of transcription start sites) were kept, to select for enhancer regions. The top 200,000 most variable enhancers were selected and the z-scores per peak and across the cell type means were calculated. The scaled matrix was used as input for gimme maelstrom from the GimmeMotifs package (v0.18.0), the same motif position frequency matrix file as used in SCEPIA (gimme.vertebrate.v5.0.pfm), and hg38 as the reference genome. Maelstrom results with a z-score > 3.5 in at least one of the cell types were selected for visualization in a heatmap.

Benchmark SCEPIA with scATAC-seq motif analysis

To establish how reproducible the SCEPIA runs are, and to benchmark the results by comparing with Maelstrom analysis, we split up the dataset into seven *randomly selected* sets of 100K cells. Each of these subsets was preprocessed, by selecting and filtering on highly variable genes, scaling, selecting neighborhood connectivities and performing PCA analysis, as described above for the full dataset. Each of these subsets were used in seperate SCEPIA runs, also run with the same settings as above. These same 7 scRNA-seq subsets underwent geometric sketching using geosketch (v1.2), prior to the SCEPIA analysis. First, the scaled data was used for PCA to enable the geometric sketching of 20K cells within the subset. On the raw data the highly variable genes were selected and after scaling the data, neighborhood connectivities selection, PCA and SCEPIA were performed in the same way as described before. In the same way, the full dataset was subsampled using geosketch to 20K cells on which SCEPIA was run again.
For each of the seven scRNA-seq subsets of the hHCA, the mean motif activities from the maelstrom analysis on cell type-averaged scATAC-seq, were correlated with the cell type-averaged expression levels of all of its potential binding transcription factors, as linked in the GimmeMotifs motif2factor

table. This comparison is referred to as the naive comparison. Only highly variable genes were selected for this analysis. For the SCEPIA comparison, all significant hits from the analysis (p-adj < 0.05) were selected and the predicted motif activities were averaged per cell type. These inferred cell type motif activities were correlated with the activities predicted by Maelstrom analysis for each cell type. The same was done for the SCEPIA results of the geosketched subsets. Correlations in all of these comparisons were performed using the Pearson correlation method.

6.6 Supplemental data

Table S6.1: **Regulatory potential and TPMs match robustly.** This table shows the effect of different weight curves for the calculation of regulation potential and their effect on the correlation coefficient and accuracy as a classifier. The table displays the correlation coefficient between identical cell types, and the percentages of samples where the correlation coefficient between TPM and Regulatory potential was the highest between identical cell types (specific), or belonging to the same ontology term (broad). The Ananse weight curve is the curve used by SCEPIA. The Wang weight curve is the curve proposed by the original study of regulatory potential. The promoter curve considers only the signal around the TSS with equal weight. The enhancer weight curve makes use of the curve of ANANSE, but does not count signal within 2kb of the TSS. Finally, we show the accuracy of a random classifier.

Weight curve	Correlation between	Correct specific	Correct broad
Ananse[29]	0.53 ± 0.14	64%	77%
Wang[352]	0.54 ± 0.15	64%	75%
Promoter (2kb)	0.54 ± 0.14	66%	77%
Enhancer	0.43 ± 0.14	60%	72%
Random	NaN	2%	3%

Table S6.2: **SCEPIA output for *MEF2*-family factors**. Correlation table output of SCEPIA run on the whole hHCA dataset, selected for the absolute highest correlating motif for each of the *MEF2* factors.

Factor	Motif	Correlation	Abs Correlation	P-value	P Adj	Motif Stddev
MEF2A	GM.5.0.MADS_box.0007	0.3169	0.3169	0.0076	0.0874	0.0119
MEF2C	GM.5.0.MADS_box.0019	0.2123	0.2123	0.0279	0.1446	0.0159
MEF2B	GM.5.0.MADS_box.0003	-0.0640	0.0640	0.1763	0.1446	0.0177
MEF2D	GM.5.0.MADS_box.0002	-0.0284	0.0284	0.3082	0.4337	0.0171

Table S6.3: **SCEPIA output for *MEF2*-family factors**. Correlation table output of SCEPIA + GeoSketch run on the whole hHCA dataset, selected for the absolute highest correlating motif for each of the *MEF2* factors.

Factor	Motif	Correlation	Abs Correlation	P-value	P Adj	Motif Stddev
MEF2C	GM.5.0.MADS_box.0019	0.1503	0.1503	0.0380	0.1612	0.0120
MEF2A	GM.5.0.MADS_box.0014	0.1501	0.1501	0.0381	0.1667	0.0144
MEF2D	GM.5.0.MADS_box.0014	0.1054	0.1054	0.0756	0.1024	0.0144
MEF2B	GM.5.0.MADS_box.0003	-0.0663	0.0663	0.1433	0.1591	0.0280

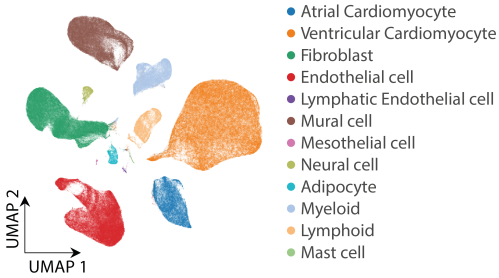


Figure S6.1: **Original annotation of cell types in the human Heart Cell Atlas (hHCA)** as provided by the authors[354]

.

Table S6.4: **The significance of improvements in the single-cell SCEPIA analyses per cell type.** Testing the significance of improvements between Pearson correlations as calculated for the baseline comparison and SCEPIA with Maelstrom, for each of the 7 scRNA-seq subsets and per cell type. The same was also done for the SCEPIA with Maelstrom correlations, with or without the GeoSketch run (right column). P-values were calculated with a one-sided Wilcoxon signed-rank test.

	Median Baseline	Median SCEPIA	Median SCEPIA + GeoSketch	Baseline vs SCEPIA	SCEPIA vs SCEPIA+ GeoSketch
Atrial Cardiomyocyte	0.08	0.27	0.55	0.0078	0.0078
Ventricular Cardiomyocyte	0.12	0.38	0.54	0.0078	0.0078
Fibroblast	-0.02	0.34	0.45	0.0078	0.1094
Endothelial cell	0.12	0.30	0.38	0.0078	0.0078
Lymphatic Endothelial cell	0.08	0.18	0.34	0.0547	0.0156
Mural cell	0.08	0.23	0.22	0.1484	0.5313
Mesothelial cell	0.15	0.14	0.11	0.5938	0.8516
Neural cell	0.13	0.37	0.31	0.0078	0.6563
Adipocyte	0.10	0.34	0.34	0.0156	0.0781
Myeloid	0.26	0.58	0.66	0.0078	0.0391
Lymphoid	0.00	0.39	0.55	0.0078	0.1094
Mast cell	0.08	0.45	0.52	0.0078	0.1094

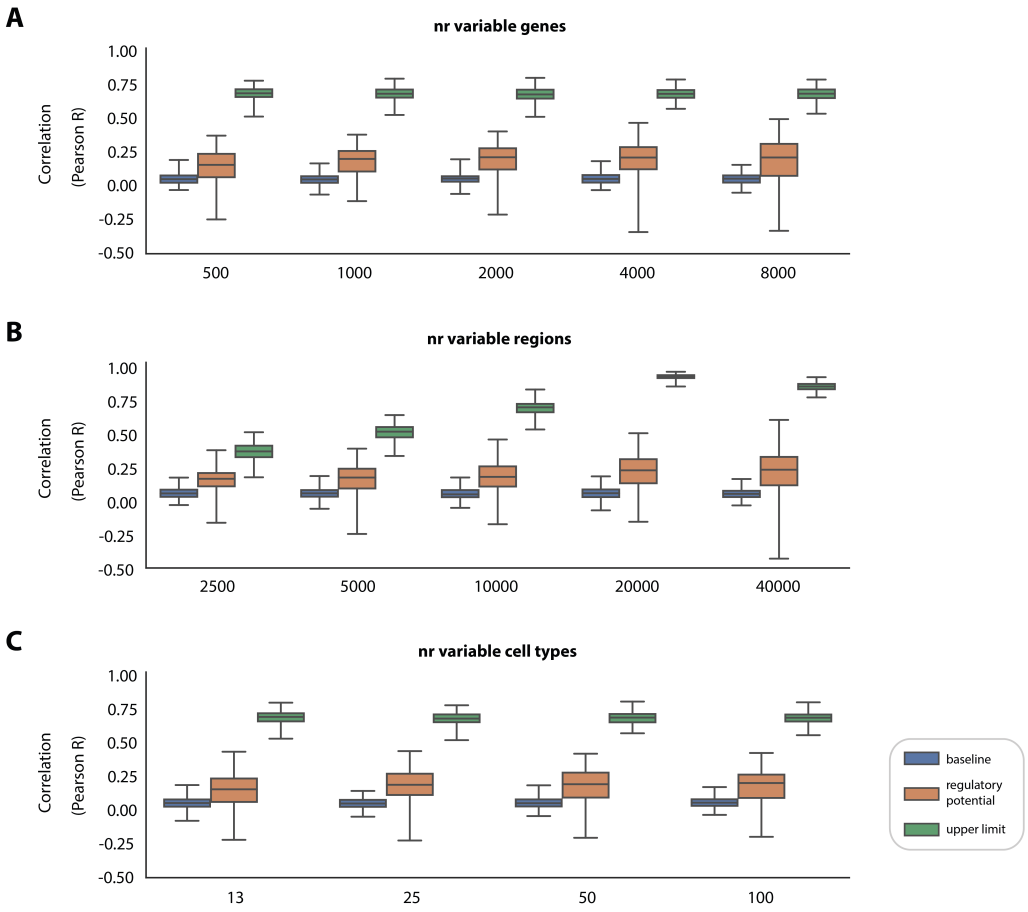


Figure S6.2: **One factor at a time benchmark of SCEPIA shows that the method is robust against a range of settings.** Each comparison has 100 permutations, and shows the baseline, the regulatory potential-based approach (SCEPIA), and the upper limit. The upper limit is calculated by taking the top-n differential enhancers and applying gimme maelstrom on those. The top-n is the same number of enhancers as the regulatory potential-based approach **(A)**, **(B)**, and **(C)** show the Pearson correlation coefficient when changing the number of variable genes, number of variable enhancers, and the number of cell types SCEPIA considers respectively.

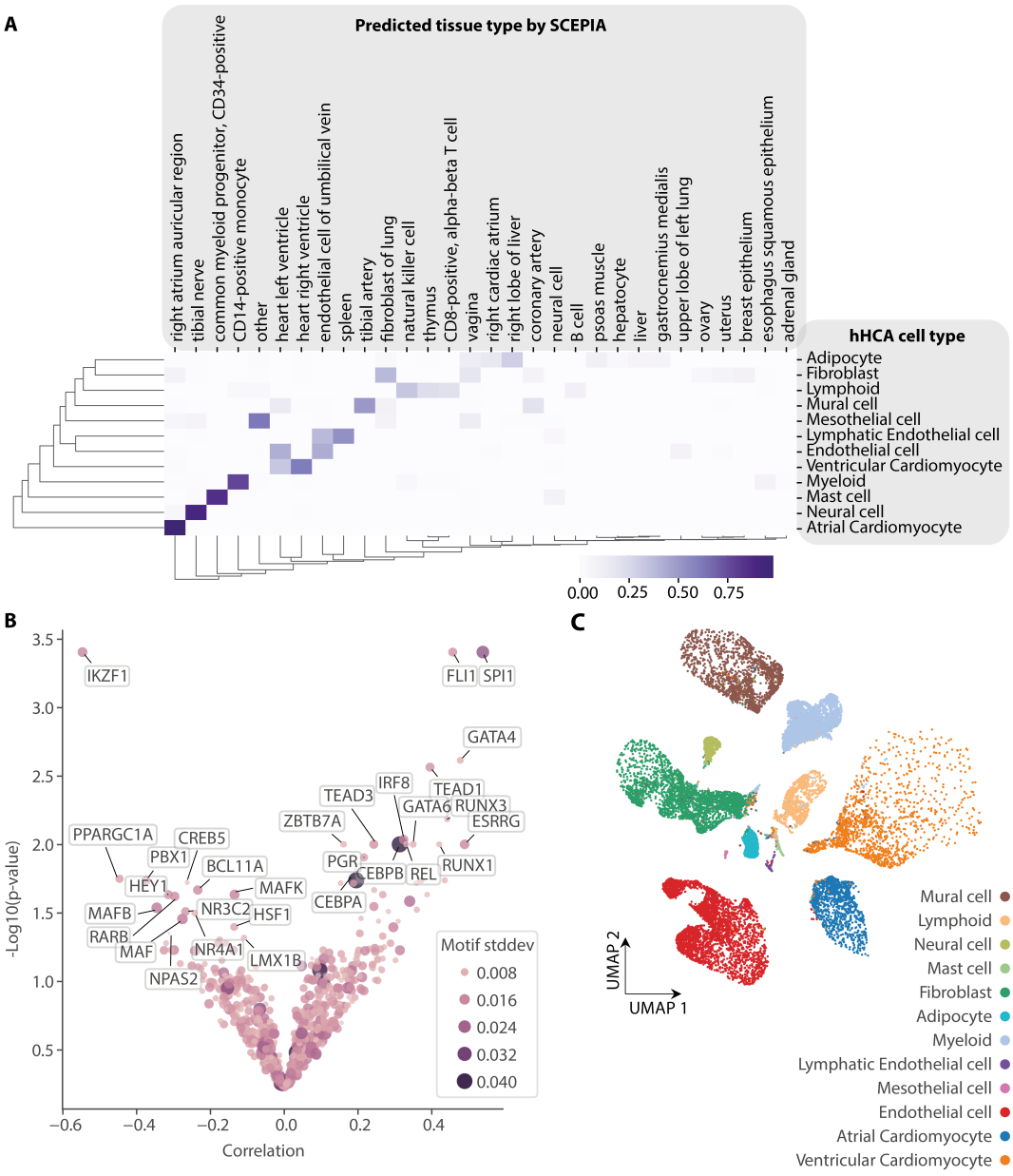


Figure S6.3: **Specifics and outcome of the SCEPIA + geosketch run on 20K cells.** (**A**) SCEPIA annotation of the geosketched cell clusters. Clusters in the columns as annotated by the hHCA, and cell type annotations in the rows as predicted by SCEPIA. The frequency is the number of cells per cluster annotated for each cell type. (**B**) Highest scoring hits inferred with SCEPIA + geosketch run. (**C**) UMAP and cluster annotation of the hHCA 20K cells geosketch.

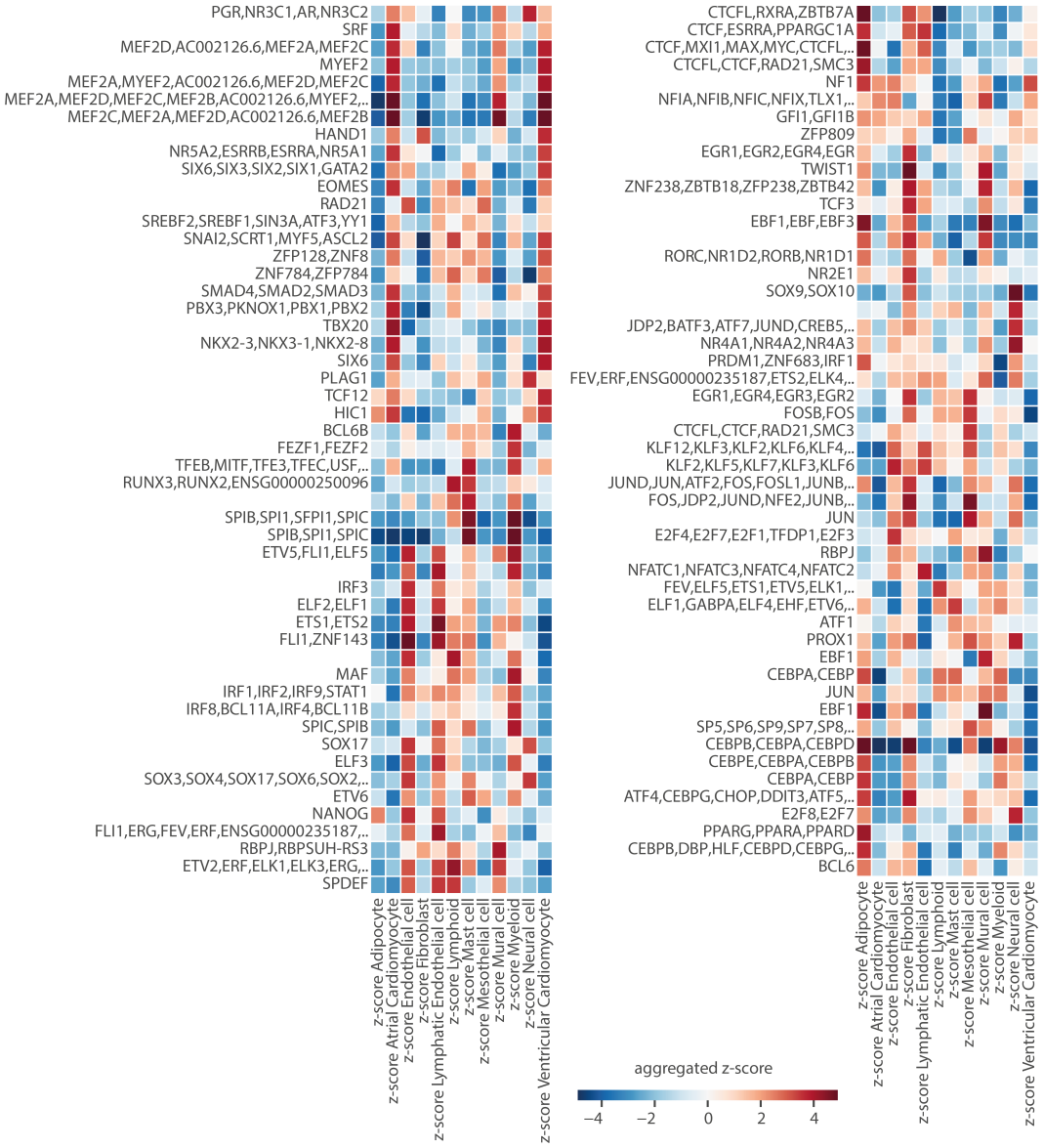


Figure S6.4: **Results of the Maelstrom motif analysis on the hHCA scATAC-seq.** Output of the Maelstrom motif analysis on peak intensities averaged per hHCA cluster. The Maelstrom hits were filtered on a z-score > 3.5 in at least one of the clusters.

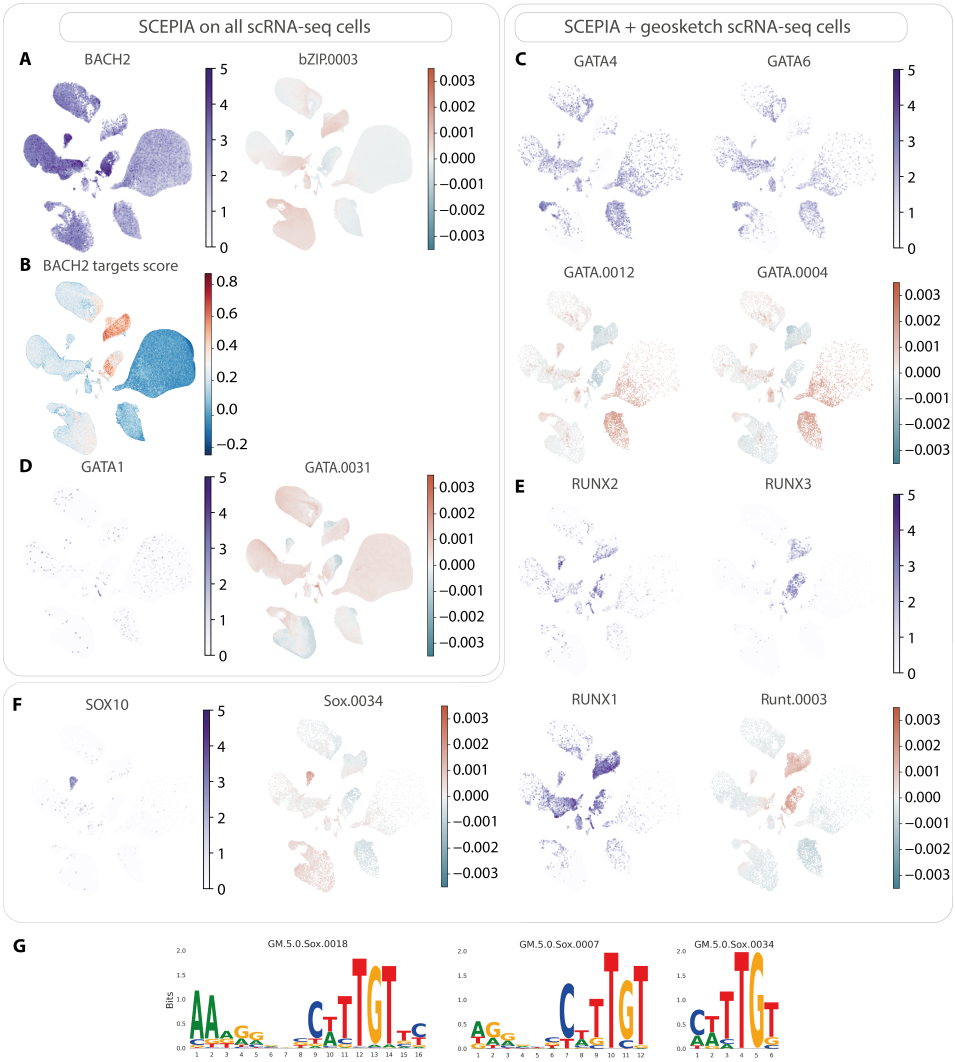


Figure S6.5: **Examples of results from the SCEPIA and geosketch + SCEPIA analysis.** (**A**) Expression levels of *BACH2* plotted across the cells of the UMAP (left) and *BACH2* motif (GM.5.0.bZIP.0003) activity as predicted by SCEPIA (right). (**B**) Compound expression levels as calculated with *gene_score* function in scanpy, of 57 *BACH2* target genes. Target genes taken from ChIP Atlas[393,394] results for *BACH2* ChIP-seq experiments in differring cell lines and with a "ChIP score" 100. (**C**) Expression levels of *GATA4* (upper left) and *GATA6* (upper right) plotted across the geosketched hHCA data and motif activity for the GM.5.0.GATA.0012 (bottom left) and GM.5.0.GATA.0004 (bottom right) motifs as predicted by SCEPIA for this dataset. (**D**) Expression levels of *GATA1* plotted across the cells of the UMAP (left) and *GATA1* motif (GM.5.0.GATA.0031) activity as predicted by SCEPIA (right). (**E**) Expression levels of *RUNX2* (upper left), *RUNX3* (upper right) and *RUNX1* (bottom left) across the cells of the geosketched dataset, and the GM.5.0.Runt.0003 motif activity as predicted by SCEPIA for this dataset. (**F**) Expression levels of *SOX10* plotted across the geosketched dataset (left) and *SOX10* motif (GM.5.0.SOX.0034) activity (right) as predicted by SCEPIA for this dataset.

Figure S6.6: **Examples of results from the SCEPIA and geosketch + SCEPIA analysis.** (**A**) Expression levels of *ESRRG* plotted across the cells of the hHCA UMAP (upper) and *ESRRG* motif (GM.5.0.Nuclear_receptor.0039) activity as predicted by SCEPIA (bottom). (**B**) Expression levels of *IKZF1* plotted across the cells of the hHCA UMAP (upper) and *IKZF1* motif (GM.5.0.C2H2_ZF.0239) activity as predicted by SCEPIA (bottom). (**C-D**) Same plots as in **A** and **B**, respectively, shown for the geosketched hHCA dataset and motifs linked in the analysis (in this case GM.5.0.Nuclear_receptor.0070 for *ESRRG*). (**E**) Expression levels of *MEF2*-family factors, with *MEF2A* (upper left), *MEF2B* (upper right), *MEF2C* (bottom left) and *MEF2D* (bottom right), across the cells of the hHCA UMAP.

Chapter 7

General Discussion

All cells in our body contain the same DNA, yet the cells in our liver express a completely different set of genes than the cells in our heart. Gene expression regulation is the mechanism that controls the activation and repression of genes in the genome. The spatial and temporal control of gene regulation is an essential part of (embryonic) development. Moreover, genes and their spatio-temporal expression patterns and regulation are remarkably conserved between species.
In this thesis, I studied gene expression regulation from an evolutionary-developmental point of view. I first discussed three recommendations for gene regulatory network inference in development; multi-omics data, single-cell sequencing, and artificial neural nets (**Chapter 2**). Next, I presented seq2science, a preprocessing workflow for functional genomics analysis (**Chapter 3**). This chapter is followed by a description of how the definition of the phylotypic stage is ambiguous and how these analyses can benefit from statistical controls. Applying these new controls to previous studies exposes flaws in the methodology and the subsequent interpretation of their results (**Chapters 4 and 5**). Finally, in **Chapter 6**, I presented a novel method to infer transcription factor activity based on single-cell transcriptomic data. In this chapter, I present my final concluding remarks.

7.1 Inferring gene regulatory networks

Gene regulatory network inference is the computational and statistical approach to decipher the regulatory interactions between genes. Understanding how certain gene products, for instance, transcription factors, regulate the expression of other genes helps with gaining a fundamental understanding of cellular processes. It has become a standard analysis step in most (single-cell) studies, yet most inferred networks perform barely any better than randomly generated networks[183,209,210]. In **Chapter 2** I discuss three (potential) improvements over "traditional" gene regulatory network inference; the inclusion of multi-omics data, single-cell sequencing, and artificial neural networks.
As a first recommendation, I suggest the inclusion of multiple sequencing modalities. Historically, gene regulatory networks have mainly been inferred based on transcriptomic data alone. Apart from the issue that these networks perform poorly[183,209,210], there is the fundamental issue of using transcripts both as input and as output in these models. Transcript abundance is the result of transcription regulation and degradation. Transcripts, however, are typically not responsible for signal transduction, chromatin remodeling, translation, and transcription regulation and degradation. The protein product synthesized from these transcripts, however, is responsible for these types of regulation and is assumed to correlate closely with transcript abundance. The correlation coefficient between transcript and protein, however, ranges between 0.3-0.4[349,350], with transcript abundance thus approximately explaining only between 0.09-0.16 of the variance of protein abundance. Similarly, in **Chapter 6** I show a poor correlation between transcription factor expression level and its motif activity. High-throughput proteomics is more costly than sequencing data and lacks its accuracy, with single-cell proteomics lagging behind single-cell transcriptomic and (epi)genomic techniques[396]. More accessible and accurate metrics for gene regulation are epigenomic assays that measure genome accessibility and histone modifications[29,30,188], which in turn improve the performance of these models. Ideally, future GRN inference methods include (epi)genomic, transcriptomic, and proteomic information, and using only transcriptomic data was perhaps in hindsight a naive endeavor.
The second recommendation I suggest is the use of single-cell data. Single-cell sequencing is important because it allows for the separation of the different signals present in tissues. Moreover, by separating cells in samples it is possible to order them on a fine-grained (pseudo)time[397], allowing for new ways to study this problem and increasing the detail of these analyses. The analysis of single-cell data, however, has turned out to be problematic with major issues of batch effects[398–400]

and sparsity[400,401]. Furthermore, gene regulatory networks inference methods specifically designed for single-cell transcriptomic data are actually not more accurate than their "bulk" counterparts[209]. So whilst single-cell data has major potential to improve GRN inference, so far the provided improvements to gene regulatory network inference have been marginal in practice.
The final recommendation I make is the use of artificial neural networks, which is a recommendation I would like to redefine here. One of the reasons I recommended the use of ANNs is because they can incorporate and train on data over multiple conditions. The networks that come out of these models can thus be based on a wide range of conditions, and are known as foundation models. As long as enough data is used we can expect these foundation-model-based networks to be able to extrapolate to new unseen conditions[402]. This makes these types of networks powerful methods for hypothesis generation, especially compared to the usual approach of inferring gene regulatory networks based on the difference between two conditions. However, ANNs are not the only type of method that can produce foundation models. For instance, one can develop (mechanistic) gene networks based on ordinary differential equations that model changes in gene expression over time[208]. These types of "universal" networks make conceptually the most sense, as ultimately, a single set of instructions (DNA) is shared across all cells of an individual. As such I would redefine the last recommendation to be to design foundation gene networks, with the caveat that current foundation models fail to outperform their simpler counterparts[403].
In **Chapter 6** I introduce the computational tool SCEPIA, which infers differentially active transcription factors in single-cell transcriptomic data. It matches transcriptomic data to a reference H3K27ac database. In this manner, the transcriptomic data can be used to find differential transcription factors and link them to the differential motifs in the reference database. Even though the basic assumption of SCEPIA, the linking of epigenomic data to transcriptomic data by regulatory potential[352], is relatively simple, it seems to perform reasonably well. Nonetheless, the approach of SCEPIA contains two major downsides. First, it produces combinations of differential transcription factors and motifs, but generates no potential gene regulatory network. This means that it is difficult to infer cause-and-effect relationships based on the output of SCEPIA. Second, the implementation is based on numerous assumptions and heuristics. We have limited understanding which of these heuristics are beneficial or detrimental to its performance, and as such we have not gained a new fundamental understanding of which types of interactions are important or how they interact. Moreover, due to the lack of a benchmark against the alternatives to SCEPIA[188,404], it is unclear how this approach compares to alternative methods. Ideally, any new gene regulatory network inference method performs an extensive parameter sweep over all design choices as in[405]. Nevertheless, this might not always be feasible, because of computational constraints and the lack of a gold standard.
Finally, the gene regulatory network field lacks a formal language by which to describe gene interactions[406]. Studies presenting their new methodology usually report a complex graph of nodes (genes) and vertices (their interactions)[28,29,177,187,188,191,192,204,220]. It is not uncommon for studies to report figures with dozens of genes connected by hundreds of crossing lines. Moreover, what these lines, and their color and thickness, mean is decided on a case-by-case basis. Understanding and remembering these networks poses a challenge. It has often been suggested to outsource this complexity to computational biologists[406–408], and whilst I agree with the general disposition in these articles that all biologists should learn to program, I think a better solution should be possible. During my doctoral studies, I have always looked with envy at the networks that the fields of metabolomics and electrical engineering produce. These fields have managed to condense extremely complex interactions into organized and even visually pleasing networks. I believe the difference lies in their ability to formally simplify their interactions and subnetworks. For instance, in major metabolomic pathway visualizations, not all the interactions are shown, but for example, a circle is drawn to represent the citric acid cycle. When one is specifically interested in the citric

acid cycle one can then look this cycle up in another diagram. Similarly, all standard components of electrical networks, such as batteries, capacitors, and resistors, are defined by the International Electrotechnical Commission[409]. As far as I am aware, the only systematic approach to GRN visualization is BioTapestry[410], but this visualization tool has not gained a substantial foothold in the field. To gain a comprehensive understanding of gene networks, the field has to adopt the practices of modular subnetworks and standardized symbols. Without these changes, the interpretation and analysis of gene networks will remain limited to computers and computational biologists exclusively.

7.2 The scientific dogma of the phylotypic stage

> “All models are wrong, but some are useful.”
>
> *George Box*

While historically the phylotypic stage has predominantly been examined and described through qualitative methods, the 21st century started a paradigm shift towards a more quantitative and data-driven approach to understanding this phenomenon[319]. The first notable quantitative investigation into the phylotypic stage was done by Bininda-Emonds *et al.*, where they calculated temporal conservation as the order in which morphological embryonic features appear in vertebrates[318]. However, it wasn't until the early 2010s that the field truly embraced quantitative methodologies with the simultaneous publication of two groundbreaking studies in Nature[60,62]. In these works, Domazet-Lošo *et al.* investigated the average developmental age of transcripts in *D. rerio* and *D. melanogaster*, whilst at the same time, Kalinka *et al.* explored the temporal transcriptome similarities across different *Drosophila* species. These molecular studies opened a new line of research to the quantitative basis of the phylotypic stage, where support for the hourglass model appeared stronger and stronger with each new study.

The Transcriptome Age Index (TAI), as introduced by Domazet-Lošo *et al.*, is a metric of the average evolutionary age of transcripts over time[60]. In this study, evolutionary age is determined as the count of taxonomic branches that can be traced back to a gene. The central idea of the TAI is that temporal changes in the average transcript age provide insights into the degree of conservation during development. Domazet-Lošo *et al.* found that both zebrafish and *Drosophila* expressed the oldest transcriptome at their respective phylotypic stages and concluded that an old transcriptome marks the phylotypic phase. However, an independent re-analysis conducted by Piasecka *et al.* raised some critical points about the methodology[63]. Their investigation revealed that the TAI is heavily influenced by a relatively small subset of genes due to major differences in transcript levels per gene (transcriptomic data is notoriously heteroscedastic[411]). Log transforming the data, which is a standard processing step for this type of data, completely invalidates the results of the original study. Nonetheless, the original study introducing the TAI has been cited 88 times between 2010 and 2013, but has been cited 359 times since Piasecka *et al.*'s publication (covering the years 2014-2023). As it turns out, you can now analyze your data with and without transformation, and keep the results that reinforce your preferred evo-devo hypothesis. A notable example is Wu *et al.*'s study on Spiralian development[412]. In their analysis of untransformed data for *Crassostrea gigas*, *Haliotis discus hannai*, and *Perinereis aibuhitensis*, they claim to have found an inverse hourglass pattern of evolutionary conservation and speculate on why spiralia have a different temporal selection pressure than other species. However, their supplementary data reveals a different pattern for *Crassostrea gigas* after square root transformation, shifting from an inverse hourglass to a funnel shape. The transformed TAI of the other two species is not shown. Moreover, the TAI of the phylotypic stages of *Haliotis discus hannai* and *Perinereis aibuhitensis* both lie within one standard

deviation of the TAIs of pre-phylotypic stages, which means that it cannot be excluded that these patterns are caused by stochastic fluctuations

The work of Barbara Piasecka, where she showed that the pattern of the TAI is caused by a subset of all genes, was led by Marc Robinson-Rechavi. The main work of this study was not about the TAI, but about using a multitude of metrics to estimate temporal evolutionary conservation. They conclude that different metrics produce different results. In his personal blog, Marc Robinson-Rechavi even states "First, that biology is complicated, and insisting on answers such as « the hourglass exists (and explains diverse data) » or « it doesn't » may not be the best strategy. Second, that the technical details are very important"[413]. In a later study led by Marc Robinson-Rechavi, they suggest that the *Drosophila* phylotypic stage is the most conserved stage based on the enrichment of regulatory DNA. Moreover, they claim that the high conservation at the phylotypic stage is caused by the lack of positive selection compared to other stages. In **Chapter 4** I show that the technical details are indeed important, and demonstrate that the high conservation at the phylotypic stage is a statistical artifact, and express my concerns about their methodology of studying positive selection.

In 2003, Bininda-Emonds *et al.* conducted a study of the phylotypic stage, which was revisited seventeen years later by Gerardo A. Cordero *et al*[317,318]. Both studies are about the quantitative temporal analysis of morphological characteristics. To the best of my knowledge, these are the only quantitative analyses of morphological characteristics concerning the phylotypic stage. The initial findings of Bininda-Emonds *et al.* revealed an unexpected inverse hourglass pattern in morphological rankings. This discovery challenges the existing assumption of mid-development being the most conserved period. It took seventeen years for a follow-up study, that surprisingly showed precisely the opposite - an hourglass pattern. Unfortunately, Cordero *et al.* only comment that the difference between these two studies *could* be caused by a difference in morphological characteristics, methodology, and species used, without any further analysis of the differences. They were correct in their assessment though, as in **Chapter 5** I show through simulated data that the inverse hourglass pattern of Bininda-Emonds *et al.* is likely caused by the methodology's sensitivity to edge effects, and does not represent a biological signal.

In 2016, Levin *et al.* introduced the transcriptomic inverse hourglass model as a potential method for distinguishing between different phyla[56]. This study has been criticized by Casey Dunn and Andreas Hejnol for its lack of a within-phylum control[322] and incorrect statistical methods[320]. Addressing these concerns seems imperative, given the possibility of a "universal" phylotypic stage. The only other between-phyla evo-devo study I am aware of is between deuterostomes with the chordate amphioxus[314]. This study finds an hourglass-like pattern, directly contradicting the work of Levin *et al.*, but do not comment on this. In **Chapter 4**, I present evidence that the transcriptomic inverse hourglass is a statistical artifact resulting from standardization rather than an accurate representation of temporal conservation.

Yoav Mayshar *et al.* studied the phylotypic stage and the hourglass model from a single-cell point of view[58]. Their research involved a comparative analysis of cell type proportions during the development of rabbit and mouse embryos. However, in **Chapter 4**, I show that both the rabbit time series as well as the mouse time series exhibit discontinuous temporal patterns. These discontinuities influence the temporal correlations between the two species. Without a better distributed temporal sampling, a direct comparison with this data concerning temporal conservation is difficult. Moreover, a mid-developmental transition between pairwise comparisons of two time series visually display an hourglass. Thus confusingly, if the data looks like an hourglass in a pairwise heatmap, this represents the inverse hourglass model of conservation (Fig. 4.1). This means that while the observed pattern by Mayshar *et al.*, a "mid-developmental transition", is described as confirming the traditional hourglass model, it actually shows exactly the opposite.

These examples exemplify that the evidence for the existence of a phylotypic stage is not as strong

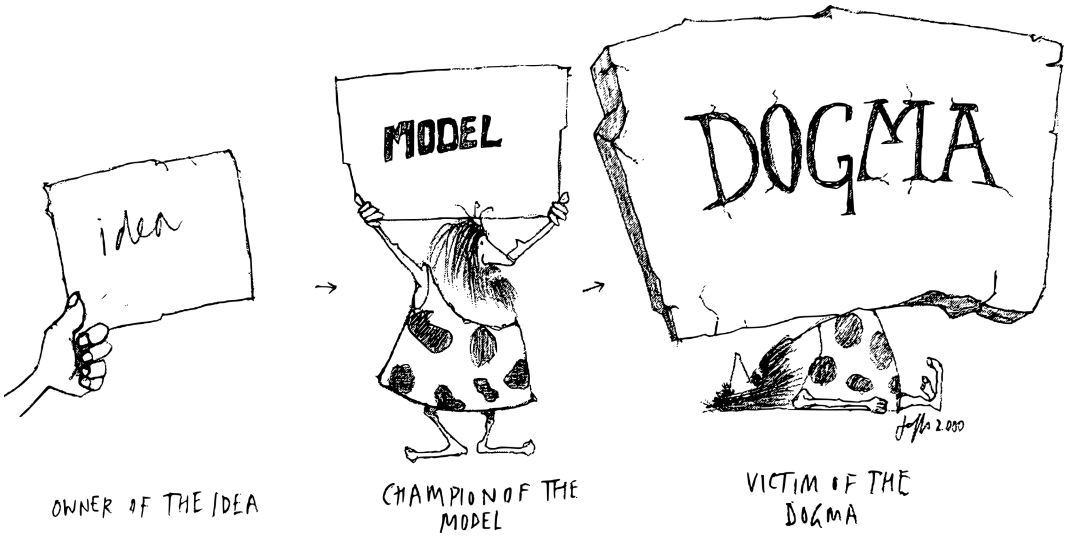


Figure S7.1: **The phylotypic stage as a scientific dogma?**[414]

as often presented. On top of that, in **Chapters 4 and 5**, I discuss the methodological issues in comparative analyses concerning the phylotypic stage. The root problem is that the phylotypic stage and its related models are ill-defined, and as a consequence, there is a lack of appropriate (statistical) controls in these studies. Is temporal conservation in these models present in comparisons within and between phyla? And is temporal conservation already observable when comparing a species against itself? By systematically addressing these ambiguities in previous studies through within-species, between-species, and between-phyla comparisons I show examples where the conclusions are not supported by the data. To summarize my main findings:

- The transcriptomic hourglass-like pattern between zebrafish and frogs[57] can be explained by within-species correlations alone.
- The transcriptomic between-phyla inverse hourglass pattern[56] is a statistical artifact and can be reproduced by simulated data with no specific temporal conservation.
- The pattern of cell type proportion similarity between rabbits and mice[58] can be explained by discontinuous temporal sampling.
- The *Drosophila* enhancer conservation re-analysis results in a "mid-developmental" stage of maximum similarity, albeit at a different point than found by the original authors. Moreover, in **Chapter 4** I highlight additional problems with the methodology.
- The morphological within-phylum inverse hourglass pattern is caused by edge effects and can be reproduced by simulated data with no specific temporal conservation.

The null model for evolutionary embryonic development would be that there is no specific stage of higher or lower temporal conservation. Altogether, I have found little evidence to reject the null hypothesis of constant temporal conservation based on quantitative data, and thus see no reason to support a (molecular) phylotypic stage.
Moreover, in **Chapter 4** I discuss further ambiguities in the models of the phylotypic stage. If a molecular phylotypic stage would exist, which genetic features are expected to be conserved and which are not? The original observation that vertebrate embryos, perhaps, look more like each

other *externally* at certain points in development, says little about the whole-embryo molecular basis for it. While the morphological observations of Haeckel are almost 200 years old, there have been only two quantitative studies about the morphological phylotypic stage, and these two studies were in direct contradiction with each other. Instead, there appears to be a pursuit to find new (quantitative) methodologies that produce a more straight-forward confirmation of the existence of the phylotypic stage, such as embryonic lethality[311], DNA sequence conservation[59,63,307] and activation order[64], cell type proportion[58], and whole-transcriptome similarity[57,61–63,312,314,315], with little effort to integrate these results across studies. Of these methodologies, the whole-embryo transcriptome has become the most popular method to asses quantitative similarity. The implicit justification for this popularity is that transcriptomics is an unbiased way to study genes during development. Yet, a select group of genes, transcription factors, play a pivotal role the regulation of transcriptomic levels. Coincidentally, the genes thought to be responsible for the phylotypic stage are the HOX transcription factors[55], creating a major dependence in the data. Moreover, the HOX genes are mainly responsible for the spatial organization of an organism, but no spatial information is measured during whole-embryo sequencing. It remains unclear whether the current whole-embryo whole-transcriptome methodology is able to measure a relevant evolutionary pattern.
Throughout scientific history, certain theories, such as taxonomic phyla and the phylotypic stage, have evolved from initial concepts into widely accepted truths, creating a demand for a molecular explanation along the way. However, a fundamental issue arises from the loose and ambiguous definitions on which these theories are based, leading to their lack of predictive power and falsifiability, rendering them, by Popperian standards, non-scientific in nature. For instance, the concept of phyla hinges on the notion that animals sharing a common basic body plan are part of the same phylum, yet paradoxically, the basic body plan is defined as the morphological characteristics shared by all animals within the same phyla[415,416]. The definition of the phylotypic stage is similarly ambiguous[417]. Historically, the pharyngula stage[418], early somite embryo[419], and the tail bud stage[420] have all been proposed as the vertebrate phylotypic stage. Conversely, in quantitative studies, the choice of definition in turn depends on which stage exhibits the highest quantitative conservation. Consequently, the pharyngula[57,61,63], the early somite embryo[60], or simply the stage(s) with the highest conservation metric[58,62,317] have all been identified as phylotypic stages. In conclusion, our current approach to studying the phylotypic stage, where we selectively include definitions and ignore non-conforming results, is not only wrong, but also not useful.

7.3 The technical debt of bioinformatics

The complexity of molecular biological systems naturally makes their (bioinformatic) analysis a complex task as well. Their analysis is further complicated by the continuous addition of new insights, which leads to updated methodologies and analytical approaches, keeping the field of bioinformatics in a state of perpetual change. Often, in the rush to generate new insights, tools are developed quickly, leading to solutions that, while functional, may not fully capture the nuances of the underlying biology. In turn, the bioinformatics community builds upon these tools, perpetuating their inefficiencies[421,422]. Eventually, the burden of these suboptimal solutions becomes untenable, which requires major revisions of the original methodology. In software development, this concept is known as technical debt. Technical debt, similar to financial debt, means that shortcuts taken today must be paid back in the future, often at a greater cost. I believe that in the field of bioinformatics, the technical debt has become so large that it has become prohibitive for most analyses.
A significant source of technical debt exists in the data structures bioinformatics uses. During my doctoral studies, I have extensively made use of the FASTA format, Position Weight Matrices (PWMs), and positional genomic formats such as the GFF and GTF. The General Feature Format

(GFF), and its derived Gene Transfer Format (GTF) are formats that are not fit for their use anymore. The GFF and GTF format specify genomic features and their locations, where each line represents a feature and the columns are different attributes. One of these columns indicates the position position of the feature in a one-indexed coordinate system. This differs from the widely held view, albeit slightly subjective, that zero-indexing is the norm for coordinate systems[423]. This makes one-off errors between these formats and other positional formats, such as BED, easy to make. Moreover, as our understanding of genes got more complex, so did their subsequent analysis. This meant that it became necessary to express more complex relationships between features within these GFF and GTF files. Flat formats such as GTF and GFF can not naturally represent this, yet the newer GFF and GTF formats now specify these relationships in a "attribute-value" column. This has made these formats less suitable to be parsed line-by-line, something for which they were originally designed. Even the first sentence of the GFF3 format starts disagreeing with its own design; "although there are many richer ways of representing genomic features ..., the stubborn persistence ... of ad-hoc tab-delimited flat file formats declares the bioinformatics community's need for a simple format that can be modified with a text editor and processed with shell tools like grep"[424]. Moreover, the unstructured manner in which these new formats were released resulted in multiple incompatible dialects. The frustration of the field about the GFF and GTF formats is best illustrated by a Tweet from Lior Pachter, where he asked if there is a citation for the GTF format. The most popular response covered the collective frustration of the field, stating "Ain't no one gonna take responsibility for that mess"[425]. A transition to a relational data format, capable of adequately handling the increasing complexity and relationships in genomic data, seems not only logical but inevitable.

FASTA is a text-based format that represents nucleotide sequences or amino acid sequences, in which nucleotides are represented using a single-letter[426]. Single nucleotide polymorphisms are defined by their IUPAC codes, but uncertainties longer than a single nucleotide (haplotypes) can not be naturally represented. This problem is solved by appending alternative sequences to the file. These sequences can then be selectively used in the downstream analysis. However, bwa mem(2)[271,272] and Illumina Dragen are currently the only genomic aligners that support this natively. Consequently, despite alternative regions being part of the human genome since 2013, structural variation is typically overlooked in studies. Moreover, for some researchers the difference between a primary assembly (no alternative sequences) and a toplevel assembly (includes alternative sequences) might not be clear, causing them to align their reads against a toplevel assembly. This results in many multimapped reads, and consequently a low coverage in these regions. Thus, the limitations of FASTA in representing haplotypes highlight the necessity for adopting more sophisticated file formats like graph genomes[427], which represent haplotypes more effectively.

Finally, PWMs quantify the log-likelihood of each nucleotide's presence in a motif. While PWMs are straightforward to understand and visually interpret, they oversimplify transcription factor binding by assuming that nucleotides within a binding site act independently[428–430]. Higher-order structures, such as DNA methylation and DNA shape, can not be naturally represented in PWMs. As a consequence, PWM databases consist of numerous nearly identical PWMs, with one PWM for each dependence between nucleotides. A notable example of this is the SCENIC+ motif database[431], which claims to have compiled the largest collection of PWMs to date. Their database contains over 30,000 "unique" PWMs. However, these PWMs are distributed among the approximately 3,000 DNA binding domains of TFs between humans, mice, and fruit flies, demonstrating a considerable amount of redundancy. Alternatives, such as graph-based PWMs[432] and neural networks[433,434], do incorporate dependence between nucleotides and as such offer a more complete understanding of TF binding.

Whilst the outdated data formats represent a more systematic and difficult-to-solve technical debt, there is also a considerable amount of low-effort technical debt. This includes the use of old assem-

blies, unresolved bugs, and unoptimized software. For example, in 2023, Google Scholar identified 7,630 publications that used the older hg19 or GRCh37 genome assemblies (released in 2009), while there were only 8,960 publications that used the newer hg38 or GRCh38 assemblies (released in 2013). Notably, the T2T-CHM13 assembly (released in 2022) has only been mentioned in 222 publications in 2023. Whilst I recognize that using an older genome version might be preferable in comparison to previously analyzed data or databases, not updating your assembly perpetuates the use of outdated versions. Not surprisingly, a newer genome assembly results in more accurate variant calls and discoveries[435–437].

Another easily addressed problem is resolving identified bugs in popular software. For example, MACS2 is one of the most popular tools in the field of regulatory genomics and is used to discover enriched regions in the genome. MACS2 has been cited more than 14,000 times[37]. Despite its popularity, it contains some unexpected default settings. The recommended way to use MACS2 in the case of ATAC-seq discards all mates from paired-end data, effectively removing half of the cleavage information[438]. Whilst this can be overcome by some specific processing on the user side, most users are not aware of this setting, resulting in suboptimal peak sets in presumably the majority of ATAC-seq studies that use MACS2. Similarly, a popular toolkit for the analysis of single-cell data, Seurat, which has been cited over 8,000 times, contains a particularly painful bug. Its implementation of the log fold change calculates the log of the means (as opposed to the mean of the logs), which makes it sensitive to outliers. This results in unexpected log fold changes and vastly different outcomes depending on whether one used Seurat or any alternative tool such as Scanpy for their analysis. The bug was reported in November 2022, got an incomplete fix in March 2023, and as yet, January 2024, has not been fixed[439].

Finally, most software is usually poorly optimized, especially in the field of single-cell analyses, taking days to run and requiring enormous amounts of memory[210]. The usual solution in the field to this problem is, surprisingly, not to improve the software implementation, but instead to increase the hardware capabilities. This practice not only causes enormous computational waste but also creates a substantial barrier to researchers from labs without such resources.

A personal example of how technical debt directly affected my doctoral studies is through the Sequence Read Archive (SRA). The SRA is a public repository that serves as a centralized archive for storing and sharing high-throughput sequencing data. The SRA has been essential for my doctoral studies, as I have relied solely on public data. For **Chapter 6**, I aimed to download all available human H3K27ac data (+/- 12,000 samples) from the SRA as a reference database. Processing these samples meant that I had to download 20TB from the SRA and spent approximately 300,000 CPU hours processing them on Cartesius, costing an enormous amount of computational resources and emitting an estimated one-and-a-half tonnes of CO_2[†]. In the analysis that followed, however, it turned out that obtaining sample annotations, such as tissue or cell type, is challenging due to the lack of standardized metadata on the SRA. Third-party tools like MetaSRA[442], and PredictMEE[443] have been developed to automatically infer this metadata but ultimately failed to provide accurate results. In the end, I decided to discard the experiment, wasting an enormous amount of resources. Whilst the wasted computational resources could have been avoided by a pilot study on my side, an unstructured SRA is an enormous wasted opportunity. My experience with the SRA is exemplary of a broader trend in big data bioinformatics research, where ENCODE is preferred over the SRA despite its smaller size, in part because it offers structured metadata.

Not only is the SRA lacking crucial metadata, but the associated software sra-tools, a toolkit for interacting programmatically with the SRA, is particularly hard to use. As a consequence, many alternative tools have been developed for downloading data from the SRA, such as the SRA-explorer[444],

[†]300,000 CPU hours, spread over the 8 cores of an Intel Xeon Processor E5-2670 uses 115W[440], which results in a total estimated energy usage of 4312 kWh. The production of one kWh emitted an average of 369 grams of CO_2 in the Netherlands in 2019[441], which in turn leads to a total CO_2 production of 1591 kg. Furthermore, this calculation ignores the significant electricity usage of data transfer and working memory.

pysradb[266], FetchFastQ[445], nf-core/fetchngs[446], parallel-fastq-dump[447], and finally our download-fastq workflow of seq2science[337] (**Chapter 3**). The poor implementation of the sra-tools results in additional and unnecessary work, wasting the time and resources of third-party researchers. Similarly, the submission of new sequencing data is notoriously complicated which has even led others to develop tools to streamline the upload process[448].

I consider all of the aforementioned examples, like inadequate data formats, outdated assemblies, unresolved bugs, and the SRA, as technical debt. Addressing these issues is crucial for advancing the field; it's not a matter of whether they should be addressed, but rather how and when. Even though addressing some of these problems will require substantial effort, postponing their resolution will not simplify the task. Moreover, delaying the resolution hinders all current and future analyses with a less efficient process. In my opinion, there are three main reasons for the widespread technical debt in bioinformatics; (i) insufficient incentives to maintain and develop high-quality software, (ii) the lack of interdepartmental software collaborations, and (iii) the lack of formal bioinformatics training.

It appears that writing high-quality software is not incentivized, although frankly, I have little personal experience with this part of science. Typically, the end product of an analysis is a scientific publication, and in the rush to generate new insights, software, and tools emerge as byproducts. Consequently, their design rarely prioritizes adaptability or user-friendliness for broader applications[449]. This problem is reinforced by the fact that funding decisions are based predominantly on the impact of previous research and the novelty of the newly proposed work. Prominent grant opportunities like the NWO Veni, Vidi, and Vici grants, EMBO postdoctoral fellowships, and Marie Skłodowska-Curie Postdoctoral Fellowships all emphasize the importance of innovative applications. This emphasis on innovation creates a strong selection pressure for researchers doing novel work, inadvertently disadvantaging researchers dedicated to maintaining and enhancing existing software. While maintaining one's own software can potentially lead to increased citations and, by extension, improved funding prospects, this is not the case for the maintenance of crucial third-party software. The current funding system undervalues the maintenance of scientific software in bioinformatics, which by extension perpetuates its technical debt.

A different cause for technical debt is the fact that most tools are developed independently, most of the time by a single doctoral student or postdoc. This has two major downsides. The first is that the maintenance of tools depends on individuals. If someone decides to leave academia, their tools, how useful they might be, will soon become obsolete as no one will maintain them. A clear example is the dissolution of the "van Heeringen" group, where soon tools such as genomepy[267], gimmemotifs[139], ANANSE[29], SCEPIA (**Chapter 6**), seq2science[337] (**Chapter 3**), and qnorm[342] will become outdated as they will no longer be maintained. Second, as most tools are a byproduct of an analysis, they are often highly specialized for specific tasks. This, in turn, leads to many near-identical tools that solve similar problems, for example, all the different SRA FASTQ downloading tools. This is in stark contrast to the formulation of the SAM/BAM format[450] and the subsequent implementation of HTSLIB[451]. The SAM/BAM formats are the *de facto* standard for genomic alignments in the field, with the HTSLIB as the library to work with these files. This project was led by the 1000 Genomes project, which means that numerous bioinformaticians were and still are involved in their design. This guarantees that the SAM/BAM format is useful for a wide range of different applications. Moreover, an important detail is that the developers of the SAM/BAM format and HTSLIB are all researchers and use their own implementations. This guarantees that the developers continuously test their own implementations for ease of use and correctness. Even though HTSLIB was started over a decade ago, it is still receiving regular updates.

Finally, as the field of bioinformatics continues to grow, the demand for skilled bioinformaticians is outpacing the current supply. This dynamic has led to an inclusive approach in the field, where biologists from various backgrounds are engaging in bioinformatics analyses. This inclusivity is ex-

emplified by projects like Galaxy[250], which have been instrumental in making bioinformatics accessible. However, whilst this inclusivity is beneficial, we should not neglect the need for specialized skills in certain areas of bioinformatics. For instance, the development of tools and file formats, as well as the execution of complex analyses, often require a deeper understanding of programming, statistics, and modeling. A telling example of the lack of formal training is the widespread use of Excel in bioinformatics analysis. Excel notoriously used to mangle gene names[452], for example, the gene alias *SEPT1* was automatically converted into September 1. Consequently, approximately 30% of studies report these mangled gene aliases in their supplementary data[453]! In 2020 the HGNC opted to change 27 gene names to avoid Excel name mangling, and the *SEPT1* alias became *SEPTIN1*. Thus, while inclusivity solves the demand for bioinformaticians in the short term, the intricacies of biology and bioinformatics do also demand more formally trained bioinformaticians.
In conclusion, addressing the technical debt in bioinformatics is crucial for the advancement of the field. This requires a collective effort to update and refine data representation models, resolve software bugs, and optimize existing tools. Additionally, a fundamental shift in the scientific culture of biology is needed, where the development and maintenance of software are given the same importance as novel research. This shift should include incentivizing high-quality software development, fostering interdepartmental collaborations, and emphasizing the need for formal training in bioinformatics. Only through these measures can we ensure that bioinformatics research remains reliable, accurate, and efficient.

7.4 Concluding Remarks

Transcription is a highly complex process, which depends on numerous factors such as chromatin context and transcription factor gene expression. The regulation of transcription is dynamic and changes throughout (embryonic) development, and is prone to changes through evolution. To study this phenomenon, experimental biologists have generated enormous amounts of data, which is typically analyzed by bioinformaticians and computational biologists. In this thesis I propose three improvements to gene regulatory network inference; the use of multiple omics, the use of single-cell data over bulk sequencing, and the use of foundation models. Additionally, I introduce two computational tools that aid in the investigation of transcription regulation. Seq2science is an end-to-end functional genomics preprocessing tool, and SCEPIA is a tool that facilitates the linking of single-cell transcriptomic information to an extensive database of chromatin context to improve motif activity inference. Furthermore, in this thesis, I present my findings on the phylotypic stage, a stage that is supposedly highly conserved between species. I explain how its many definitions are ambiguous, which leads to unsupported conclusions from the data. Altogether, this thesis provides computational perspectives on transcription regulation throughout evolution and development.

Chapter 8

Appendix

8.1 Summary in Dutch

Een volwassen mens bestaat gemiddeld uit ongeveer 37 biljoen (37.000.000.000.000) cellen, die allemaal oorspronkelijk afkomstig zijn van één enkele eicel. Tijdens de embryonale ontwikkeling vermenigvuldigt deze eicel zich, wat leidt tot steeds meer cellen en een toenemende diversiteit tussen deze cellen. Vroeg in de ontwikkeling ontstaan drie belangrijke kiemlijnen: het endoderm, ectoderm en mesoderm. Grof genomen ontwikkelen endodermale cellen zich tot de ingewanden, ectodermale cellen tot de huid en het zenuwstelsel, en mesodermale cellen tot het skelet, bloed en spieren. Hoewel al deze cellen hetzelfde DNA delen, verschillen ze uiteindelijk sterk van elkaar en voeren ze verschillende taken uit. Dit is mogelijk door een sterke regulatie van genexpressie. In dit proefschrift bespreek ik de computationele analyse van genexpressie, met een focus op transcriptieregulatie.

In **hoofdstuk 1** geef ik een inleiding over de onderwerpen die worden behandeld in dit proefschrift. Eerst behandel ik de regulatie van genexpressie via transcriptiefactoren, de relevantie van de staat van chromatine rondom genen, en geef ik een beknopte beschrijving van andere regulatoire mechanismen. Vervolgens bespreek ik hoe genexpressie wordt onderzocht, waarbij zowel het experimentele deel in het laboratorium als de daaropvolgende computationele analyse aan bod komen. Tot slot introduceer ik het vakgebied van evolutionaire ontwikkelingsbiologie (evo-devo). Evolutionair ontwikkelingsbiologen zijn geïnteresseerd in zowel de conservering als de verschillen in genregulatie tijdens de (embryonale) ontwikkeling tussen organismen.

In **hoofdstuk 2** bespreek ik de huidige stand van zaken rondom de analyse van genregulatoire netwerken (GRNs) en doe ik drie aanbevelingen aan het veld. Mijn eerste aanbeveling is de integratie van meerdere bronnen van informatie over genregulatie. Huidige GRNs zijn vaak uitsluitend gebaseerd op RNA transcripten, terwijl het aantal transcripten slechts beperkt correleert met het aantal eiwitten van hetzelfde gen. Mijn tweede aanbeveling is het gebruik van single-cell technologie. Single-cell sequencing biedt een zuiverder signaal dan "bulk" sequencing, wat een positief effect kan hebben op GRN-analyse. Als laatste aanbeveling stel ik voor om universele GRN-modellen te ontwikkelen. De huidige netwerken zijn vaak gebaseerd op verschillen tussen twee condities, terwijl uiteindelijk dezelfde set instructies (DNA) wordt gedeeld door alle mogelijke GRNs.

In **hoofdstuk 3** introduceer ik seq2science, een computationeel hulpmiddel voor de preprocessing van sequencing data uit de functionele genetica. Het is zowel mogelijk om transcriptionele data (RNA-seq) als genomische data (ChIP-/ATAC-seq) te processen. Het is geïntegreerd met de meeste gangbare publieke databases en genereert een uitgebreid rapport aan het einde van het proces.

Hoofdstukken 4 en 5 zijn gewijd aan de definitie en analyse van het fylogenetische stadium. Het fylogenetische stadium is een fase tijdens de embryonale ontwikkeling waarin embryo's uit dezelfde fylogenetische stam sterk op elkaar zouden lijken, wat heeft geleid tot het idee dat dit stadium evolutionair gezien geconserveerd is. Recentelijk zijn er kwantitatieve studies uitgevoerd naar zowel de morfologische als moleculaire gelijkenis tussen embryo's. In deze hoofdstukken bespreek ik de methodologische uitdagingen van deze analyses en heranalyseer ik eerdere studies. Zo laat ik onder andere zien dat de temporele conservering binnen soorten vaak sterker is dan tussen soorten, wat de analyse compliceert. Bovendien weerleg ik het omgekeerde zandlopermodel voor vergelijkingen tussen fylogenetische stammen als een methodologisch artefact. Al met al is er weinig kwantitatief bewijs voor het bestaan van het fylogenetische stadium.

In **hoofdstuk 6** introduceer ik SCEPIA, een computationele methode die ondersteunt bij het schatten van de motiefactiviteit van transcriptiefactoren voor single-cell data. Aangezien transcripten slechts beperkt correleren met het aantal eiwitten van hetzelfde gen, koppelt SCEPIA eerst de transcriptie-informatie van een cel aan een referentiedatabase van H3K27ac. Door vervolgens op deze database de motiefactiviteit te schatten en deze te correleren met het aantal transcripten van dezelfde transcriptiefactor, levert SCEPIA aanzienlijk nauwkeurigere schattingen dan als je alleen naar transcripten kijken. Daarnaast valideer ik de resultaten van SCEPIA met behulp van een

dataset van het menselijk hart.
Ten slotte vat ik in **hoofdstuk 7** mijn eerdere resultaten samen en bespreek ik ze in een bredere context. Ik benadruk daarbij de noodzaak van een systematische weergave van genregulatoire netwerken. Daarnaast uit ik kritiek op het onderzoek naar het fylogenetische stadium, waarbij ik voorbeelden geef van inconsistent gedrag van zowel onderzoekers als hun resultaten, en het naar mijn inzicht dogmatische geloof in het bestaan van het fylogenetische stadium. Tot slot deel ik mijn zorgen over de huidige stand van de bioinformatica, waarin veel computationele methoden en bestanden gebaseerd zijn op verourderde en simplistische aannames.

8.2 Data management

All studies in this thesis are based on publicly available data, as such no raw data is stored. Processed files and scripts however are stored. The seq2science codebase (**Chapter 3**) is publicly hosted on GitHub (https://github.com/vanheeringen-lab/seq2science) and Zenodo (https://zenodo.org/records/10137648). The code for analysis of the phylotypic stage of **Chapter 4** is hosted on GitHub (https://github.com/vanheeringen-lab/phylotypic_hourglass). For **Chapter 6** the code of SCEPIA is hosted on Github (https://github.com/vanheeringen-lab/scepia)

Big Thanks!

8.3 Acknowledgements

While I like to pretend it was all a solo project, it is clear that I would not have been able to undertake this project without the help and support of many people.

First and foremost, I would like to thank my daily supervisor, **Simon**. Simon, thank you for providing a space where I could freely research whatever I wanted, and where you always took the time and effort to help me. Moreover, something I never expressed, but I sincerely want to thank you is for staying on as my supervisor and spending your Friday mornings with me even though you got another job. In a similar spirit, I would like to thank my promoter, **Gert Jan**, for his guidance and insights throughout this process.

Next, I would like to thank all my close bioinformatics colleagues. First, my paranymphs **Rebecca** and **Siebren**, with whom I spent our time in Leuven intoxicated. Rebecca, thanks for sharing your many extended coffee breaks with me, where we discussed a wide variety of topics such as load-bearing walls, ideal coffee-machine settings, and ¡collaborashíon! And Siebren, for our fun venting sessions when everyone else had already left the office. Next, a shoutout to my different deskmates over the years, **Quan**, **Tilman**, and **Sybren**, who have patiently suffered my sighs and moans during the day and were always in for a nice chat. Up next are **Jos** and **Janou** for your (relative) optimism and positivity. Finally, "my" students **Marlien**, **Okan**, **Emma**, **Jules**, and **Justin**, for helping me with my projects, but most of all for the fun of having you around.

Big-up to my close friends **Mart**, **Sjoerd**, **Jonas**, **Elias**, **Vincent**, and **Jim** for sympathizing and pretending to understand my ramblings about my PhD over the years.

Many thanks to **Josephine** for her delicious curries, keeping me well-fed during the process. Thanks to my outdoor pals **Guusje**, **Puk**, **Sam**, **Nuno**, **Paco**, **Ginkgo**, and **Barry** for their encouragements. And thanks to my new colleagues at **Solynta** for the warm welcome and encouraging environment to finish my PhD.

I want to thank my parents, **Rob** and **Janine**, for their support during my doctoral studies, always asking engaging questions such as "why haven't you finished yet?" and "what's taking so long?". Jokes aside, I appreciate the freedom provided to pursue my academic interests and your confidence in my abilities.

Lastly, to **Magda**, for your help and support, but most of all, for loving me.

8.4 List of scientific publications

van der Sande M*, Frölich S*, Schäfers T, Smits JGA, Snabel RR, Rinzema S, van Heeringen SJ. **Seq2science: an end-to-end workflow for functional genomics analysis.** PeerJ 11:e16380 2024.

Frölich S, van der Sande M, Schäfers T, van Heeringen SJ. **Genomepy: genes and genomes at your fingertips.** Bioinformatics 39 (3), btad119, 2023.[267]

van der Sande M*, Frölich* S, van Heeringen SJ. **Computational approaches to understand transcription regulation in development.** Biochemical Society Transactions 51 (1), 1-12, 2023.

Wester RA, van Voorthuijsen L, Neikes HK, Dijkstra JJ, Lamers LA, Frölich S, van der Sande M, Logie C, Lindeboom RGH, Vermeulen M. **Retinoic acid signaling drives differentiation toward the absorptive lineage in colorectal cancer.** Iscience 24 (12), 103444, 2021.

Xu Q, Georgiou G, Frölich S, van der Sande M, Veenstra GJC, Zhou H, van Heeringen SJ. **ANANSE: An enhancer network-based computational approach for predicting key transcription factors in cell fate determination.** Nucleic Acids Research 49 (14), 7966-7985, 2021.

Bright AR, van Genesen S, Li Q, Grasso A, Frölich S, van der Sande M, van Heeringen SJ, Veenstra GJC. **Combinatorial transcription factor activities on open chromatin induce embryonic heterogeneity in vertebrates.** The EMBO Journal 40 (9), e104913, 2021.

Vierboom MPM, Dijkman K, Sombroek CC, Hofman SO, Boot C, Vervenne RAW, Haanstra KG, van der Sande M, van Emst L, Domínguez-Andrés J, Moorlag SJCFM, Kocken CHM, Thole J, Rodriguez E, Puentes E, Martes JHA, van Crevel R, Netea MG, Aguila N, Martin C, Verreck FAW. **Stronger induction of trained immunity by mucosal BCG or MTBVAC vaccination compared to standard intradermal vaccination.** Cell Reports Medicine 2 (1), 100185, 2021.

van der Sande M, Kraus Y, Houliston E, Kaandorp J. **A cell-based boundary model of gastrulation by unipolar ingression in the hydrozoan cnidarian Clytia hemisphaerica.** Developmental biology 460 (2), 176-186, 2020.

* These authors contributed equally to this work.

8.5 List of open-source scientific software contributions

Qnorm: The development of a Python quantile normalization package. The pre-existing implementations did not resolve ties correctly. This implementation is correct, fast-ish, and scales to infinitely large tables. It has been installed 40k+ times through BioConda, and is a requirement of 30+ tools such as Basenji, Gimmemotifs, and TF-COMB.

GimmeMotifs: 17 merged Pull Requests; most notably the addition of the motif2factors command which automatically converts a motif database to any species based on orthology. Moreover, these PRs include numerous bug fixes and documentation improvements.

Snakemake: 9 merged Pull Requests; most notably the addition of automated keyword unpacking and the addition of LockExpection, but also include numerous bug fixes and documentation improvements.

MultiQC: 5 merged Pull Requests; with the addition of the BUStools module, the addition of show/hide buttons, and improvements to the MACS2 module.

BioConda & Conda-Forge; maintaining 9 packages (seq2science, ngs-tools, kb-python, peaksql, pyseq-align, pretty_html_table, qnorm, iteround, and loompy).

Pysradb: 4 merged pull requests; including performance improvement by multithreaded loading, and bug fixes.

Snakemake wrappers. The addition and maintenance of the Genomepy and encode-download wrappers.

MACS2: 1 merged pull request; which adds the option to include more than 2 replicates for peak calling.

gffread: 1 merged pull request; which removes duplicate code.

fluff: 1 merged pull request; fixing a bug concerning not correctly removing temporary files.

kbpython: 1 merged pull request; extending the recognized nucleotides with the full IUPAC code.

8.6 Curriculum Vitae

Maarten van der Sande was born on August 13th 1993, in the municipality of Renkum, the Netherlands. After completing his secondary education at Dorenweerd College, he started his bachelor's at the Universiteit van Amsterdam. Here, he simultaneously obtained two bachelor's degrees, one in Biology, and one in Liberal Arts and Sciences. Maarten, however, spent most of his time doing extracurricular activities such as competitive rowing, race biking, fencing, and trying to remember where he last parked his bike. He finished his studies with a master's degree in Computational Science, a joint program of the Universiteit van Amsterdam and the Vrije Universiteit. At this time Maarten developed a healthy aversion to busy work, like writing Curriculum Vitae sections for PhD theses. After his studies in Amsterdam, Maarten went on to get a PhD degree at the Radboud Universiteit. Here, under the supervision of Simon van Heeringen, Maarten studied the computational analysis of transcription regulation during (evolutionary) development. During this time, Maarten also conducted aerodynamic field tests between Nijmegen and Wageningen, and concluded that he is particularly accident-prone. In December 2023, Maarten started a new job as a bioinformatician at Solynta, the world's most innovative potato breeding company.

8.7 References

[1] Eva Bianconi et al. "An estimation of the number of cells in the human body". In: *Annals of Human Biology* 40.6 (July 2013), pp. 463–471. DOI: 10.3109/03014460.2013.807878.

[2] Ron Sender, Shai Fuchs, and Ron Milo. "Revised Estimates for the Number of Human and Bacteria Cells in the Body". In: *PLOS Biology* 14.8 (Aug. 2016), e1002533. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1002533.

[3] Matthew Cobb. "60 years ago, Francis Crick changed the logic of biology". In: *PLOS Biology* 15.9 (Sept. 2017), e2003243. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2003243.

[4] Allison Piovesan et al. "Human protein-coding genes and gene feature statistics in 2019". In: *BMC Research Notes* 12.1 (June 2019). DOI: 10.1186/s13104-019-4343-8.

[5] D. Graur et al. "On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE". In: *Genome Biology and Evolution* 5.3 (Feb. 2013), pp. 578–590. DOI: 10.1093/gbe/evt028.

[6] The ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. DOI: 10.1038/nature11247.

[7] Manolis Kellis et al. "Defining functional DNA elements in the human genome". In: *Proceedings of the National Academy of Sciences* 111.17 (Apr. 2014), pp. 6131–6138. ISSN: 1091-6490. DOI: 10.1073/pnas.1318948111.

[8] Vanja Haberle and Alexander Stark. "Eukaryotic core promoters and the functional basis of transcription initiation". In: *Nature Reviews Molecular Cell Biology* 19.10 (June 2018), pp. 621–637. DOI: 10.1038/s41580-018-0028-8.

[9] François Spitz and Eileen E. M. Furlong. "Transcription factors: from enhancer binding to developmental control". In: *Nature Reviews Genetics* 13.9 (Aug. 2012), pp. 613–626. DOI: 10.1038/nrg3207.

[10] Gergely Nagy and Laszlo Nagy. "Motif grammar: The basis of the language of gene expression". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2026–2032. DOI: 10.1016/j.csbj.2020.07.007.

[11] David N. Arnosti and Meghana M. Kulkarni. "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" In: *Journal of Cellular Biochemistry* 94.5 (2005), pp. 890–898. DOI: 10.1002/jcb.20352.

[12] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution". In: *Cell* 167.5 (Nov. 2016), pp. 1170–1187. DOI: 10.1016/j.cell.2016.09.018.

[13] Menno P. Creyghton et al. "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences* 107.50 (Nov. 2010), pp. 21931–21936. DOI: 10.1073/pnas.1016071107.

[14] Artem Barski et al. "High-Resolution Profiling of Histone Methylations in the Human Genome". In: *Cell* 129.4 (May 2007), pp. 823–837. DOI: 10.1016/j.cell.2007.05.009.

[15] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* 20.4 (Jan. 2019), pp. 207–220. DOI: 10.1038/s41576-018-0089-8.

[16] Cizhong Jiang and B. Franklin Pugh. "Nucleosome positioning and gene regulation: advances through genomics". In: *Nature Reviews Genetics* 10.3 (Mar. 2009), pp. 161–172. DOI: 10.1038/nrg2522.

[17] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA Methylation and Its Basic Function". In: *Neuropsychopharmacology* 38.1 (July 2012), pp. 23–38. DOI: 10.1038/npp.2012.112.

[18] Maxim V. C. Greenberg and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease". In: *Nature Reviews Molecular Cell Biology* 20.10 (Aug. 2019), pp. 590–607. DOI: 10.1038/s41580-019-0159-6.

[19] Jia Yu and J. Eric Russell. "Structural and Functional Analysis of an mRNP Complex That Mediates the High Stability of Human beta-Globin mRNA". In: *Molecular and Cellular Biology* 21.17 (Sept. 2001), pp. 5879–5888. DOI: 10.1128/mcb.21.17.5879-5888.2001.

[20] Shahin Ramazi and Javad Zahiri. "Post-translational modifications in proteins: resources, tools and prediction methods". In: *Database* 2021 (Jan. 2021). DOI: 10.1093/database/baab012.

[21] Philip Cohen. "The origins of protein phosphorylation". In: *Nature Cell Biology* 4.5 (May 2002), E127–E130. DOI: 10.1038/ncb0502-e127.

[22] Magdalena J. Mazur and Harrold A. van den Burg. "Global SUMO Proteome Responses Guide Gene Regulation, mRNA Biogenesis, and Plant Stress Responses". In: *Frontiers in Plant Science* 3 (2012). DOI: 10.3389/fpls.2012.00215.

[23] Michael Levin. "Endogenous bioelectrical networks store non-genetic patterning information during development and regeneration". In: *The Journal of Physiology* 592.11 (May 2014), pp. 2295–2305. DOI: 10.1113/jphysiol.2014.271940.

[24] Roy J. Britten and Eric H. Davidson. "Gene Regulation for Higher Cells: A Theory". In: *Science* 165.3891 (July 1969), pp. 349–357. DOI: 10.1126/science.165.3891.349.

[25] Eric H. Davidson et al. "A Provisional Regulatory Gene Network for Specification of Endomesoderm in the Sea Urchin Embryo". In: *Developmental Biology* 246.1 (June 2002), pp. 162–190. DOI: 10.1006/dbio.2002.0635.

[26] Alan Turing. "The chemical basis of morphogenesis". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 237.641 (Aug. 1952), pp. 37–72. DOI: 10.1098/rstb.1952.0012.

[27] Bin Zhang and Steve Horvath. "A General Framework for Weighted Gene Co-Expression Network Analysis". In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (Jan. 2005). DOI: 10.2202/1544-6115.1128.

[28] Adam A Margolin et al. "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". In: *BMC Bioinformatics* 7.S1 (Mar. 2006). DOI: 10.1186/1471-2105-7-s1-s7.

[29] Quan Xu et al. "ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination". In: *Nucleic Acids Research* 49.14 (July 2021), pp. 7966–7985. DOI: 10.1093/nar/gkab598.

[30] Aryan Kamal et al. "GRaNIE and GRaNPA: Inference and evaluation of enhancer-mediated gene regulatory networks applied to study macrophages". In: (Dec. 2021). DOI: 10.1101/2021.12.18.473290.

[31] Jason D. Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Current Protocols in Molecular Biology* 109.1 (Jan. 2015). DOI: 10.1002/0471142727.mb2129s109.

[32] Gordon Robertson et al. "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing". In: *Nature Methods* 4.8 (June 2007), pp. 651–657. DOI: 10.1038/nmeth1068.

[33] Ugrappa Nagalakshmi et al. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing". In: *Science* 320.5881 (June 2008), pp. 1344–1349. DOI: 10.1126/science.1158441.

[34] Bruno Canard and Robert S. Sarfati. "DNA polymerase fluorescent substrates with reversible 3′-tags". In: *Gene* 148.1 (Oct. 1994), pp. 1–6. DOI: 10.1016/0378-1119(94)90226-7.

[35] Miten Jain et al. "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community". In: *Genome Biology* 17.1 (Nov. 2016). DOI: 10.1186/s13059-016-1103-0.

[36] M. J. Levene et al. "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations". In: *Science* 299.5607 (Jan. 2003), pp. 682–686. DOI: 10.1126/science.1079700.

[37] Yong Zhang et al. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (Sept. 2008). DOI: 10.1186/gb-2008-9-9-r137.

[38] Endre Bakken Stovner and Pål Sætrom. "PyRanges: efficient comparison of genomic intervals in Python". In: *Bioinformatics* 36.3 (Aug. 2019). Ed. by John Hancock, pp. 918–919. DOI: 10.1093/bioinformatics/btz615.

[39] Evan D Tarbell and Tao Liu. "HMMRATAC: a Hidden Markov ModeleR for ATAC-seq". In: *Nucleic Acids Research* 47.16 (June 2019), e91–e91. DOI: 10.1093/nar/gkz533.

[40] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DE-Seq2". In: *Genome Biology* 15.12 (Dec. 2014). DOI: 10.1186/s13059-014-0550-8.

[41] Jason D. Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561 (June 2015), pp. 486–490. DOI: 10.1038/nature14590.

[42] Fuchou Tang et al. "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5 (Apr. 2009), pp. 377–382. DOI: 10.1038/nmeth.1315.

[43] Ruochen Jiang et al. "Statistics or biology: the zero-inflation controversy about scRNA-seq data". In: *Genome Biology* 23.1 (Jan. 2022). DOI: 10.1186/s13059-022-02601-5.

[44] Stephan Schneuwly, Roman Klemenz, and Walter J. Gehring. "Redesigning the body plan of Drosophila by ectopic expression of the homoeotic gene Antennapedia". In: *Nature* 325.6107 (Feb. 1987), pp. 816–818. DOI: 10.1038/325816a0.

[45] Scott D. Weatherbee et al. "Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the Drosophila haltere". In: *Genes &amp Development* 12.10 (May 1998), pp. 1474–1482. DOI: 10.1101/gad.12.10.1474.

[46] Cynthia L. Hughes and Thomas C. Kaufman. "Hox genes and the evolution of the arthropod body plan1". In: *Evolution and Development* 4.6 (Nov. 2002), pp. 459–499. DOI: 10.1046/j.1525-142x.2002.02034.x.

[47] Stephen J. Gaunt. "The significance of Hox gene collinearity". In: *The International Journal of Developmental Biology* 59.4-5-6 (2015), pp. 159–170. DOI: 10.1387/ijdb.150223sg.

[48] Robert L. Chow et al. "Pax6 induces ectopic eyes in a vertebrate". In: *Development* 126.19 (Oct. 1999), pp. 4213–4222. DOI: 10.1242/dev.126.19.4213.

[49] Ernst Haeckel. *Generelle morphologie der organismen*. Berlin: G. Reimer, 1866.

[50] Georgy S. Levit and Uwe Hossfeld. "Ernst Haeckel, Nikolai Miklucho-Maclay and the racial controversy over the Papuans". In: *Frontiers in Zoology* 17.1 (May 2020). DOI: 10.1186/s12983-020-00358-w.

[51] Robert J. Richards. "Haeckel's embryos: fraud not proven". In: *Biology and Philosophy* 24.1 (Nov. 2008), pp. 147–154. DOI: 10.1007/s10539-008-9140-z.

[52] Elizabeth Pennisi. "Haeckel's Embryos: Fraud Rediscovered". In: *Science* 277.5331 (Sept. 1997), pp. 1435–1435. DOI: 10.1126/science.277.5331.1435a.

[53] Karl Ernst von Baer. *Über Entwickelungsgeschichte der Thiere*. Königsberg: Bornträger, 1828.

[54] P. B. Medawar. "The Significance of Inductive Relationships in the Development of Vertebrates". In: *Development* 2.2 (June 1954), pp. 172–174. DOI: 10.1242/dev.2.2.172.

[55] Denis Duboule. "Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony". In: *Development* 1994.Supplement (Jan. 1994), pp. 135–142. DOI: 10.1242/dev.1994.supplement.135.

[56] Michal Levin et al. "The mid-developmental transition and the evolution of animal body plans". In: *Nature* 531.7596 (Feb. 2016), pp. 637–641. DOI: 10.1038/nature16994.

[57] Ferdinand Marlétaz et al. "Amphioxus functional genomics and the origins of vertebrate gene regulation". In: *Nature* 564.7734 (Nov. 2018), pp. 64–70. DOI: 10.1038/s41586-018-0734-6.

[58] Yoav Mayshar et al. "Time-aligned hourglass gastrulation models in rabbit and mouse". In: *Cell* (May 2023). DOI: 10.1016/j.cell.2023.04.037.

[59] Jialin Liu et al. "The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection". In: *Genome Research* 31.9 (July 2021), pp. 1573–1581. DOI: 10.1101/gr.275212.121.

[60] Tomislav Domazet-Lošo and Diethard Tautz. "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns". In: *Nature* 468.7325 (Dec. 2010), pp. 815–818. DOI: 10.1038/nature09632.

[61] Naoki Irie and Shigeru Kuratani. "Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis". In: *Nature Communications* 2.1 (Mar. 2011). DOI: 10.1038/ncomms1248.

[62] Alex T. Kalinka et al. "Gene expression divergence recapitulates the developmental hourglass model". In: *Nature* 468.7325 (Dec. 2010), pp. 811–814. DOI: 10.1038/nature09634.

[63] Barbara Piasecka et al. "The Hourglass and the Early Conservation Models—Co-Existing Patterns of Developmental Constraints in Vertebrates". In: *PLoS Genetics* 9.4 (Apr. 2013). Ed. by Gregory S. Barsh, e1003476. DOI: 10.1371/journal.pgen.1003476.

[64] Masahiro Uesaka et al. "Recapitulation-like developmental transitions of chromatin accessibility in vertebrates". In: *Zoological Letters* 5.1 (Nov. 2019). DOI: 10.1186/s40851-019-0148-9.

[65] Ann Rose Bright et al. "Combinatorial transcription factor activities on open chromatin induce embryonic heterogeneity in vertebrates". In: *The EMBO Journal* 40.9 (Feb. 2021). DOI: 10.15252/embj.2020104913.

[66] Michel P.M. Vierboom et al. "Stronger induction of trained immunity by mucosal BCG or MTBVAC vaccination compared to standard intradermal vaccination". In: *Cell Reports Medicine* 2.1 (Jan. 2021), p. 100185. DOI: 10.1016/j.xcrm.2020.100185.

[67] Jos G.A. Smits et al. "Characterization of In Vitro Differentiation of Human Primary Keratinocytes by RNA-Seq Analysis". In: *Journal of Visualized Experiments* 159 (May 2020). DOI: 10.3791/60905.

[68] Jos GA Smits et al. "Multi-omics analyses identify transcription factor interplay in corneal epithelial fate determination and disease". In: (July 2022). DOI: 10.1101/2022.07.13.499857.

[69] B. M. H. Heuts et al. "Identification of transcription factors dictating blood cell development using a bidirectional transcription network-based computational framework". In: *Scientific Reports* 12.1 (Nov. 2022). DOI: 10.1038/s41598-022-21148-w.

[70] Nils Rother et al. "#3830 TRAINED INNATE IMMUNITY IN RESPONSE TO NUCLEAR ANTIGENS IN SYSTEMIC LUPUS ERYTHEMATOSUS". In: *Nephrology Dialysis Transplantation* 38.Supplement_1 (June 2023). DOI: 10.1093/ndt/gfad063c_3830.

[71] Nataschja I Ho et al. "Saponin-based adjuvants enhance antigen cross-presentation in human CD11c+ CD1c+ CD5- CD163+ conventional type 2 dendritic cells". In: *Journal for ImmunoTherapy of Cancer* 11.8 (Aug. 2023), e007082. DOI: 10.1136/jitc-2023-007082.

[72] Mark W. D. Sweep et al. "Case Report: A severe case of immunosuppressant-refractory immune checkpoint inhibitor-mediated colitis rescued by tofacitinib". In: *Frontiers in Immunology* 14 (June 2023). DOI: 10.3389/fimmu.2023.1212432.

[73] Glenn F. van Wigcheren et al. "Myeloid-derived suppressor cells and tolerogenic dendritic cells are distinctively induced by PI3K and Wnt signaling pathways". In: *Journal of Biological Chemistry* (Sept. 2023), p. 105276. DOI: 10.1016/j.jbc.2023.105276.

[74] Jos G. A. Smits et al. "Identification of the regulatory circuit governing corneal epithelial fate determination and disease". In: *PLOS Biology* 21.10 (Oct. 2023). Ed. by Sui Wang, e3002336. DOI: 10.1371/journal.pbio.3002336.

[75] Meri Vattulainen et al. "Deciphering the heterogeneity of differentiating hPSC-derived corneal limbal stem cells through single-cell RNA-sequencing". In: (Nov. 2023). DOI: 10.1101/2023.11.24.568553.

[76] Roelof A. Wester et al. "Retinoic acid signaling drives differentiation toward the absorptive lineage in colorectal cancer". In: *iScience* 24.12 (Dec. 2021), p. 103444. DOI: 10.1016/j.isci.2021.103444.

[77] Irene Santos-Barriopedro, Guido van Mierlo, and Michiel Vermeulen. "Off-the-shelf proximity biotinylation for interaction proteomics". In: *Nature Communications* 12.1 (Aug. 2021). DOI: 10.1038/s41467-021-25338-4.

[78] Branco M. H. Heuts et al. "Inducible MLL-AF9 Expression Drives an AML Program during Human Pluripotent Stem Cell-Derived Hematopoietic Differentiation". In: *Cells* 12.8 (Apr. 2023), p. 1195. DOI: 10.3390/cells12081195.

[79] Lotte E. Tholen et al. "Transcription factor HNF1Beta controls a transcriptional network regulating kidney cell structure and tight junction integrity". In: *American Journal of Physiology-Renal Physiology* 324.2 (Feb. 2023), F211–F224. DOI: 10.1152/ajprenal.00199.2022.

[80] Niels Harlaar et al. "Conditional immortalization of human atrial myocytes for the generation of in vitro models of atrial fibrillation". In: *Nature Biomedical Engineering* 6.4 (Jan. 2022), pp. 389–402. DOI: 10.1038/s41551-021-00827-5.

[81] Maria V Luna Velez et al. "ONECUT2 regulates RANKL-dependent enterocyte and microfold cell differentiation in the small intestine & a multi-omics study". In: *Nucleic Acids Research* 51.3 (Jan. 2023), pp. 1277–1296. DOI: 10.1093/nar/gkac1236.

[82] Hannah K. Neikes et al. "Quantification of absolute transcription factor binding affinities in the native chromatin context using BANC-seq". In: *Nature Biotechnology* (Mar. 2023). DOI: 10.1038/s41587-023-01715-w.

[83] Akiko Mammoto, Tadanori Mammoto, and Donald E. Ingber. "Mechanosensitive mechanisms in transcriptional regulation". In: *Journal of Cell Science* (Jan. 2012). DOI: 10.1242/jcs.093005.

[84] R A Cameron et al. "Lineage and fate of each blastomere of the eight-cell sea urchin embryo." In: *Genes & Development* 1.1 (Mar. 1987), pp. 75–85. DOI: 10.1101/gad.1.1.75.

[85] Geoffrey M. Cooper. *The Cell: A Molecular Approach. 2nd edition.* Sinauer Associates, 2000.

[86] Ya-Jun Li et al. "Opening the chromatin for transcription". In: *The International Journal of Biochemistry & Cell Biology* 36.8 (Aug. 2004), pp. 1411–1423. DOI: 10.1016/j.biocel.2003.11.003.

[87] A P McMahon et al. "Inducible expression of a cloned heat shock fusion gene in sea urchin embryos." In: *Proceedings of the National Academy of Sciences* 81.23 (Dec. 1984), pp. 7490–7494. DOI: 10.1073/pnas.81.23.7490.

[88] Raluca Gordân, Alexander J. Hartemink, and Martha L. Bulyk. "Distinguishing direct versus indirect transcription factor extendashDNA interactions". In: *Genome Research* 19.11 (Aug. 2009), pp. 2090–2100. DOI: 10.1101/gr.094144.109.

[89] Chen Chen et al. "DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks". In: *BMC Bioinformatics* 22.1 (Feb. 2021). DOI: 10.1186/s12859-020-03952-1.

[90] Juan M. Vaquerizas et al. "A census of human transcription factors: function, expression and evolution". In: *Nature Reviews Genetics* 10.4 (Apr. 2009), pp. 252–263. DOI: 10.1038/nrg2538.

[91] Samuel A. Lambert et al. "The Human Transcription Factors". In: *Cell* 172.4 (Feb. 2018), pp. 650–665. DOI: 10.1016/j.cell.2018.01.029.

[92] Arnau Seb'e-Pedr'os et al. "Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq". In: *Cell* 173.6 (May 2018), 1520–1534.e20. DOI: 10.1016/j.cell.2018.05.019.

[93] Kazuhiro R Nitta et al. "Conservation of transcription factor binding specificities across 600 million years of bilateria evolution". In: *eLife* 4 (Mar. 2015). DOI: 10.7554/elife.04837.

[94] Dominic Schmidt et al. "Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding". In: *Science* 328.5981 (May 2010), pp. 1036–1040. DOI: 10.1126/science.1186176.

[95] Diego Villar, Paul Flicek, and Duncan T. Odom. "Evolution of transcription factor binding in metazoans extemdash mechanisms and functional implications". In: *Nature Reviews Genetics* 15.4 (Mar. 2014), pp. 221–233. DOI: 10.1038/nrg3481.

[96] Michael Levine and Eric H. Davidson. "Gene regulatory networks for development". In: *Proceedings of the National Academy of Sciences* 102.14 (Mar. 2005), pp. 4936–4942. DOI: 10.1073/pnas.0408031102.

[97] Ziga Avsec et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax". In: *Nature Genetics* 53.3 (Feb. 2021), pp. 354–366. DOI: 10.1038/s41588-021-00782-6.

[98] Christopher D. Brown, David S. Johnson, and Arend Sidow. "Functional Architecture and Evolution of Transcriptional Elements That Drive Gene Coexpression". In: *Science* 317.5844 (Sept. 2007), pp. 1557–1560. DOI: 10.1126/science.1145893.

[99] Emma K. Farley et al. "Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers". In: *Proceedings of the National Academy of Sciences* 113.23 (May 2016), pp. 6508–6513. DOI: 10.1073/pnas.1605085113.

[100] Emily S. Wong et al. "Deep conservation of the enhancer regulatory code in animals". In: *Science* 370.6517 (Nov. 2020). DOI: 10.1126/science.aax8137.

[101] Julia Zeitlinger. "Seven myths of how transcription factors read the cis-regulatory code". In: *Current Opinion in Systems Biology* 23 (Oct. 2020), pp. 22–31. DOI: 10.1016/j.coisb.2020.08.002.

[102] Mark Q. Martindale. "The evolution of metazoan axial properties". In: *Nature Reviews Genetics* 6.12 (Nov. 2005), pp. 917–927. DOI: 10.1038/nrg1725.

[103] Gregory A. Cary et al. "Systematic comparison of sea urchin and sea star developmental gene regulatory networks explains how novelty is incorporated in early development". In: *Nature Communications* 11.1 (Dec. 2020). DOI: 10.1038/s41467-020-20023-4.

[104] Isabelle S. Peter and Eric H. Davidson. "A gene regulatory network controlling the embryonic specification of endoderm". In: *Nature* 474.7353 (May 2011), pp. 635–639. DOI: 10.1038/nature10100.

[105] Alexandra Saudemont et al. "Ancestral Regulatory Circuits Governing Ectoderm Patterning Downstream of Nodal and BMP2/4 Revealed by Gene Regulatory Network Analysis in an Echinoderm". In: *PLoS Genetics* 6.12 (Dec. 2010). Ed. by Mark Martindale, e1001259. DOI: 10.1371/journal.pgen.1001259.

[106] Rebekah M. Charney et al. "A gene regulatory program controlling early Xenopus mesendoderm formation: Network conservation and motifs". In: *Seminars in Cell & Developmental Biology* 66 (June 2017), pp. 12–24. DOI: 10.1016/j.semcdb.2017.03.003.

[107] Tetsuya Koide, Tadayoshi Hayata, and Ken W. Y. Cho. "Xenopus as a model system to study transcriptional regulatory networks". In: *Proceedings of the National Academy of Sciences* 102.14 (Mar. 2005), pp. 4943–4948. DOI: 10.1073/pnas.0408125102.

[108] Scott A. Rankin et al. "A gene regulatory network controlling hhex transcription in the anterior endoderm of the organizer". In: *Developmental Biology* 351.2 (Mar. 2011), pp. 297–310. DOI: 10.1016/j.ydbio.2010.11.037.

[109] D'ebora Sinner et al. "Global analysis of the transcriptional network controlling Xenopus endoderm formation". In: *Development* 133.10 (May 2006), pp. 1955–1966. DOI: 10.1242/dev.02358.

[110] Martyna Lukoseviciute et al. "From Pioneer to Repressor: Bimodal foxd3 Activity Dynamically Remodels Neural Crest Regulatory Landscape In Vivo". In: *Developmental Cell* 47.5 (Dec. 2018), 608–628.e6. DOI: 10.1016/j.devcel.2018.11.009.

[111] Ruth M. Williams et al. "Reconstruction of the Global Neural Crest Gene Regulatory Network In Vivo". In: *Developmental Cell* 51.2 (Oct. 2019), 255–276.e7. DOI: 10.1016/j.devcel.2019.10.003.

[112] Johannes Jaeger. "The gap gene network". In: *Cellular and Molecular Life Sciences* 68.2 (Oct. 2010), pp. 243–274. DOI: 10.1007/s00018-010-0536-y.

[113] Hao Yuan Kueh and Ellen V. Rothenberg. "Regulatory gene network circuits underlying T cell development from multipotent progenitors". In: *WIREs Systems Biology and Medicine* 4.1 (Oct. 2011), pp. 79–102. DOI: 10.1002/wsbm.162.

[114] John E. Pimanda and Berthold Gttgens. "Gene regulatory networks governing haematopoietic stem cell development and identity". In: *The International Journal of Developmental Biology* 54.6-7 (2010), pp. 1201–1211. DOI: 10.1387/ijdb.093038jp.

[115] Harinder Singh, Aly A. Khan, and Aaron R. Dinner. "Gene regulatory networks in the immune system". In: *Trends in Immunology* 35.5 (May 2014), pp. 211–218. DOI: 10.1016/j.it.2014.03.006.

[116] Genevieve E. Ryan and Emma K. Farley. "Functional genomic approaches to elucidate the role of enhancers during development". In: *WIREs Systems Biology and Medicine* 12.2 (Dec. 2019). DOI: 10.1002/wsbm.1467.

[117] "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. DOI: 10.1038/nature11247.

[118] Trey Ideker et al. "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network". In: *Science* 292.5518 (May 2001), pp. 929–934. DOI: 10.1126/science.292.5518.929.

[119] Dominic J Allocco, Isaac S Kohane, and Atul J Butte. In: *BMC Bioinformatics* 5.1 (2004), p. 18. DOI: 10.1186/1471-2105-5-18.

[120] Michael B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25 (Dec. 1998), pp. 14863–14868. DOI: 10.1073/pnas.95.25.14863.

[121] Deborah Chasman and Sushmita Roy. "Inference of cell type specific regulatory networks on mammalian lineages". In: *Current Opinion in Systems Biology* 2 (Apr. 2017), pp. 130–139. DOI: 10.1016/j.coisb.2017.04.001.

[122] Fernando M. Delgado and Francisco G'omez-Vela. "Computational methods for Gene Regulatory Networks reconstruction and analysis: A review". In: *Artificial Intelligence in Medicine* 95 (Apr. 2019), pp. 133–145. DOI: 10.1016/j.artmed.2018.10.006.

[123] Daniele Mercatelli et al. "Gene regulatory network inference resources: A practical overview". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863.6 (June 2020), p. 194430. DOI: 10.1016/j.bbagrm.2019.194430.

[124] Qiye He, Jeff Johnston, and Julia Zeitlinger. "ChIP-nexus enables improved detection of in vivo transcription factor binding footprints". In: *Nature Biotechnology* 33.4 (Mar. 2015), pp. 395–401. DOI: 10.1038/nbt.3121.

[125] Ho Sung Rhee and B. Franklin Pugh. "Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution". In: *Cell* 147.6 (Dec. 2011), pp. 1408–1419. DOI: 10.1016/j.cell.2011.11.013.

[126] Hatice S. Kaya-Okur et al. "CUT&Tag for efficient epigenomic profiling of small samples and single cells". In: *Nature Communications* 10.1 (Apr. 2019). DOI: 10.1038/s41467-019-09982-5.

[127] Jens Keilwagen, Stefan Posch, and Jan Grau. "Accurate prediction of cell type-specific transcription factor binding". In: *Genome Biology* 20.1 (Jan. 2019). DOI: 10.1186/s13059-018-1614-y.

[128] Fatemeh Behjati Ardakani, Florian Schmidt, and Marcel H. Schulz. "Predicting transcription factor binding using ensemble random forest models". In: *F1000Research* 7 (Sept. 2019), p. 1603. DOI: 10.12688/f1000research.16200.2.

[129] Meredith V. Trotter et al. *Epigenomic language models powered by Cerebras*. 2021. DOI: 10.48550/ARXIV.2112.07571.

[130] Ren Yi, Kyunghyun Cho, and Richard Bonneau. "NetTIME: a multitask and base-pair resolution framework for improved transcription factor binding site prediction". In: *Bioinformatics* 38.20 (Aug. 2022). Ed. by Anthony Mathelier, pp. 4762–4770. DOI: 10.1093/bioinformatics/btac569.

[131] Gergely Pap et al. "Transcription factor binding site detection using convolutional neural networks with a functional group-based data representation". In: *Journal of Physics: Conference Series* 1824.1 (Mar. 2021), p. 012001. DOI: 10.1088/1742-6596/1824/1/012001.

[132] Mehran Karimzadeh and Michael M. Hoffman. "Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome". In: *Genome Biology* 23.1 (June 2022). DOI: 10.1186/s13059-022-02690-2.

[133] Peter K. Koo and Matt Ploenzke. "Deep learning for inferring transcription factor binding sites". In: *Current Opinion in Systems Biology* 19 (Feb. 2020), pp. 16–23. DOI: 10.1016/j.coisb.2020.04.001.

[134] (Author Name Not Available). *ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge*. 2018. DOI: 10.7303/SYN6131484.

[135] Hongyang Li, Daniel Quang, and Yuanfang Guan. "Anchor: trans-cell type prediction of transcription factor binding sites". In: *Genome Research* 29.2 (Dec. 2018), pp. 281–292. DOI: 10.1101/gr.237156.118.

[136] Daniel Quang and Xiaohui Xie. "FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data". In: *Methods* 166 (Aug. 2019), pp. 40–47. DOI: 10.1016/j.ymeth.2019.03.020.

[137] Luca Mariani et al. "MEDEA: analysis of transcription factor binding motifs in accessible chromatin". In: *Genome Research* 30.5 (May 2020), pp. 736–748. DOI: 10.1101/gr.260877.120.

[138] Hongyang Li and Yuanfang Guan. "Fast decoding cell type extendash specific transcription factor binding landscape at single-nucleotide resolution". In: *Genome Research* 31.4 (Mar. 2021), pp. 721–731. DOI: 10.1101/gr.269613.120.

[139] Niklas Bruse and Simon J. van Heeringen. "GimmeMotifs: an analysis framework for transcription factor motif analysis". In: (Nov. 2018). DOI: 10.1101/474403.

[140] Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. "Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples". In: *Genome Biology* 21.1 (Mar. 2020). DOI: 10.1186/s13059-020-01978-5.

[141] Jiandong Cheng et al. "AttBind: Prediction of Transcription Factor Binding Sites Across Cell-types Based on Attention Mechanism". In: *2022 7th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, Apr. 2022. DOI: 10.1109/icccs55155.2022.9846215.

[142] Kelly Cochran et al. "Domain-adaptive neural networks improve cross-species prediction of transcription factor binding". In: *Genome Research* 32.3 (Jan. 2022), pp. 512–523. DOI: 10.1101/gr.275394.121.

[143] Alan P. Boyle et al. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell* 132.2 (Jan. 2008), pp. 311–322. DOI: 10.1016/j.cell.2007.12.014.

[144] Jason D Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10.12 (Oct. 2013), pp. 1213–1218. DOI: 10.1038/nmeth.2688.

[145] Hongbo Yang et al. "A map of cis-regulatory elements and 3D genome structures in zebrafish". In: *Nature* 588.7837 (Nov. 2020), pp. 337–343. DOI: 10.1038/s41586-020-2962-9.

[146] M'at'e P'alfy et al. "Chromatin accessibility established by Pou5f3, Sox19b and Nanog primes genes for activity during zebrafish genome activation". In: *PLOS Genetics* 16.1 (Jan. 2020). Ed. by A. Aziz Aboobaker, e1008546. DOI: 10.1371/journal.pgen.1008546.

[147] Tanvi Shashikant, Jian Ming Khor, and Charles A. Ettensohn. "Global analysis of primary mesenchyme cell cis-regulatory modules by chromatin accessibility profiling". In: *BMC Genomics* 19.1 (Mar. 2018). DOI: 10.1186/s12864-018-4542-z.

[148] Alicia Madgwick et al. "Evolution of embryonic cis-regulatory landscapes between divergent Phallusia and Ciona ascidians". In: *Developmental Biology* 448.2 (Apr. 2019), pp. 71–87. DOI: 10.1016/j.ydbio.2019.01.003.

[149] Melody Esmaeili et al. "Chromatin accessibility and histone acetylation in the regulation of competence in early development". In: *Developmental Biology* 462.1 (June 2020), pp. 20–35. DOI: 10.1016/j.ydbio.2020.02.013.

[150] Axel Visel et al. "ChIP-seq accurately predicts tissue-specific activity of enhancers". In: *Nature* 457.7231 (Feb. 2009), pp. 854–858. DOI: 10.1038/nature07730.

[151] Nathaniel D Heintzman et al. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome". In: *Nature Genetics* 39.3 (Feb. 2007), pp. 311–318. DOI: 10.1038/ng1966.

[152] Antonio Simeone et al. "At least three human homeoboxes on chromosome 12 belong to the same transcription unit". In: *Nucleic Acids Research* 16.12 (1988), pp. 5379–5390. DOI: 10.1093/nar/16.12.5379.

[153] Pamela J. Mitchell and Robert Tjian. "Transcriptional Regulation in Mammalian Cells by Sequence-Specific DNA Binding Proteins". In: *Science* 245.4916 (July 1989), pp. 371–378. DOI: 10.1126/science.2667136.

[154] Jay R Hesselberth et al. "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting". In: *Nature Methods* 6.4 (Mar. 2009), pp. 283–289. DOI: 10.1038/nmeth.1313.

[155] Shane Neph et al. "Circuitry and Dynamics of Human Transcription Factor Regulatory Networks". In: *Cell* 150.6 (Sept. 2012), pp. 1274–1286. DOI: 10.1016/j.cell.2012.04.040.

[156] Housheng Hansen He et al. "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification". In: *Nature Methods* 11.1 (Dec. 2013), pp. 73–78. DOI: 10.1038/nmeth.2762.

[157] Myong-Hee Sung, Songjoon Baek, and Gordon L Hager. "Genome-wide footprinting: ready for prime time?" In: *Nature Methods* 13.3 (Feb. 2016), pp. 222–228. DOI: 10.1038/nmeth.3766.

[158] Myong-Hee Sung et al. "DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence". In: *Molecular Cell* 56.2 (Oct. 2014), pp. 275–285. DOI: 10.1016/j.molcel.2014.08.016.

[159] Zhijian Li et al. "Identification of transcription factor binding sites using ATAC-seq". In: *Genome Biology* 20.1 (Feb. 2019). DOI: 10.1186/s13059-019-1642-2.

[160] Mette Bentsen et al. "ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation". In: *Nature Communications* 11.1 (Aug. 2020). DOI: 10.1038/s41467-020-18035-1.

[161] Emily R. Miraldi et al. "Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells". In: *Genome Research* 29.3 (Jan. 2019), pp. 449–463. DOI: 10.1101/gr.238253.118.

[162] Alireza F. Siahpirani and Sushmita Roy. "A prior-based integrative framework for functional transcriptional regulatory network inference". In: *Nucleic Acids Research* (Oct. 2016), gkw963. DOI: 10.1093/nar/gkw963.

[163] Abhijeet Rajendra Sonawane et al. "Constructing gene regulatory networks using epigenetic data". In: *npj Systems Biology and Applications* 7.1 (Dec. 2021). DOI: 10.1038/s41540-021-00208-3.

[164] Jesper Grud Skat Madsen et al. "Integrated analysis of motif activity and gene expression changes of transcription factors". In: *Genome Research* 28.2 (Dec. 2017), pp. 243–255. DOI: 10.1101/gr.227231.117.

[165] Florian Schmidt et al. "TEPIC 2 extemdashan extended framework for transcription factor binding prediction and integrative epigenomic analysis". In: *Bioinformatics* 35.9 (Oct. 2018). Ed. by Bonnie Berger, pp. 1608–1609. DOI: 10.1093/bioinformatics/bty856.

[166] Saba Ghaffari et al. "An integrated multi-omics approach to identify regulatory mechanisms in cancer metastatic processes". In: *Genome Biology* 22.1 (Jan. 2021). DOI: 10.1186/s13059-020-02213-x.

[167] M. S. Vijayabaskar et al. "Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets". In: *PLOS Computational Biology* 15.11 (Nov. 2019). Ed. by Sushmita Roy, e1007337. DOI: 10.1371/journal.pcbi.1007337.

[168] Mike Levine. "Transcriptional Enhancers in Animal Development and Evolution". In: *Current Biology* 20.17 (Sept. 2010), R754–R763. DOI: 10.1016/j.cub.2010.06.070.

[169] Elphège P. Nora et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485.7398 (Apr. 2012), pp. 381–385. DOI: 10.1038/nature11049.

[170] Job Dekker et al. "Capturing Chromosome Conformation". In: *Science* 295.5558 (Feb. 2002), pp. 1306–1311. DOI: 10.1126/science.1067799.

[171] Erez Lieberman-Aiden et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome". In: *Science* 326.5950 (Oct. 2009), pp. 289–293. DOI: 10.1126/science.1181369.

[172] Borbala Mifsud et al. "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". In: *Nature Genetics* 47.6 (May 2015), pp. 598–606. DOI: 10.1038/ng.3286.

[173] Amartya Sanyal et al. "The long-range interaction landscape of gene promoters". In: *Nature* 489.7414 (Sept. 2012), pp. 109–113. DOI: 10.1038/nature11279.

[174] Guoliang Li et al. "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation". In: *Cell* 148.1-2 (Jan. 2012), pp. 84–98. DOI: 10.1016/j.cell.2011.12.014.

[175] Daniel Marbach et al. "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases". In: *Nature Methods* 13.4 (Mar. 2016), pp. 366–370. DOI: 10.1038/nmeth.3799.

[176] Charles P. Fulco et al. "Activity-by-contact model of enhancer extendashpromoter regulation from thousands of CRISPR perturbations". In: *Nature Genetics* 51.12 (Nov. 2019), pp. 1664–1669. DOI: 10.1038/s41588-019-0538-0.

[177] Kimberly Glass et al. "Passing Messages between Biological Networks to Refine Predicted Interactions". In: *PLoS ONE* 8.5 (May 2013). Ed. by Szabolcs Semsey, e64832. DOI: 10.1371/journal.pone.0064832.

[178] Ellen Jensen. "Technical Review: In Situ Hybridization". In: *The Anatomical Record* 297.8 (May 2014), pp. 1349–1353. DOI: 10.1002/ar.22944.

[179] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1 (Jan. 2009), pp. 57–63. DOI: 10.1038/nrg2484.

[180] Sophia K. Longo et al. "Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics". In: *Nature Reviews Genetics* 22.10 (June 2021), pp. 627–644. DOI: 10.1038/s41576-021-00370-8.

[181] Lars E. Borm et al. "Scalable in situ single-cell profiling by electrophoretic capture of mRNA using EEL FISH". In: *Nature Biotechnology* (Sept. 2022). DOI: 10.1038/s41587-022-01455-3.

[182] Lam-Ha Ly and Martin Vingron. "Effect of imputation on gene network reconstruction from single-cell RNA-seq data". In: *Patterns* 3.2 (Feb. 2022), p. 100414. DOI: 10.1016/j.patter.2021.100414.

[183] Sunnie Grace McCalla et al. "Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data". In: (June 2021). DOI: 10.1101/2021.06.01.446671.

[184] Kip D. Zimmerman, Mark A. Espeland, and Carl D. Langefeld. "A practical solution to pseudoreplication bias in single-cell studies". In: *Nature Communications* 12.1 (Feb. 2021). DOI: 10.1038/s41467-021-21038-1.

[185] Yael Baran et al. "MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions". In: *Genome Biology* 20.1 (Oct. 2019). DOI: 10.1186/s13059-019-1812-2.

[186] Vân Anh Huynh-Thu et al. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods". In: *PLoS ONE* 5.9 (Sept. 2010). Ed. by Mark Isalan, e12776. DOI: 10.1371/journal.pone.0012776.

[187] Thalia E. Chan, Michael P.H. Stumpf, and Ann C. Babtie. "Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures". In: *Cell Systems* 5.3 (Sept. 2017), 251–267.e3. DOI: 10.1016/j.cels.2017.08.014.

[188] Sara Aibar et al. "SCENIC: single-cell regulatory network inference and clustering". In: *Nature Methods* 14.11 (Oct. 2017), pp. 1083–1086. DOI: 10.1038/nmeth.4463.

[189] Camden Jansen et al. "Building gene regulatory networks from scATAC-seq and scRNA-seq using Linked Self Organizing Maps". In: *PLOS Computational Biology* 15.11 (Nov. 2019). Ed. by Christina S. Leslie, e1006555. DOI: 10.1371/journal.pcbi.1006555.

[190] Carmen Bravo Gonz'alez-Blas et al. "SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks". In: (Aug. 2022). DOI: 10.1101/2022.08.19.504505.

[191] Junyao Jiang et al. "IReNA: integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles". In: (Nov. 2021). DOI: 10.1101/2021.11.22.469628.

[192] Kenji Kamimoto, Christy M. Hoffmann, and Samantha A. Morris. "CellOracle: Dissecting cell identity via network inference and in silico gene perturbation". In: (Feb. 2020). DOI: 10.1101/2020.02.17.947416.

[193] Jonathan S. Packer et al. "A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution". In: *Science* 365.6459 (Sept. 2019). DOI: 10.1126/science.aax1971.

[194] F. Alexander Wolf et al. "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells". In: *Genome Biology* 20.1 (Mar. 2019). DOI: 10.1186/s13059-019-1663-x.

[195] Xiaojie Qiu et al. "Reversed graph embedding resolves complex single-cell trajectories". In: *Nature Methods* 14.10 (Aug. 2017), pp. 979–982. DOI: 10.1038/nmeth.4402.

[196] Kelly Street et al. "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC Genomics* 19.1 (June 2018). DOI: 10.1186/s12864-018-4772-0.

[197] Volker Bergen et al. "Generalizing RNA velocity to transient cell states through dynamical modeling". In: *Nature Biotechnology* 38.12 (Aug. 2020), pp. 1408–1414. DOI: 10.1038/s41587-020-0591-3.

[198] Gioele La Manno et al. "RNA velocity of single cells". In: *Nature* 560.7719 (Aug. 2018), pp. 494–498. DOI: 10.1038/s41586-018-0414-6.

[199] Pierre-Cyril Aubin-Frankowski and Jean-Philippe Vert. "Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference". In: *Bioinformatics* 36.18 (June 2020). Ed. by Jan Gorodkin, pp. 4774–4780. DOI: 10.1093/bioinformatics/btaa576.

[200] Hirotaka Matsumoto et al. "SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation". In: *Bioinformatics* 33.15 (Apr. 2017). Ed. by Ziv Bar-Joseph, pp. 2314–2321. DOI: 10.1093/bioinformatics/btx194.

[201] Atul Deshpande et al. "Network inference with Granger causality ensembles on single-cell transcriptomics". In: *Cell Reports* 38.6 (Feb. 2022), p. 110333. DOI: 10.1016/j.celrep.2022.110333.

[202] Nan Papili Gao et al. "SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles". In: *Bioinformatics* 34.2 (Sept. 2017). Ed. by Alfonso Valencia, pp. 258–266. DOI: 10.1093/bioinformatics/btx575.

[203] Xiaojie Qiu et al. "Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe". In: *Cell Systems* 10.3 (Mar. 2020), 265–274.e11. DOI: 10.1016/j.cels.2020.02.003.

[204] Steven Woodhouse et al. "SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data". In: *BMC Systems Biology* 12.1 (May 2018). DOI: 10.1186/s12918-018-0581-y.

[205] M Sanchez-Castillo et al. "A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data". In: *Bioinformatics* 34.6 (Sept. 2017). Ed. by Oliver Stegle, pp. 964–970. DOI: 10.1093/bioinformatics/btx605.

[206] Jonathan R. Chubb et al. "Transcriptional Pulsing of a Developmental Gene". In: *Current Biology* 16.10 (May 2006), pp. 1018–1025. DOI: 10.1016/j.cub.2006.03.092.

[207] Anton J. M. Larsson et al. "Genomic encoding of transcriptional burst kinetics". In: *Nature* 565.7738 (Jan. 2019), pp. 251–254. DOI: 10.1038/s41586-018-0836-1.

[208] Elias Ventre. "Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics". In: *In Silico Biology* 14.3-4 (Jan. 2022), pp. 89–113. DOI: 10.3233/isb-210226.

[209] Shuonan Chen and Jessica C. Mar. "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data". In: *BMC Bioinformatics* 19.1 (June 2018). DOI: 10.1186/s12859-018-2217-z.

[210] Aditya Pratapa et al. "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data". In: *Nature Methods* 17.2 (Jan. 2020), pp. 147–154. DOI: 10.1038/s41592-019-0690-6.

[211] Daniel S. Kim et al. "The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation". In: *Nature Genetics* 53.11 (Oct. 2021), pp. 1564–1576. DOI: 10.1038/s41588-021-00947-3.

[212] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (July 2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[213] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. DOI: 10.1145/3065386.

[214] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961.

[215] Gökcen Eraslan et al. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* 20.7 (Apr. 2019), pp. 389–403. DOI: 10.1038/s41576-019-0122-6.

[216] Babak Alipanahi et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". In: *Nature Biotechnology* 33.8 (July 2015), pp. 831–838. DOI: 10.1038/nbt.3300.

[217] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning extendashbased sequence model". In: *Nature Methods* 12.10 (Aug. 2015), pp. 931–934. DOI: 10.1038/nmeth.3547.

[218] David R. Kelley et al. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks". In: *Genome Research* 28.5 (Mar. 2018), pp. 739–750. DOI: 10.1101/gr.227819.117.

[219] Hantao Shu et al. "Modeling gene regulatory networks using neural network architectures". In: *Nature Computational Science* 1.7 (July 2021), pp. 491–501. DOI: 10.1038/s43588-021-00099-8.

[220] M Rubiolo, D H Milone, and G Stegmayer. "Extreme learning machines for reverse engineering of gene regulatory networks from expression time series". In: *Bioinformatics* 34.7 (Nov. 2017). Ed. by Cenk Sahinalp, pp. 1253–1260. DOI: 10.1093/bioinformatics/btx730.

[221] Francis Dutil et al. *Towards Gene Expression Convolutions using Gene Interaction Graphs*. 2018. DOI: 10.48550/ARXIV.1806.06975.

[222] Juexin Wang et al. "Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3335–3343. DOI: 10.1016/j.csbj.2020.10.022.

[223] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314. DOI: 10.1007/bf02551274.

[224] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[225] Yu Zhang et al. "A Survey on Neural Network Interpretability". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (Oct. 2021), pp. 726–742. DOI: 10.1109/tetci.2021.3100641.

[226] François Chollet et al. *Keras*. https://keras.io. 2015.

[227] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. DOI: 10.48550/ARXIV.1912.01703.

[228] Raquel de Sousa Abreu et al. "Global signatures of protein and mRNA expression levels". In: *Molecular BioSystems* (2009). DOI: 10.1039/b908315d.

[229] Arun Krishnan, Alessandro Giuliani, and Masaru Tomita. "Indeterminacy of Reverse Engineering of Gene Regulatory Networks: The Curse of Gene Elasticity". In: *PLoS ONE* 2.6 (June 2007). Ed. by Raya Khanin, e562. DOI: 10.1371/journal.pone.0000562.

[230] Thomas Cokelaer et al. "DREAMTools: a Python package for scoring collaborative challenges". In: *F1000Research* 4 (Apr. 2016), p. 1030. DOI: 10.12688/f1000research.7118.2.

[231] Dream challenge. Available from: https://www.synapse.org/#\protect\protect\leavevmode@ifvmode\kern-.1667em\relaxSynapse:syn21760283/wiki/603540.

[232] Wenbin Guo et al. "Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size". In: *BMC Systems Biology* 11.1 (June 2017). DOI: 10.1186/s12918-017-0440-2.

[233] Daniel Marbach et al. "Wisdom of crowds for robust gene network inference". In: *Nature Methods* 9.8 (July 2012), pp. 796–804. DOI: 10.1038/nmeth.2016.

[234] G. Stolovitzky, D. Monroe, and A. Califano. "Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of High-Throughput Pathway Inference". In: *Annals of the New York Academy of Sciences* 1115.1 (Oct. 2007), pp. 1–22. DOI: 10.1196/annals.1407.021.

[235] Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. "TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach". In: *BMC Bioinformatics* 11.1 (Mar. 2010). DOI: 10.1186/1471-2105-11-154.

[236] Jason Ernst et al. "Reconstructing dynamic regulatory maps". In: *Molecular Systems Biology* 3.1 (Jan. 2007). DOI: 10.1038/msb4100115.

[237] Marcel H Schulz et al. "DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data". In: *BMC Systems Biology* 6.1 (2012), p. 104. DOI: 10.1186/1752-0509-6-104.

[238] Jun Ding et al. "iDREM: Interactive visualization of dynamic regulatory networks". In: *PLOS Computational Biology* 14.3 (Mar. 2018). Ed. by Dina Schneidman, e1006019. DOI: 10.1371/journal.pcbi.1006019.

[239] Ashley Mae Conard et al. "TIMEOR: a web-based tool to uncover temporal regulatory mechanisms from multi-omics data". In: *Nucleic Acids Research* 49.W1 (June 2021), W641–W653. DOI: 10.1093/nar/gkab384.

[240] Sara Bühlmann Peter; van de Geer. "Statistics for High-Dimensional Data: Methods, Theory and Applications". In: (2011).

[241] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958.

[242] Sai Ma et al. "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin". In: *Cell* 183.4 (Nov. 2020), 1103–1116.e20. DOI: 10.1016/j.cell.2020.09.056.

[243] Junyue Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. DOI: 10.1126/science.aau0730.

[244] Song Chen, Blue B. Lake, and Kun Zhang. "High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell". In: *Nature Biotechnology* 37.12 (Oct. 2019), pp. 1452–1457. DOI: 10.1038/s41587-019-0290-0.

[245] Lucy L. Gao, Jacob Bien, and Daniela Witten. *Selective Inference for Hierarchical Clustering*. 2020. DOI: 10.48550/ARXIV.2012.02936.

[246] Payam Dibaeinia and Saurabh Sinha. "SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks". In: *Cell Systems* 11.3 (Sept. 2020), 252–271.e11. DOI: 10.1016/j.cels.2020.08.003.

[247] Hechen Li et al. "scMultiSim: simulation of multi-modality single cell data guided by cell-cell interactions and gene regulatory networks". In: (Oct. 2022). DOI: 10.1101/2022.10.15.512320.

[248] *We want to hear from you about changes to NIHś sequence read archive data format and storage*. https://ncbiinsights.ncbi.nlm.nih.gov/2020/06/30/sra-rfi/. 2020.

[249] David S. Johnson et al. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions". In: *Science* 316.5830 (June 2007), pp. 1497–1502. DOI: 10.1126/science.1141319.

[250] The Galaxy Community. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update". In: *Nucleic Acids Research* 50.W1 (Apr. 2022), W345–W351. ISSN: 0305-1048. DOI: 10.1093/nar/gkac247. eprint: https://academic.oup.com/nar/article-pdf/50/W1/W345/45189566/gkac247.pdf.

[251] *Snakemake-workflows*. https://github.com/snakemake-workflows/. n.d.

[252] Philip A. Ewels et al. "The nf-core framework for community-curated bioinformatics pipelines". In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 276–278. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0439-x.

[253] Vivek Bhardwaj et al. "snakePipes: facilitating flexible, scalable and integrative epigenomic analysis". In: *Bioinformatics* 35.22 (May 2019), pp. 4757–4759. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz436. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/22/4757/30706717/btz436.pdf.

[254] Benjamin C. Hitz et al. "The ENCODE Uniform Analysis Pipelines". In: *bioRxiv* (Apr. 2023). DOI: 10.1101/2023.04.04.535623.

[255] Grace X. Y. Zheng et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8.1 (Jan. 2017). DOI: 10.1038/ncomms14049.

[256] Stephany Orjuela et al. "ARMOR: An Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNA-seq Data". In: *G3 Genes|Genomes|Genetics* 9.7 (July 2019), pp. 2089–2096. DOI: 10.1534/g3.119.400185.

[257] Jason P Smith et al. "PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments". In: *NAR Genomics and Bioinformatics* 3.4 (Oct. 2021). DOI: 10.1093/nargab/lqab101.

[258] Felix Mölder et al. "Sustainable data analysis with Snakemake". In: *F1000Research* 10 (Apr. 2021), p. 33. DOI: 10.12688/f1000research.29032.2.

[259] *Anaconda Software Distribution*. n.d.

[260] R. Leinonen et al. "The European Nucleotide Archive". In: *Nucleic Acids Research* 39.Database (Oct. 2010), pp. D28–D31. DOI: 10.1093/nar/gkq967.

[261] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". In: *Nucleic Acids Research* 41.D1 (Nov. 2012), pp. D991–D995. DOI: 10.1093/nar/gks1193.

[262] R. Leinonen, H. Sugawara, and M. Shumway and. "The Sequence Read Archive". In: *Nucleic Acids Research* 39.Database (Nov. 2010), pp. D19–D21. DOI: 10.1093/nar/gkq1019.

[263] Y. Kodama, M. Shumway, and R. Leinonen and. "The sequence read archive: explosive growth of sequencing data". In: *Nucleic Acids Research* 40.D1 (Oct. 2011), pp. D54–D56. DOI: 10.1093/nar/gkr854.

[264] Yanqing Wang et al. "GSA: Genome Sequence Archive". In: *Genomics, Proteomics & Bioinformatics* 15.1 (Feb. 2017), pp. 14–18. DOI: 10.1016/j.gpb.2017.01.001.

[265] Yunhai Luo et al. "New developments on the Encyclopedia of DNA Elements (ENCODE) data portal". In: *Nucleic Acids Research* 48.D1 (Nov. 2019), pp. D882–D889. DOI: 10.1093/nar/gkz1062.

[266] Saket Choudhary. "pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive". In: *F1000Research* 8 (Apr. 2019), p. 532. DOI: 10.12688/f1000research.18676.1.

[267] Siebren Frölich et al. "genomepy: genes and genomes at your fingertips". In: *Bioinformatics* 39.3 (Mar. 2023). Ed. by Tobias Marschall. DOI: 10.1093/bioinformatics/btad119.

[268] Felix Krueger et al. *TrimGalore*. 2023. DOI: 10.5281/ZENODO.5127898.

[269] Shifu Chen et al. "fastp: an ultra-fast all-in-one FASTQ preprocessor". In: *Bioinformatics* 34.17 (Sept. 2018), pp. i884–i890. DOI: 10.1093/bioinformatics/bty560.

[270] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (Mar. 2012), pp. 357–359. DOI: 10.1038/nmeth.1923.

[271] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv* (2013). DOI: 10.48550/ARXIV.1303.3997.

[272] Md. Vasimuddin et al. "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems". In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, May 2019. DOI: 10.1109/ipdps.2019.00041.

[273] Daehwan Kim et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 907–915. DOI: 10.1038/s41587-019-0201-4.

[274] Heng Li. "New strategies to improve minimap2 alignment accuracy". In: *Bioinformatics* 37.23 (Oct. 2021). Ed. by Can Alkan, pp. 4572–4574. DOI: 10.1093/bioinformatics/btab705.

[275] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (Oct. 2012), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[276] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: *Scientific Reports* 9.1 (June 2019). DOI: 10.1038/s41598-019-45839-z.

[277] W. James Kent et al. "The Human Genome Browser at UCSC". en. In: *Genome Research* 12.6 (June 2002). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 996–1006. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.229102.

[278] Simon Andrews et al. *FastQC: a quality control tool for high throughput sequence data*. 2010.

[279] H. Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (June 2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[280] Broad Institute. *Picard Tools*. http://broadinstitute.github.io/picard/. n.d.

[281] Fidel Ramirez et al. "deepTools: a flexible platform for exploring deep-sequencing data". In: *Nucleic Acids Research* 42.W1 (May 2014), W187–W191. DOI: 10.1093/nar/gku365.

[282] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (June 2016), pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354.

[283] Feng Yan et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology* 21.1 (Feb. 2020). DOI: 10.1186/s13059-020-1929-3.

[284] John M. Gaspar. *Genrich: detecting sites of genomic enrichment*. https://github.com/jsh58/Genrich. n.d.

[285] Qunhua Li et al. "Measuring reproducibility of high-throughput experiments". In: *The Annals of Applied Statistics* 5.3 (Sept. 2011). DOI: 10.1214/11-aoas466.

[286] Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization". In: *Bioinformatics* 31.14 (Mar. 2015), pp. 2382–2383. DOI: 10.1093/bioinformatics/btv145.

[287] Yang Liao, Gordon K. Smyth, and Wei Shi. "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote". In: *Nucleic Acids Research* 41.10 (Apr. 2013), e108–e108. DOI: 10.1093/nar/gkt214.

[288] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (Sept. 2014), pp. 166–169. DOI: 10.1093/bioinformatics/btu638.

[289] Liguo Wang, Shengqin Wang, and Wei Li. "RSeQC: quality control of RNA-seq experiments". In: *Bioinformatics* 28.16 (June 2012), pp. 2184–2185. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts356. eprint: https://academic.oup.com/bioinformatics/article-pdf/28/16/2184/48870333/bioinformatics_28_16_2184.pdf.

[290] Rob Patro et al. "Salmon provides fast and bias-aware quantification of transcript expression". en. In: *Nature Methods* 14.4 (Apr. 2017). Number: 4 Publisher: Nature Publishing Group, pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197.

[291] Siebren Frölich and Simon van Heeringen. *pytxi - Transcript-level to gene-level quantification*. https://github.com/vanheeringen-lab/pytxi. n.d.

[292] Michael I. Love et al. "Tximeta: reference sequence checksums for provenance identification in RNA-seq". In: *bioRxiv* (2019). DOI: 10.1101/777888. eprint: https://www.biorxiv.org/content/early/2019/09/25/777888.full.pdf.

[293] Simon Anders, Alejandro Reyes, and Wolfgang Huber. *DEXSeq: Inference of differential exon usage in RNA-Seq*. 2023. DOI: 10.18129/B9.bioc.DEXSeq.

[294] Sergi Sayols, Denise Scherzinger, and Holger Klein. "dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data". In: *BMC Bioinformatics* 17.1 (Oct. 2016), p. 428. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1276-2.

[295] Hongbo Yang et al. "A map of cis-regulatory elements and 3D genome structures in zebrafish". In: *Nature* 588.7837 (Nov. 2020), pp. 337–343. DOI: 10.1038/s41586-020-2962-9.

[296] Eduardo Soares and Huiqing Zhou. "Master regulatory role of p63 in epidermal development and disease". In: *Cellular and Molecular Life Sciences* 75.7 (Nov. 2017), pp. 1179–1190. DOI: 10.1007/s00018-017-2701-z.

[297] Lingjie Li et al. "TFAP2C- and p63-Dependent Networks Sequentially Rearrange Chromatin Landscapes to Drive Human Epidermal Lineage Commitment". In: *Cell Stem Cell* 24.2 (Feb. 2019), 271–284.e8. DOI: 10.1016/j.stem.2018.12.012.

[298] Andrew R. Gehrke et al. "Acoel genome reveals the regulatory landscape of whole-body regeneration". In: *Science* 363.6432 (Mar. 2019). DOI: 10.1126/science.aau6173.

[299] Wouter Meuleman et al. "Index and biological spectrum of human DNase I hypersensitive sites". In: *Nature* 584.7820 (July 2020), pp. 244–251. DOI: 10.1038/s41586-020-2559-3.

[300] Michael K. Richardson. "Heterochrony and the Phylotypic Period". In: *Developmental Biology* 172.2 (Dec. 1995), pp. 412–421. ISSN: 0012-1606. DOI: 10.1006/dbio.1995.8041.

[301] Alex T. Kalinka and Pavel Tomancak. "The evolution of early animal embryos: conservation or divergence?" In: *Trends in Ecology and Evolution* 27.7 (July 2012), pp. 385–393. DOI: 10.1016/j.tree.2012.03.007.

[302] N. Irie and S. Kuratani. "The developmental hourglass model: a predictor of the basic body plan?" In: *Development* 141.24 (Dec. 2014), pp. 4649–4655. DOI: 10.1242/dev.107318.

[303] Hajk-Georg Drost et al. "Cross-kingdom comparison of the developmental hourglass". In: *Current Opinion in Genetics and Development* 45 (Aug. 2017), pp. 69–75. DOI: 10.1016/j.gde.2017.03.003.

[304] A. Aristotle and A. L. Peck. *Generation of animals*. Harvard University Press, 1943.

[305] Paul Ehrlich and Dennis R. Parnell. *The Process of Evolution*. McGraw-Hill Companies, 1974.

[306] Wilhelm His. *Unsere Körperform und das physiologische Problem ihrer Entstehung: Briefe an einen befreundeten Naturforscher*. Vogel, Leipzig, 1875.

[307] Marcel Quint et al. "A transcriptomic hourglass in plant embryogenesis". In: *Nature* 490.7418 (Sept. 2012), pp. 98–101. DOI: 10.1038/nature11394.

[308] Xuanjin Cheng et al. "A "Developmental Hourglass" in Fungi". In: *Molecular Biology and Evolution* 32.6 (Mar. 2015), pp. 1556–1566. DOI: 10.1093/molbev/msv047.

[309] Jack Cohen. *Living embryos*. Oxford Pergamon, 1963.

[310] F. Seidel. *Körpergrundgestalt und Keimstruktur. Eine Erörterung über die Grundlagen der vergleichenden und experimentellen Embryologie und deren Gültigkeit bei phylogenetischen Berlegungen*. 1960.

[311] Yui Uchida et al. “Embryonic lethality is not sufficient to explain hourglass-like conservation of vertebrate embryos”. In: *EvoDevo* 9.1 (Mar. 2018). DOI: 10.1186/s13227-018-0095-0.

[312] Jialin Liu et al. “Inter-embryo gene expression variability recapitulates the hourglass pattern of evo-devo”. In: *BMC Biology* 18.1 (Sept. 2020). DOI: 10.1186/s12915-020-00842-z.

[313] Yui Uchida et al. “Potential contribution of intrinsic developmental stability toward body plan conservation”. In: *BMC Biology* 20.1 (Apr. 2022). DOI: 10.1186/s12915-022-01276-5.

[314] Alberto Perez-Posada et al. “Insights into deuterostome evolution from the biphasic transcriptional programmes of hemichordates”. In: *bioRxiv* (June 2022). DOI: 10.1101/2022.06.10.495707.

[315] Jason Cheok Kuan Leong et al. “Derivedness Index for Estimating Degree of Phenotypic Evolution of Embryos: A Study of Comparative Transcriptomic Analyses of Chordates and Echinoderms”. In: *Frontiers in Cell and Developmental Biology* 9 (Nov. 2021). DOI: 10.3389/fcell.2021.749963.

[316] Haiyang Hu et al. “Constrained vertebrate evolution by pleiotropic genes”. In: *Nature Ecology and Evolution* 1.11 (Sept. 2017), pp. 1722–1730. DOI: 10.1038/s41559-017-0318-0.

[317] Gerardo A. Cordero, Marcelo R. Sánchez-Villagra, and Ingmar Werneburg. “An irregular hourglass pattern describes the tempo of phenotypic development in placental mammal evolution”. In: *Biology Letters* 16.5 (May 2020), p. 20200087. DOI: 10.1098/rsbl.2020.0087.

[318] Bininda-Emonds Olaf R. P., Jeffery Jonathan E., and Michael K. Richardson. “Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.1513 (Feb. 2003), pp. 341–346. DOI: 10.1098/rspb.2002.2242.

[319] Megan E Chan et al. “Comparative Transcriptomics Reveals Distinct Patterns of Gene Expression Conservation through Vertebrate Embryogenesis”. In: *Genome Biology and Evolution* 13.8 (July 2021). Ed. by Mar Alba. DOI: 10.1093/gbe/evab160.

[320] Casey W. Dunn et al. “Pairwise comparisons across species are problematic when analyzing functional genomic data”. In: *Proceedings of the National Academy of Sciences* 115.3 (Jan. 2018). DOI: 10.1073/pnas.1707515115.

[321] Richard J White et al. “A high-resolution mRNA expression time course of embryonic development in zebrafish”. In: *eLife* 6 (Nov. 2017). DOI: 10.7554/elife.30860.

[322] Andreas Hejnol and Casey W. Dunn. “Animal Evolution: Are Phyla Real?” In: *Current Biology* 26.10 (2016), R424–R426. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2016.03.058.

[323] Edoardo Saccenti. “What can go wrong when observations are not independently and identically distributed: A cautionary note on calculating correlations on combined data sets from different experiments or conditions”. In: *Frontiers in Systems Biology* 3 (Jan. 2023). DOI: 10.3389/fsysb.2023.1042156.

[324] Stephen G. Landt et al. “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia”. In: *Genome Research* 22.9 (Sept. 2012), pp. 1813–1831. DOI: 10.1101/gr.136184.111.

[325] Sean Whalen et al. “Navigating the pitfalls of applying machine learning in genomics”. In: *Nature Reviews Genetics* 23.3 (Nov. 2021), pp. 169–181. DOI: 10.1038/s41576-021-00434-9.

[326] Olaf R. P. Bininda-Emonds et al. “From Haeckel to event-pairing: the evolution of developmental sequences”. In: *Theory in Biosciences* 121.3 (Nov. 2002), pp. 297–320. DOI: 10.1007/s12064-002-0016-5.

[327] Itai Yanai et al. “Mapping Gene Expression in Two Xenopus Species: Evolutionary Constraints and Developmental Flexibility”. In: *Developmental Cell* 20.4 (Apr. 2011), pp. 483–496. DOI: 10.1016/j.devcel.2011.03.015.

[328] Michal Levin et al. “Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo”. In: *Developmental Cell* 22.5 (May 2012), pp. 1101–1108. DOI: 10.1016/j.devcel.2012.04.004.

[329] Suraj Kannan et al. “Transcriptomic entropy benchmarks stem cell-derived cardiomyocyte maturation against endogenous tissue at single cell level”. In: *PLOS Computational Biology* 17.9 (Sept. 2021). Ed. by Jeffrey J. Saucerman, e1009305. DOI: 10.1371/journal.pcbi.1009305.

[330] Yun-Kyo Kim et al. “Absolute scaling of single-cell transcriptomes identifies pervasive hypertranscription in adult stem and progenitor cells”. In: *Cell Reports* 42.1 (Jan. 2023), p. 111978. DOI: 10.1016/j.celrep.2022.111978.

[331] Michelle Percharde, Aydan Bulut-Karslioglu, and Miguel Ramalho-Santos. “Hypertranscription in Development, Stem Cells, and Regeneration”. In: *Developmental Cell* 40.1 (Jan. 2017), pp. 9–21. DOI: 10.1016/j.devcel.2016.11.010.

[332] Assaf Malik et al. “Parallel embryonic transcriptional programs evolve under distinct constraints and may enable morphological conservation amidst adaptation”. In: *Developmental Biology* 430.1 (Oct. 2017), pp. 202–213. DOI: 10.1016/j.ydbio.2017.07.019.

[333] Tsvia Gildor et al. “Developmental transcriptomes of the sea star, Patiria miniata, illuminate how gene expression changes with evolutionary distance”. In: *Scientific Reports* 9.1 (Nov. 2019). DOI: 10.1038/s41598-019-52577-9.

[334] Koh Onimaru et al. “Developmental hourglass and heterochronic shifts in fin and limb development”. In: *eLife* 10 (Feb. 2021). DOI: 10.7554/elife.62865.

[335] Nick D.L. Owens et al. “Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development”. In: *Cell Reports* 14.3 (Jan. 2016), pp. 632–647. DOI: 10.1016/j.celrep.2015.12.050.

[336] Harel Zalts and Itai Yanai. “Developmental constraints shape the evolution of the nematode mid-developmental transition”. In: *Nature Ecology and Evolution* 1.5 (Mar. 2017). DOI: 10.1038/s41559-017-0113.

[337] Maarten van der Sande et al. “Seq2science: an end-to-end workflow for functional genomics analysis”. In: *PeerJ* 11 (Nov. 2023), e16380. ISSN: 2167-8359. DOI: 10.7717/peerj.16380.

[338] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (Oct. 2012), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[339] Petr Danecek et al. “Twelve years of SAMtools and BCFtools”. In: *GigaScience* 10.2 (Jan. 2021). DOI: 10.1093/gigascience/giab008.

[340] Geo Pertea and Mihaela Pertea. “GFF Utilities: GffRead and GffCompare”. In: *F1000Research* 9 (Sept. 2020), p. 304. DOI: 10.12688/f1000research.23297.2.

[341] David M. Emms and Steven Kelly. “OrthoFinder: phylogenetic orthology inference for comparative genomics”. In: *Genome Biology* 20.1 (Nov. 2019). DOI: 10.1186/s13059-019-1832-y.

[342] Maarten Van Der Sande and Simon Van Heeringen. *Qnorm: fast-ish (and correct!) quantile normalization in Python*. 2021. DOI: 10.5281/ZENODO.5546255.

[343] Michael R. Crusoe et al. “The khmer software package: enabling efficient nucleotide sequence analysis”. In: *F1000Research* 4 (Sept. 2015), p. 900. DOI: 10.12688/f1000research.6924.1.

[344] Maria Tsompana and Michael J Buck. “Chromatin accessibility: a window into the genome”. In: *Epigenetics and Chromatin* 7.1 (Nov. 2014). ISSN: 1756-8935. DOI: 10.1186/1756-8935-7-33.

[345] Jonas Simon Fleck et al. "Inferring and perturbing cell fate regulomes in human brain organoids". In: *Nature* 621.7978 (Oct. 2022), pp. 365–372. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05279-8.

[346] Kenji Kamimoto et al. "Dissecting cell identity via network inference and in silico gene perturbation". In: *Nature* 614.7949 (Feb. 2023), pp. 742–751. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05688-9.

[347] Vinay K. Kartha et al. "Functional inference of gene regulation using single-cell multi-omics". In: *Cell Genomics* 2.9 (Sept. 2022), p. 100166. ISSN: 2666-979X. DOI: 10.1016/j.xgen.2022.100166.

[348] Maarten Van Der Sande and Simon Van Heeringen. *Fast-ish (and correct!) quantile normalization in Python*. 2021. DOI: 10.5281/ZENODO.5546255.

[349] Nikolaus Fortelny et al. "Can we predict protein from mRNA levels?" In: *Nature* 547.7664 (July 2017), E19–E20. DOI: 10.1038/nature22293.

[350] Alexander Franks, Edoardo Airoldi, and Nikolai Slavov. "Post-transcriptional regulation across human tissues". In: *PLOS Computational Biology* 13.5 (May 2017). Ed. by Christine Vogel, e1005535. DOI: 10.1371/journal.pcbi.1005535.

[351] Zhijian Li et al. "Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen". In: *Nature Communications* 12.1 (Nov. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-26530-2.

[352] Su Wang et al. "Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles". In: *Genome Research* 26.10 (July 2016), pp. 1417–1429. DOI: 10.1101/gr.201574.115.

[353] Mar González-Ramírez et al. "Differential contribution to gene expression prediction of histone modifications at enhancers or promoters". In: *PLOS Computational Biology* 17.9 (Sept. 2021). Ed. by Chongzhi Zang, e1009368. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009368.

[354] Kazumasa Kanemaru et al. "Spatially resolved multiomics of human cardiac niches". In: *Nature* 619.7971 (July 2023), pp. 801–810. DOI: 10.1038/s41586-023-06311-1.

[355] Meenakshi S. Kagda et al. "Data navigation on the ENCODE portal". In: (2023). DOI: 10.48550/ARXIV.2305.00006.

[356] The FANTOM Consortium & Riken Omics Science Center. "The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line". In: *Nature Genetics* 41.5 (Apr. 2009), pp. 553–562. ISSN: 1546-1718. DOI: 10.1038/ng.375.

[357] Piotr J. Balwierz et al. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs". In: *Genome Research* 24.5 (Feb. 2014), pp. 869–884. DOI: 10.1101/gr.169508.113.

[358] Jeanne Chèneby et al. "ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments". In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D267–D275. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1092.

[359] Stryder M. Meadows, Matthew C. Salanga, and Paul A. Krieg. "Krüppel-like factor 2 cooperates with the ETS family protein ERG to activateFlk1expression during vascular development". In: *Development* 136.7 (Apr. 2009), pp. 1115–1125. ISSN: 0950-1991. DOI: 10.1242/dev.029538.

[360] Hao Zhao et al. "FLI1 and PKC co-activation promote highly efficient differentiation of human embryonic stem cells into endothelial-like cells". In: *Cell Death & Disease* 9.2 (Jan. 2018). ISSN: 2041-4889. DOI: 10.1038/s41419-017-0162-9.

[361] Yaacov Ben-David et al. "Current insights into the role of Fli-1 in hematopoiesis and malignant transformation". In: *Cellular and Molecular Life Sciences* 79.3 (Feb. 2022). ISSN: 1420-9071. DOI: 10.1007/s00018-022-04160-1.

[362] J.D. Steimle and I.P. Moskowitz. "TBX5". In: *T-box Genes in Development*. Elsevier, 2017, pp. 195–221. ISBN: 9780128013809. DOI: 10.1016/bs.ctdb.2016.08.008.

[363] Panagiota Siatra et al. "Return of the Tbx5; lineage-tracing reveals ventricular cardiomyocyte-like precursors in the injured adult mammalian heart". In: *npj Regenerative Medicine* 8.1 (Mar. 2023). ISSN: 2057-3995. DOI: 10.1038/s41536-023-00280-9.

[364] Hiroki Kokubo et al. "Hesr1 and Hesr2 regulate atrioventricular boundary formation in the developing heart through the repression of Tbx2". In: *Development* 134.4 (Feb. 2007), pp. 747–755. ISSN: 0950-1991. DOI: 10.1242/dev.02777.

[365] Xinran Ma et al. "Deciphering the Roles of PPARγ in Adipocytes via Dynamic Change of Transcription Complex". In: *Frontiers in Endocrinology* 9 (Aug. 2018). ISSN: 1664-2392. DOI: 10.3389/fendo.2018.00473.

[366] Seung-Hun Lee et al. "Notch Pathway Targets Proangiogenic Regulator Sox17 to Restrict Angiogenesis". In: *Circulation Research* 115.2 (July 2014), pp. 215–226. ISSN: 1524-4571. DOI: 10.1161/circresaha.115.303142.

[367] William Schachterle et al. "Sox17 drives functional engraftment of endothelium converted from non-vascular cells". In: *Nature Communications* 8.1 (Jan. 2017). ISSN: 2041-1723. DOI: 10.1038/ncomms13963.

[368] Raman Sood, Yasuhiko Kamikubo, and Paul Liu. "Role of RUNX1 in hematological malignancies". In: *Blood* 129.15 (Apr. 2017), pp. 2070–2082. ISSN: 1528-0020. DOI: 10.1182/blood-2016-10-687830.

[369] H. Hoshino and K. Igarashi. "Expression of the Oxidative Stress-Regulated Transcription Factor Bach2 in Differentiating Neuronal Cells". In: *Journal of Biochemistry* 132.3 (Sept. 2002), pp. 427–431. ISSN: 0021-924X. DOI: 10.1093/oxfordjournals.jbchem.a003239.

[370] Guo Liu and Feng Liu. "Bach2: A Key Regulator in Th2-Related Immune Cells and Th2 Immune Response". In: *Journal of Immunology Research* 2022 (Mar. 2022). Ed. by Huaxi Xu, pp. 1–10. ISSN: 2314-8861. DOI: 10.1155/2022/2814510.

[371] Ari Itoh-Nakadai et al. "The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program". In: *Nature Immunology* 15.12 (Oct. 2014), pp. 1171–1180. ISSN: 1529-2916. DOI: 10.1038/ni.3024.

[372] Kyoko Ochiai and Kazuhiko Igarashi. "Exploring novel functions of BACH2 in the acquisition of antigen-specific antibodies". In: *International Immunology* 35.6 (Dec. 2022), pp. 257–265. ISSN: 1460-2377. DOI: 10.1093/intimm/dxac065.

[373] Brian Hie et al. "Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape". In: *Cell Systems* 8.6 (June 2019), 483–493.e7. ISSN: 2405-4712. DOI: 10.1016/j.cels.2019.05.003.

[374] Edward E. Morrisey et al. "GATA-6: A Zinc Finger Transcription Factor That Is Expressed in Multiple Cell Lineages Derived from Lateral Mesoderm". In: *Developmental Biology* 177.1 (July 1996), pp. 309–322. ISSN: 0012-1606. DOI: 10.1006/dbio.1996.0165.

[375] Mengyi Song et al. "GATA4/5/6 family transcription factors are conserved determinants of cardiac versus pharyngeal mesoderm fate". In: *Science Advances* 8.10 (Mar. 2022). ISSN: 2375-2548. DOI: 10.1126/sciadv.abg0834.

[376] Anna Rita Migliaccio et al. "GATA-1 as a Regulator of Mast Cell Differentiation Revealed by the Phenotype of the GATA-1low Mouse Mutant". In: *The Journal of Experimental Medicine* 197.3 (Feb. 2003), pp. 281–296. ISSN: 0022-1007. DOI: 10.1084/jem.20021149.

[377] Juehua Gao, Yi-Hua Chen, and LoAnn C. Peterson. "GATA family transcriptional factors: emerging suspects in hematologic disorders". In: *Experimental Hematology & Oncology* 4.1 (Oct. 2015). ISSN: 2162-3619. DOI: 10.1186/s40164-015-0024-z.

[378] Xingqiang Lai et al. "SOX10 ablation severely impairs the generation of postmigratory neural crest from human pluripotent stem cells". In: *Cell Death & Disease* 12.9 (Aug. 2021). ISSN: 2041-4889. DOI: 10.1038/s41419-021-04099-4.

[379] René Marke, Frank N. van Leeuwen, and Blanca Scheijen. "The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia". In: *Haematologica* 103.4 (Mar. 2018), pp. 565–574. ISSN: 1592-8721. DOI: 10.3324/haematol.2017.185603.

[380] San-Pin Wu et al. "Tbx18 regulates development of the epicardium and coronary vessels". In: *Developmental Biology* 383.2 (Nov. 2013), pp. 307–320. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2013.08.019.

[381] Nicole R. Stone et al. "Context-Specific Transcription Factor Functions Regulate Epigenomic and Transcriptional Dynamics during Cardiac Reprogramming". In: *Cell Stem Cell* 25.1 (July 2019), 87–102.e9. ISSN: 1934-5909. DOI: 10.1016/j.stem.2019.06.012.

[382] Masaki Ieda et al. "Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors". In: *Cell* 142.3 (Aug. 2010), pp. 375–386. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.07.002.

[383] Cody Desjardins and Francisco Naya. "The Function of the MEF2 Family of Transcription Factors in Cardiac Development, Cardiogenomics, and Direct Reprogramming". In: *Journal of Cardiovascular Development and Disease* 3.3 (Aug. 2016), p. 26. ISSN: 2308-3425. DOI: 10.3390/jcdd3030026.

[384] Hao Li et al. "Transcription factor MEF2C influences neural stem/progenitor cell differentiation and maturation in vivo". In: *Proceedings of the National Academy of Sciences* 105.27 (July 2008), pp. 9397–9402. ISSN: 1091-6490. DOI: 10.1073/pnas.0802876105.

[385] Bangmin Zhu and Tod Gulick. "Phosphorylation and Alternative Pre-mRNA Splicing Converge To Regulate Myocyte Enhancer Factor 2C Activity". In: *Molecular and Cellular Biology* 24.18 (Sept. 2004), pp. 8264–8275. ISSN: 1098-5549. DOI: 10.1128/mcb.24.18.8264-8275.2004.

[386] Brian L. Black and Eric N. Olson. "TRANSCRIPTIONAL CONTROL OF MUSCLE DEVELOPMENT BY MYOCYTE ENHANCER FACTOR-2 (MEF2) PROTEINS". In: *Annual Review of Cell and Developmental Biology* 14.1 (Nov. 1998), pp. 167–196. ISSN: 1530-8995. DOI: 10.1146/annurev.cellbio.14.1.167.

[387] Anna Hendrika Cornelia Vlot, Setareh Maghsudi, and Uwe Ohler. "Cluster-independent marker feature identification from single-cell omics data using SEMITONES". In: *Nucleic Acids Research* 50.18 (July 2022), e107–e107. ISSN: 1362-4962. DOI: 10.1093/nar/gkac639.

[388] Silvia Domcke et al. "A human cell atlas of fetal chromatin accessibility". In: *Science* 370.6518 (Nov. 2020). ISSN: 1095-9203. DOI: 10.1126/science.aba7612.

[389] Darren A. Cusanovich et al. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility". In: *Cell* 174.5 (Aug. 2018), 1309–1324.e18. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.06.052.

[390] Connie W. Tsao et al. "Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association". In: *Circulation* 147.8 (Feb. 2023). ISSN: 1524-4539. DOI: 10.1161/cir.0000000000001123.

[391] Kazumasa Kanemaru. *Heart Cell Atlas v2: Heart Global H5AD (log-normalised)*. https://cellgeni.cog.sanger.ac.uk/heartcellatlas/v2/Global_lognormalised.h5ad. [Online; accessed March-2023]. 2023.

[392] Kazumasa Kanemaru. *Heart Cell Atlas v2: Heart Global ATAC Peak Matrix*. https://cellgeni.cog.sanger.ac.uk/heartcellatlas/v2/Global_lognormalised.h5ad. [Online; accessed August-2023]. 2023.

[393] Shinya Oki et al. "Ch <scp>IP</scp> -Atlas: a data-mining suite powered by full integration of public Ch <scp>IP</scp> -seq data". In: *EMBO reports* 19.12 (Nov. 2018). ISSN: 1469-3178. DOI: 10.15252/embr.201846255.

[394] Zhaonan Zou et al. "ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data". In: *Nucleic Acids Research* 50.W1 (Mar. 2022), W175–W182. ISSN: 1362-4962. DOI: 10.1093/nar/gkac199.

[395] S Oki and T Ohta. *ChIP-Atlas*. https://chip-atlas.dbcls.jp/data/hg38/target/BACH2.1.tsv. [Online; accessed December-2023]. 2023.

[396] Hayley M. Bennett et al. "Single-cell proteomics enabled by next-generation sequencing or mass spectrometry". In: *Nature Methods* 20.3 (Mar. 2023), pp. 363–374. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01791-5.

[397] Wouter Saelens et al. "A comparison of single-cell trajectory inference methods". In: *Nature Biotechnology* 37.5 (Apr. 2019), pp. 547–554. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0071-9.

[398] Hoa Thi Nhu Tran et al. "A benchmark of batch-effect correction methods for single-cell RNA sequencing data". In: *Genome Biology* 21.1 (Jan. 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1850-9.

[399] Laleh Haghverdi et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature Biotechnology* 36.5 (Apr. 2018), pp. 421–427. ISSN: 1546-1696. DOI: 10.1038/nbt.4091.

[400] David Lähnemann et al. "Eleven grand challenges in single-cell data science". In: *Genome Biology* 21.1 (Feb. 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-020-1926-6.

[401] Gerard A. Bouland, Ahmed Mahfouz, and Marcel J. T. Reinders. "Consequences and opportunities arising due to sparser single-cell RNA-seq datasets". In: *Genome Biology* 24.1 (Apr. 2023). ISSN: 1474-760X. DOI: 10.1186/s13059-023-02933-w.

[402] Jacob Schreiber et al. "Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome". In: *Genome Biology* 21.1 (Mar. 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-020-01977-6.

[403] Kasia Z. Kedzierska et al. "Assessing the limits of zero-shot foundation models in single-cell biology". In: (Oct. 2023). DOI: 10.1101/2023.10.16.561085.

[404] Xin Dong et al. "Single-cell gene regulation network inference by large-scale data integration". In: *Nucleic Acids Research* 50.21 (Sept. 2022), e126–e126. ISSN: 1362-4962. DOI: 10.1093/nar/gkac819.

[405] Andreas R. Gschwind et al. "An encyclopedia of enhancer-gene regulatory interactions in the human genome". In: (Nov. 2023). DOI: 10.1101/2023.11.09.563812.

[406] Yuri Lazebnik. "Can a biologist fix a radio?—Or, what I learned while studying apoptosis". In: *Cancer Cell* 2.3 (Sept. 2002), pp. 179–182. DOI: 10.1016/s1535-6108(02)00133-2.

[407] Dennis Bray. "Reasoning for results". In: *Nature* 412.6850 (Aug. 2001), pp. 863–863. ISSN: 1476-4687. DOI: 10.1038/35091132.

[408] Florian Markowetz. "All biology is computational biology". In: *PLOS Biology* 15.3 (Mar. 2017), e2002050. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2002050.

[409] International Electrotechnical Commission. *IEC 60617 - Graphical Symbols for Diagrams*. https://std.iec.ch/iec60617. 2017.

[410] William J. R. Longabaugh. "BioTapestry: A Tool to Visualize the Dynamic Properties of Gene Regulatory Networks". In: *Gene Regulatory Networks*. Humana Press, Aug. 2011, pp. 359–394. ISBN: 9781617792922. DOI: 10.1007/978-1-61779-292-2_21.

[411] David M. Rocke and Blythe Durbin. "A Model for Measurement Error for Gene Expression Arrays". In: *Journal of Computational Biology* 8.6 (Nov. 2001), pp. 557–569. DOI: 10.1089/106652701753307485.

[412] Longjun Wu, Kailey E Ferger, and J David Lambert. "Gene Expression Does Not Support the Developmental Hourglass Model in Three Animals with Spiralian Development". In: *Molecular Biology and Evolution* 36.7 (Mar. 2019). Ed. by Ilya Ruvinsky, pp. 1373–1383. DOI: 10.1093/molbev/msz065.

[413] Marc Robinson-Rechavi. *Story behind the paper: The hourglass and the early conservation models – co-existing evolutionary patterns in vertebrate development*. https://web.archive.org/web/20160416153603/https://people.unil.ch/marcrobinsonrechavi/2013/07/story-behind-the-paper-the-hourglass-and-the-early-conservation-models-co-existing-evolutionary-patterns-in-vertebrate-development/. 2013.

[414] Caveman A. "When it is time to have an old dogma put down". In: *Journal of Cell Science* 113.14 (July 2000), pp. 2517–2518. DOI: 10.1242/jcs.113.14.2517.

[415] GRAHAM E. BUDD and SÖREN JENSEN. "A critical reappraisal of the fossil record of the bilaterian phyla". In: *Biological Reviews of the Cambridge Philosophical Society* 75.2 (May 2000), pp. 253–295. DOI: 10.1017/s000632310000548x.

[416] G Scholtz. "Bauplane versus groundpatterns, phyla versus monophyla: aspects of patterns and processes in evolutionary developmental biology". In: *Crustacean Issues* 15 (2004), pp. 3–18.

[417] Michael K Richardson et al. "Phylotypic stage theory". In: *Trends in Ecology and Evolution* 13.4 (Apr. 1998), p. 158. ISSN: 0169-5347. DOI: 10.1016/s0169-5347(98)01340-8.

[418] William W. Ballard. "Morphogenetic Movements and Fate Maps of Vertebrates". In: *American Zoologist* 21.2 (May 1981), pp. 391–399. DOI: 10.1093/icb/21.2.391.

[419] Pere Alberch. "Wolpert, L. The Triumph of the Embryo. Oxford University Press (1991), 211 pp. ISBN: 0-19-854243-7". In: *Journal of Evolutionary Biology* 6.3 (May 1993), pp. 457–459. DOI: 10.1046/j.1420-9101.1993.6030457.x.

[420] J. M. W. Slack, P. W. H. Holland, and C. F. Graham. "The zootype and the phylotypic stage". In: *Nature* 361.6412 (Feb. 1993), pp. 490–492. DOI: 10.1038/361490a0.

[421] Katalin Ferenc et al. "Empirical study on software and process quality in bioinformatics tools". In: (Mar. 2022). DOI: 10.1101/2022.03.10.483804.

[422] Sarah Killcoyne and John Boyle. "Managing Chaos: Lessons Learned Developing Software in the Life Sciences". In: *Computing in Science & Engineering* 11.6 (Nov. 2009), pp. 20–29. ISSN: 1521-9615. DOI: 10.1109/mcse.2009.198.

[423] E W Dijkstra. *E.W. Dijkstra Archive: Why numbering should start at zero (EWD 831) — cs.utexas.edu.* https://www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD831.html. [Accessed 22-12-2023]. 1982.

[424] Lincoln Stein. *Generic Feature Format Version 3 (GFF3).* Feb. 2013.

[425] Lior Pachter. *Is there a citation for GTF?* Oct. 2023.

[426] David J. Lipman and William R. Pearson. "Rapid and Sensitive Protein Similarity Searches". In: *Science* 227.4693 (Mar. 1985), pp. 1435–1441. DOI: 10.1126/science.2983426.

[427] Heng Li, Xiaowen Feng, and Chong Chu. "The design and construction of reference pangenome graphs with minigraph". In: *Genome Biology* 21.1 (Oct. 2020). DOI: 10.1186/s13059-020-02168-z.

[428] Zheng Zuo and Gary D Stormo. "High-Resolution Specificity from DNA Sequencing Highlights Alternative Modes of Lac Repressor Binding". In: *Genetics* 198.3 (Sept. 2014), pp. 1329–1343. ISSN: 1943-2631. DOI: 10.1534/genetics.114.170100.

[429] Arttu Jolma et al. "DNA-dependent formation of transcription factor pairs alters their binding specificity". In: *Nature* 527.7578 (Nov. 2015), pp. 384–388. ISSN: 1476-4687. DOI: 10.1038/nature15518.

[430] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. "Transcription factor–DNA binding: beyond binding site motifs". In: *Current Opinion in Genetics & Development* 43 (Apr. 2017), pp. 110–119. ISSN: 0959-437X. DOI: 10.1016/j.gde.2017.02.007.

[431] Carmen Bravo González-Blas et al. "SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks". In: *Nature Methods* 20.9 (July 2023), pp. 1355–1367. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01938-4.

[432] Matthias Siebert and Johannes Söding. "Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences". In: *Nucleic Acids Research* 44.13 (June 2016), pp. 6055–6069. ISSN: 1362-4962. DOI: 10.1093/nar/gkw521.

[433] Gherman Novakovsky et al. "ExplaiNN: interpretable and transparent neural networks for genomics". In: *Genome Biology* 24.1 (June 2023). ISSN: 1474-760X. DOI: 10.1186/s13059-023-02985-y.

[434] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". In: (2017). DOI: 10.48550/ARXIV.1704.02685.

[435] Sergey Aganezov et al. "A complete reference genome improves analysis of human genetic variation". In: *Science* 376.6588 (Apr. 2022). DOI: 10.1126/science.abl3533.

[436] Bohu Pan et al. "Similarities and differences between variants called with human reference genome HG19 or HG38". In: *BMC Bioinformatics* 20.S2 (Mar. 2019). ISSN: 1471-2105. DOI: 10.1186/s12859-019-2620-0.

[437] Yan Guo et al. "Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis". In: *Genomics* 109.2 (Mar. 2017), pp. 83–90. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2017.01.005.

[438] John M. Gaspar. "Improved peak-calling with MACS2". In: (Dec. 2018). DOI: 10.1101/496521.

[439] Nathan Watson-Haigh. *satijalab/Seurat - issue #6654 - Incorrect/inconsistent fold change calculation.*

[440] *Intel® Xeon® Processor E5-2670 (20M Cache, 2.60 GHz, 8.00 GT/s Intel® QPI) - Product Specifications | Intel — intel.com.* https://www.intel.com/content/www/us/en/products/sku/64595/intel-xeon-processor-e52670-20m-cache-2-60-ghz-8-00-gts-intel-qpi/specifications.html. [Accessed 22-12-2023].

[441] *Emissiefactor elektriciteit uit fossiele bronnen.* https://www.rvo.nl/sites/default/files/2022-05/CE_Delft_210338_Emissiefactor_Elektriciteit_Fossiele_Bronnen_DEF.pdf. 2021.

[442] Matthew N Bernstein, AnHai Doan, and Colin N Dewey. "MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive". In: *Bioinformatics* 33.18 (May 2017). Ed. by Jonathan Wren, pp. 2914–2923. DOI: 10.1093/bioinformatics/btx334.

[443] Adam Klie et al. "Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition". In: *Database* 2021 (Jan. 2021). DOI: 10.1093/database/baab021.

[444] Phil Ewels. *SRA-Explorer: Web application to explore the Sequence Read Archive.* https://github.com/ewels/sra-explorer.

[445] Ángel Gálvez-Merchán et al. "Metadata retrieval from sequence databases with ffq". In: (2022).

[446] Harshil Patel et al. *nf-core/fetchngs: nf-core/fetchngs v1.10.0 - Manganese Monkey.* 2023. DOI: 10.5281/ZENODO.7941940.

[447] Renan Valieris. *parallel fastq-dump wrapper.* https://github.com/rvalieris/parallel-fastq-dump.

[448] Mariam Quiñones et al. ""METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI's sequence read archive"". In: *BMC Bioinformatics* 21.1 (Sept. 2020). DOI: 10.1186/s12859-020-03694-0.

[449] Anne E. Carpenter and Shantanu Singh. "Bringing computation to biology by bridging the last mile". In: *Nature Cell Biology* 26.1 (Jan. 2024), pp. 5–7. ISSN: 1476-4679. DOI: 10.1038/s41556-023-01286-7.

[450] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (June 2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[451] James K Bonfield et al. "HTSlib: C library for reading/writing high-throughput sequencing data". In: *GigaScience* 10.2 (Jan. 2021). ISSN: 2047-217X. DOI: 10.1093/gigascience/giab007.

[452] Barry R Zeeberg et al. In: *BMC Bioinformatics* 5.1 (2004), p. 80. DOI: 10.1186/1471-2105-5-80.

[453] Mandhri Abeysooriya et al. “Gene name errors: Lessons not learned”. In: *PLOS Computational Biology* 17.7 (July 2021). Ed. by Christos A. Ouzounis, e1008984. DOI: 10.1371/journal.pcbi.1008984.