From Department for Biosciences and Nutrition
Karolinska Institutet, Stockholm, Sweden

HUNTER-GATHERER-ANNOTATOR SCIENCE: CHARACTERIZING REGULATORY ELEMENTS IN THE GENOME OF DOG AND ZEBRAFISH WITH PUBLIC AND NOT YET PUBLIC DATA

Matthias Hörtenhuber

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-661-4
Cover illustration: Generated by Disco Diffusion 5.2 with default parameters, 250 iteration steps, seed number 1905283125, and the text prompt: "a beautiful painting of Hunter-Gatherer-Annotator Science Characterizing regulatory elements in the genome of dog and zebrafish with public and not yet public data by Man Ray, Trending on artstation."

Hunter-Gatherer-Annotator Science: Characterizing regulatory elements in the genome of dog and zebrafish with public and not yet public data

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Matthias Hörtenhuber

The thesis will be defended in public at Erna Möllersalen, ground floor, NEO, Blickagången 16, Flemingsberg, on June 3rd, 2022 at 12:30.

Principal Supervisor:
Assoc. Professor Carsten O. Daub
Karolinska Institute
Department of Biosciences and Nutrition

Co-supervisor(s):
Professor Monte Westerfield
University of Oregon
Department of Biology

Opponent:
Assoc. Professor Joachim Lütken Weischenfeldt
University of Copenhagen
Biotech Research & Innovation Centre

Examination Board:
Professor Yudi Pawitan
Karolinska Institute
Department of Medical Epidemiology and Biostatistics

Professor Sven Nelander
Uppsala University
Department of Immunology, Genetics and Pathology

Dr. Ola Larsson
Karolinska Institute
Department of Oncology-Pathology

This hand was once a fin, this hand once had claws! In my human mouth I have the pointy teeth of a wolf and the chisel teeth of a rabbit and the grinding teeth of a cow! [...]
I'm made up of the memories of my parents and my grandparents, all my ancestors. They're in the way I look, in the colour of my hair. And I'm made up of everyone I've ever met who's changed the way I think.

A Hat Full of Sky
Terry Pratchett

To the people I am made up of.

POPULAR SCIENCE SUMMARY OF THE THESIS

English version

The main topic of my Ph.D. is the regulation of the gnome in zebrafish and dogs. Every cell in an organism contains the same copy of the gnome, which is unique to each individual. It contains all the information needed to construct all the organism's proteins, which are the building blocks of biological life. The process of extracting the information from the gnome is called transcription and these transcripts are the main product I looked at in my thesis. If all cells contain the same gnome, how come a nerve cell in the brain looks and behaves so different from a skin cell or a liver cell? That is because the gnome in each cell is regulated, meaning only parts of the gnome are transcribed and only for a certain amount of time. In order to better understand the different biological processes in a cell, it is therefore very useful to know if a certain region of the gnome is actually transcribed in this cell or not.

Although humans look and behave quite differently from a small, striped aquarium fish, called zebrafish, our gnomes are actually quite similar. That's why we use them to better understand our bodies and diseases. Why don't we just study our bodies in the first place? There are a number of reason, one of the simplest being that it is a lot easier to control the environment of a zebrafish compared to humans and a controlled environment makes it a lot easier to interpret the results of experiments.

If zebrafish have such similar gnome regulation, why did I also study the dog gnome then? It turns out, zebrafish are not optimal animals to study lung diseases for example. In addition, the fact that dogs live very close to us makes it possible to study the impact of our environment on gnomes. Furthermore, because we humans have been breeding them into very distinctive dog races, dogs of the same breed have a very distinctive gnome and we can more easily see the effects of smaller changes in them.

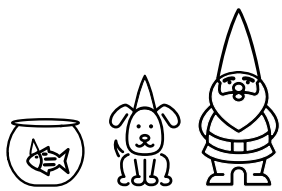
Still, even with those very nice features, the regulation of the gnome is too complex to be understood by just looking at a few cells from zebrafish or dogs. To get a good global picture, we need many, many cells. We can either hunt for these cells

by ourselves as we did in the last two studies, where we collected 116 different tissue samples from 8 different dogs. Or we use the fact that scientists usually share the data from their old experiments. That is what we did for the second study, where we gathered data from 1,802 samples from 38 different research groups to identify different regions of the zebrafish gnome, which are used to regulate the development from a fertilized egg to a full fish.

The problem with these hunter-gatherer studies is that one ends up with a lot of files and it is difficult to keep track of them and to know which file was containing the transcribed elements from the liver of a Rottweiler and which one describes the active regions of the gnome of a one-day-old zebrafish. Therefore, we also need an annotator who connects all biological information with the relevant files and makes it easy to collect, but also to share, and to analyze them. That's why I described a system in study 1 on how to do this and also created a web platform based on it, which we use for all the data in studies 2,3, and 4.

Altogether, these four studies will help scientists to handle their data better and also to understand the effects that changes in the gnome in dogs and zebrafish can have and use that knowledge for the human gnome and diseases related to it.

Oh, I think I misspelled the word genome…



German version

Jede Zelle deines Körpers ist nicht nur glücklich, sondern sie besitzt auch den gleichen Bauplan in ihrem Zellkern. Dieser Bauplan, die DNS/DNA, wird als RNA ausgelesen, die dann als Anleitung für den Bau von Proteinen dient. Proteine sind die Grundbausteine unseres Körpers. Der DNA-Abschnitt, der den Bauplan für ein bestimmtes Protein enthält, heißt (Protein-kodierendes) Gen.

Wenn alle Zellen auf dem gleichen Bauplan beruhen, weshalb schauen jedoch Nervenzellen im Gehirn anders aus als jene in der Leber und verhalten sich darüberhinaus auch noch ganz unterschiedlich? Dies wird dadurch bewerkställigt, dass nur bestimmte Gene zu bestimmten Zeitpunkten, in einer gewissen Stärke ausgelesen werden.

Interessanterweise scheinen diese Prozesse so fundamental für Lebewesen auf der Erde zu sein, dass sie nicht nur in jedem Menschen gleich funktionieren, sondern viele genauso in anderen Tieren funktionieren, je nachdem wie nahe wir zu diesen verwandt sind. Deshalb und weil genetische Experiment biologisches Material von vielen Zellen benötigen, kann man z.B. kleine Aquarienfische wie etwa Zebrafische dafür verwenden. Da diese Fische besonders gut geeignet sind um die Embryonalentwicklung besser zu verstehen, beschreiben wir in der zweiten Studie dieser Doktorarbeit u.a. bisher unbekannte Regionen in der DNA, die bei diesem Prozess aktiviert oder deaktiviert werden. Dazu haben wir Daten von 1.802 Zellproben aus 38 verschiedenen Laboren gesammelt und noch einmal neu analysiert.

Auch wenn wir Menschen überraschend viele fundamentale biologische Funktionen mit Zebrafischen teilen, ist z.B. der beste Freund des Menschen, der Hund, besser dazu geeignet um verschiedene Aspekte der Genregulation im Gehirn bei psychischen Krankheiten zu untersuchen. Deshalb stellen wir in der dritten und vierten Studie dieser Arbeit, eine Liste von bestimmten genregulativen Elementen vor, die z.B. nur in der Leber aktiv sind. Dazu haben wir mehrere hundert Proben von 88 verschiedene Geweben von 9 verschiedenen Hunden gesammelt und die darin enthaltene RNA analysiert.

Für die Analyse sowohl im Zebrafisch als auch beim Hund ist es sehr wichtig, genau zu wissen, welche Probe von welchem Tier bzw. von welchem Teil des Tieres stammt, wie die Probe weiterverarbeitet wurde und welche Datei zu welcher Probe gehört. Deshalb haben wir auch eine Datenbankstruktur entworfen,

die diese Metadaten speichert und mit den anderen Daten in Verbindung hält. Diese Struktur haben wir dann in eine Webseite implementiert, sodass Kollegen einerseits Daten von ihren Experimenten hochladen können, andererseits auch Daten von anderen herunterladen. Diese Struktur und die Webseite stellen wir in der ersten Studie dieser Arbeit vor. Eine Version dieser Plattform, die wir speziell für Zebrafischdaten eingerichtet haben, bildet den Grundstein für die zweite Studie. Eine weitere Version, jedoch spezifisch für die Daten von Hunden, wurde für die dritte und vierte Studie aufgesetzt.

Da viele Erkenntnisse über Genregulierung in Zebrafischen und Hunden auch auf den Menschen übertragbar sind, ermöglichen wir insgesamt damit neue Möglichkeiten um z.B. den genetischen Hintergrund von Krankheiten im Menschen besser zu verstehen.

ABSTRACT

In order to study gene regulation, large amounts of sequencing data are necessary. We can either generate (hunt) them ourselves or use (gather) publicly available data sets. In order to guarantee the reliability and reusability of the hunted and gathered data, we need to also annotate them with the correct metadata.

In this thesis, I will touch on all three of these aspects. I was part of two international consortia which applied these approaches to two different model organisms. The DANIO-CODE consortium was initiated to systematically annotate the zebrafish genome. Similarly, the Dog Genome Annotation (DoGA) project aims to improve the annotation of genomic elements in the dog genome. Both zebrafish and dogs are popular model organisms for studying biological processes and pathologies in humans. Despite their popularity, both organisms lack a large-scale annotation of regulatory elements.

Before analyzing any data, we designed an annotation structure that captures all aspects of a sequencing experiment that are essential for the processing and analysis of the data. We implemented this structure in a web-platform, which allows easy upload, query, and download of the sequencing data and associated metadata. We present the structure and implementation in Study I, which also contains a comparison to similar and well-established annotation schemata.

We use this annotation structure and the web platform for Study II to collect sequencing data from 1,803 samples from 38 different research groups looking from transcriptomic, epigenomic, and methylomic perspectives at different stages of zebrafish development. We identified more than 140,000 new cis-regulatory elements active during development and provide them together with the sequencing data and genome browser tracks as a resource for the community.

In Study III, we present a biobank for dog tissues established for the DoGA consortium. For both Study III and Study IV, we used 88 and 37 tissues from the biobank, respectively, to catalog promoter regions and their tissue activity using STRT and CAGE-seq. In Study III we also present the web-platform, based on the structure in Study I, where we make the data and the corresponding metadata available. In Study IV, we used the data from CAGE-seq to also identify active enhancer regions and their corresponding tissue activity. We identify regulatory

networks between enhancers and promoters and show their conservation in human.

LIST OF SCIENTIFIC PAPERS

I. **Hörtenhuber, M.**, Mukarram, A. K., Stoiber, M. H., Brown, J. B., Daub, C. O. (2020). *-DCC: A platform to collect, annotate, and explore a large variety of sequencing experiments. GigaScience, 9(3), giaa024.

II. Baranasic*, D., **Hörtenhuber*, M.**, Balwierz*, P., Zehnder*, T., Mukarram*, A. K., Nepal, C., Varnai, C., Hadziev, Y., Jimenez-Gonzalez, A., Li, N., Wragg, J., D'Orazio, F., Díaz, N., Hernández-Rodríguez, B., Chen, Z., Stoiber, M. H., Dong, M., Stevens, I., Ross, S. E., Eagle, A., Martin, R., Obasaju, P., Rastegar, S., McGarvey, A. C., Kopp, W., Chambers, E., Wang, D., Kim, H. R., Acemel, R. D., Naranjo, S., Lapinski, M., Chong, V., Mathavan, S., Peers, B., Sauka-Spengler, T., Vingron, M., Carninci, P., Ohler, U., Lacadie, S. A., Burgess, S., Winata, C., van Eeden, F., Vaquerizas, J.M., Gómez-Skarmeta, J. L., Onichtchouk, D., Brown, J.B., Bogdanovic, O., Westerfield, M., Wardle, F. C., Daub, C. O., Lenhard, B., Müller, F. (in press). Integrated annotation and analysis of genomic features reveal new types of functional elements and large-scale epigenetic phenomena in the developing zebrafish. Nature Genetics

III. Mukarram*, A.K., Arumilli*, M., Hytönen*, M.K., Araujo, C., Quintero, I., **Hörtenhuber, M.**, Syrjä, P., Airas, N., Kaukonen, M., Kyöstilä, K., Pääkkönen, T., Stevens, I., Iivanainen, A., Yoshihara, M., Gusev, O., Bannasch, D., DoGA, Consortium, Ezer, S., Sukura, A., Katayama, S., Daub, C.O., Lohi, H., Kere, J. Dog Gene Promoterome and Tissue Expression Atlas [Manuscript]

IV. **Hörtenhuber*, M.**, Roß*, F., Hytönen*, M., Mukarram, A.K, Aljelaify, R., Quintero, I., Araujo, C., Syrjä, P., Airas, N., Mathelier, A., Kaukonen, M., Kyöstilä, K., Pääkkönen, T., Stevens, I., Raman, A., DoGA Consortium, Sukura, A., Iivanainen, A., Yoshihara, M., Gusev, O., Bannasch, D., Ezer, S., Katayama, S., Kere, J., Lohi, H., Daub, C.O. Enhancer mediated gene regulation in Dog [Manuscript]

* These authors contributed equally

Scientific papers not included in the thesis

- Stevens, I., Mukarram, A.K., **Hörtenhuber, M.**, Meehan, T.F., Rung, J., Daub, C.O., 2020. Ten simple rules for annotating sequencing experiments. PLOS Computational Biology, 16(10), p.e1008260.
- **mashehu**, TrisKast, Menden, K., nf-core bot, Garcia, M., Peltzer, A., Patel, H., Ewels, P. (2021). nf-core/cageseq: [1.0.2] 13.01.2021. Zenodo.

CONTENTS

LIST OF ABBREVIATIONS

ATAC-seq	Assay for transposase-accessible chromatin using sequencing
CAGE-seq	Cap analysis gene expression sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
CTSS	CAGE tag start site
CUT&Tag	cleavage under targets and tagmentation
DCC	Data coordination center
DoGA project	Dog genome annotation project
ENA	European nucleotide archive
ERCC	External RNA controls consortium
FAIR data	Findable, accesible, interoperable and reuseable data
GEO	Gene Expression Omnibus
H3K27ac	Histone 3 lysine 27 acetylation
H3K27me3	Histone 3 lysine 27 trimethylation
H3K4me1	Histone 3 lysine 4 monomethylation
H3K4me3	Histone 3 lysine 4 trimethylation
ENCODE	Encyclopedia of DNA elements
FANTOM	Functional annotation of the mammalian genome
modENCODE	Model organism encyclopedia of DNA elements
RNA pol II	RNA polymerase II
SRA	Sequence read archive
STRT	Single-cell tagged reverse transcription
TF	Transcription factor
TC	Tag cluster
TPM	Tag per million
TSS	Transcription start site

1 INTRODUCTION

1.1 Hunter-gatherer-annotator Science

The traditional image of scientific work is that at the start of a study a scientist sets out a scientific question. To address it, they then go out and build traps, which we call experiments, to acquire data that can answer that question. I would call this *hunter science*.

The image of hunter science is hardly realistic. Scientists almost always work in the context of and with other scientists and their data sets. The sharing of data is one of the hallmarks of academic science and a core principle of Open Science. To support and compare their newly found research results, scientists gather similar, already published data sets and integrate them with their own. I would categorize this as *hunter-gatherer science*.

In ancient history, writing systems and the role of scribes evolved to manage larger societies and improve the access to resources for more people and over longer period of time, allowing further specialization of members of the society away from subsistence farming. Similarly, metadata and annotators arose to define the context of data, improving the ability to merge different data sets and allowing scientists to address new research questions, without producing any new raw data on their own. I call this *hunter-gatherer-annotator science*.

In my Ph.D. studies I worked as a hunter (Studies III & IV), a gatherer (Studies II & IV), and as an annotator (Studies I-IV).

1.2 What is a gene?

In order to talk about gene regulation, I would like to start with a definition of the term gene. As so often with definitions of scientific concepts, it depends on the scientific field, the methods used, and the background of the scientist. By taking a quick look at history, we will arrive at a suitable definition for this thesis.

Wilhelm Johannsen coined the term *gene* in 1909 as an abstract derivative of the term *pangene*, coined by de Vries in 1889.[1,2] Johannsen's goal with this new term was to use a purely conceptual term, describing "the 'something' in the

gamete and in the zygote, respectively, which is of essential significance for the character of an organism". It should come without any hypothesis about the physical counterpart of this concept.

A few years later, Thomas H. Morgan and colleagues linked the term to a physical counterpart, a point on the chromosome with unknown dimensions, on the basis of their work with Drosophila. In 1917 he defined a gene as follows.[3]

> "As a matter of fact it has been found that the many pairs of characters that follow Mendel's law are independent of each other in inheritance. [...] The germ plasm must, therefore, be made up of independent elements of some kind. It is these elements that we call genetic factors or more briefly genes."

He and his colleagues also generated the first theoretical maps of genes (then still called Mendelian factors) on chromosomes, one of the first genome annotation efforts.[4]

In the 1940s, recombination experiments in Drosophila melanogaster and mold showed that genes are regions of a certain length on the chromosome. Experiments with red bread mold by Beadle and Tatum connected genes to proteins.[5] DNA was demonstrated to be the carrier molecule for genes using pneumococcus bacteria[6] in 1944.

A few years later, in 1961 mRNA was revealed to be a midstep in the connection between genes and proteins, establishing the "one gene - one mRNA - one protein" model.[7] In the same year, Jacob and Monod discovered through experiments with bacteriophage λ and E. coli a class of regulatory genes that control the transcription of another gene through the synthesis of an "intracellular substance".[8,9] François Jacob also introduced the term *promoter* for the initiating element of transcription.[10]

The step of "one gene - one mRNA" was demonstrated to be not always true, after experiments in bacteriophage λ and also in mouse indicated alternative promoters used for the same genes.[11,12]

Two independent experiments found at the same time that parts of the adenovirus DNA template were absent in mature mRNA derived from it, introducing the concept of introns.[13,14]

Taking all these findings together, we arrive at the definition of Alberts [15], which

I will follow in this thesis. A gene is a

> ”region of DNA that controls a discrete hereditary characteristic, usually corresponding to a single protein or RNA. This definition includes the entire functional unit, encompassing coding DNA sequences, noncoding regulatory DNA sequences, and introns.”

1.3 Gene regulation

The vast majority of cells in an organism contain a copy of the same DNA. If every cell simultaneously transcribes all genes at the same time, it would be difficult to react to environmental stimuli, for example. Some regulation is required to turn some genes off or on for a certain period of time. There are numerous additional steps to modulate the transcribed RNA and even more for protein-coding genes, but for the purpose of this thesis, I will focus on when genes are transcribed and to what amount. The different aspects are summarized in Figure 1.

1.3.1 Gene transcription

In order to look at the different ways to control the transcription of a protein-coding gene, I will give a quick overview of how eucaryotic DNA is read out.

To transcribe a protein-coding gene, a pre-initiation complex of transcription forms at the start of a gene. This complex consists of many different proteins, including RNA polymerase II (RNA pol II) and general transcription factors (GTFs). After binding to DNA, a short region of the double strand is opened and RNA pol II synthesizes mRNA based on one strand of DNA. RNA pol II remains stalled at the promoter regions until activator proteins trigger the release of RNA pol II and also support the elongation of the transcript.

There are several conditions that must be met to reach this point. I will go through those most important for my thesis.

1.3.2 Cis-regulatory elements in the genome

DNA itself encodes regulatory elements. Elements regulating regions on the same DNA strand are called cis-regulatory elements. This thesis focuses especially on two of these. Promoters, which are essential for the transcription, and enhancers, which act as an amplifying element for the promoters.

Promoters

Promoters are regions on the DNA that bind and position the transcription initiation complex to promote the transcription of a proximal, downstream gene by binding transcription factors (TFs).[16]

In eukaryotes, promoters are usually separated into the core-promoter and the proximal regions. A core promoter region is in the immediate surroundings of the transcription start site (TSS), often contains the TATA box and the initiator element and is usually defined as ± 50 bp around the TSS.[16] While the core-promoter regions are essential for the recruitment of the initiation complex, the proximal promoter regions regulate the initiation rate or elongation. These regions are transcription factor binding sites (TFBSs) and regulate the initiation of transcription within several hundred base pairs upstream of the TSSs.[17]

As discussed above, the same gene can have alternative promoters. The usage of an alternative promoter can be enriched for different tissues or developmental stages.[18]

Enhancers

More than 40 years ago, Benerji et al. reported a 200-fold increase in transcription of a rabbit gene transfected into HeLa cells, if they included not only the whole gene, but also a 72 bp long repeated cis-regulatory element from viral DNA, even if the sequence was included 1,400 bp upstream or 3,300 bp downstream of the target gene.[19] They called this class of elements *enhancers*.

Today, we know that enhancers contain TFBSs for proteins that can act as transcriptional activators and enhance the transcription of a target gene. Active enhancer regions also recruit RNA pol II and are transcribed into non-coding enhancer RNA (eRNA). The question of whether the eRNA is of any function or only a by-product is still under debate.

Similarly to promoter usage, enhancer activity is highly tissue-specific, which is functionally conserved even across genomes.[20,21]

Enhancer-Promoter interactions

Chromatin confirmation is important for enhancers to regulate promoters. Loops have to be formed to bring the elements in proximity to each other. The exact

mechanism of the interaction after looping is still under debate. Three possible modes are discussed:[22]

- Stable contact model: A stable protein-DNA complex forms between enhancers, co-activators and TFs, and promoter regions.
- Kiss-and-run model: A transient contact between enhancer and promoter regions, for example, transcription factors and enhancer-bound proteins apply post-transcriptional modifications are deposited on the promoter, which reside there even after contact.
- Diffusion model: Enzymes at the enhancer site activated TFs, which diffuse to the promoter region. With a reduction in the distance between the enhancer and promoter, the number of activated TFs available would increase.

Generally, enhancer-promoter interactions are an n:n relationship, that is, enhancers can interact with multiple promoters and promoters with multiple enhancers.[23]

1.3.3 Chromatin accessibility

In order to fit the whole DNA into the nucleus, the DNA is wound around histones, and the resulting complex is called chromatin. The chromatin can be further compacted into a 30 nm thick fiber, called *heterochromatin*. These regions are too compact for the trascription preinitiation complex to form. Chromatin can also be less densely structured, which is called *euchromatin*.

The terms euchromatin and heterochromatin were coined by Emil Heitz in 1928, when he saw that some parts of chromosomes of liverworts stained more densely longitudinally than others throughout the cell cycle, while other regions become invisible after a cell has divided.[24] He already identified a correlation between genetically inert regions and their heterochromatic state and between regions with more genes and euchromatic states in Drosophila melanogaster.[25]

The accessibility of chromatin is regulated mainly by histone modifications.

1.3.4 Histone modifications

In 1964 Vincent Allfrey showed, using samples of the calf thymus, that acetylation and histone methylation affect RNA transcription.[26]

By now we know that the acetylation and methylation of different histone tails have effects on transcription. The most important with respect to promoters and enhancers are Histone 3 lysine 27 acetylation (H3K27ac), Histone 3 lysine 27 trimethylation (H3K27me3), Histone 3 lysine 4 monomethylation (H3K4me1), and Histone 3 lysine 4 trimethylation (H3K4me3). Table 1 lists the relations of these modifications with respect to the regulation of enhancer and promoter regions.

Table 1. Histone modifications for promoter and enhancer regions and functionality based on Andersson [27].

Element	State	H3K4me1	H3K4me3	H3K27ac	H3K27me3
Enhancer	active				
	primed/inactive				
	repressed/poised				
Promoter	active				
	primed/inactive				
	repressed/poised				

1.3.5 DNA Methylation

Methylation does not only play a role in the regulation of transcription through methylated histone tails but also methylation of cystine in the DNA. A large subclass of promoters contains CpG islands, which are 200-2000 bp long regions, with a C:G content greater than 50%. In general, regardless of their gene activity, these regions are not methylated.[28] If these CpG-rich promoters are methylated, they become repressed, and the same holds for enhancer regions.[29] Regulation occurs through proteins with a methyl binding domain, which support the introduction of repressive histone marks and chromatin remodeling.[30]

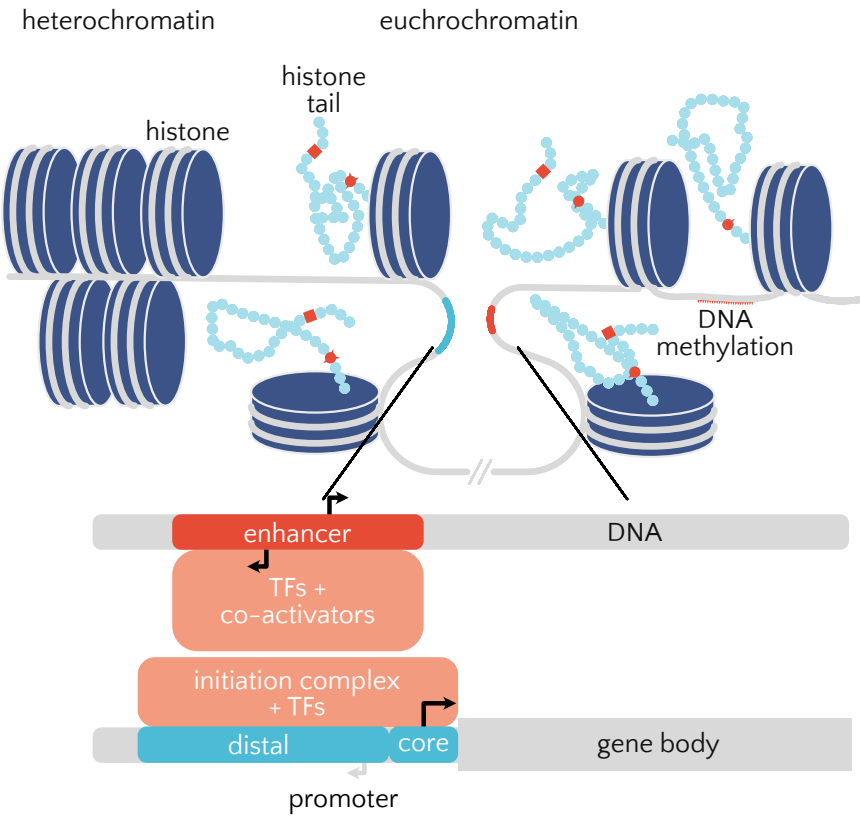


Figure 1. Schematic overview of elements regulating the initiation of transcription with arrows indicating TSSs.

1.4 Model Organisms

A model organism is any animal that we scientists use to model systems in other organisms, especially humans. A good scientific experiment consists of a rigid framework in which as many parameters as possible are fixed in order to better see the effects of the relevant parameters. Ethically and practically, this is easier with animals than with humans, although many of the results are produced to improve our understanding of human biology. As shown by the historical findings described above, researchers have relied on many different organisms to answer scientific questions. Some species are used more frequently than others,

so researchers can better compare results and share resources and methods.

My thesis focuses on the genome regulation of two model organisms, zebrafish, which is a popular model for studying, for example, development, and dog, which is commonly used to better understand hereditary diseases.

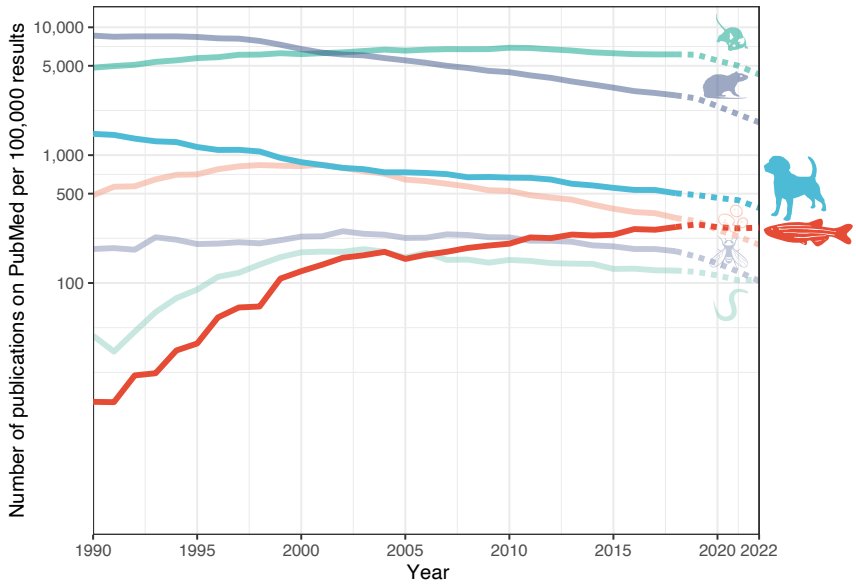


Figure 2. Number of publications with popular model organisms (mouse, rat, dog, zebrafish, fruit fly, and c. elegans), based on publications in MEDLINE journals according to PubMed. Dashed lines indicate the trend after 2019, with an assumed gap of data entry or due to the pandemic.

1.4.1 Zebrafish

The zebrafish (*Danio rerio*) is a small tropical fish that is found naturally in small rivers and side pools in South Asia, but many people are more familiar with it as an aquarium fish.[31] The latter fact was one of the reasons why George Streisinger started to try to establish them as a vertebrate model organism for genetic studies in his lab in the late 1960s.[32] The others reasons why Streisinger chose zebrafish in the end after testing multiple tropical fish, including the already better established medaka, are generally given as follows and are still the main arguments for zebrafish as a model organism:[32,33]

- Zebrafish are easy to handle, can be bred throughout the year, and a female zebrafish can produce hundreds of eggs every week.
- Zebrafish eggs are externally fertilized, allowing for control over fertilization.
- Zebrafish are small enough to monitor their development in a Petri dish, which allows screening for developmental phenotypes. At the same time, embryos are large enough for microinjections and transplantations, allowing one to perturb development at specific time points and locations.
- Zebrafish are vertebrates and contain many organs similar to humans.

An additional feature made them useful for genetic marker assays: The embryos are mostly transparent during their early development, as can be seen in Figure 3.

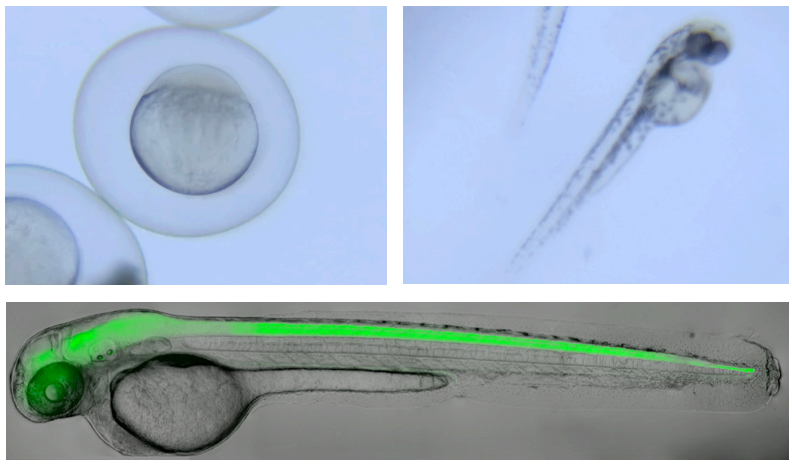


Figure 3. Top-left: Fertilized zebrafish egg, top-right: 2 day old zebrafish, bottom: Neuron-specific enhancer expression labeled with GFP, adapted from Andersson et al. [20].

The current assembly of the reference genome of the zebrafish genome, GRCz11/danRer11, is 1.4 Gb long with 25 chromosomes. The zebrafish presumably underwent a whole genome duplication event, resulting in 25,592 annotated protein-coding genes. Due to duplication, many of these genes have copies, called ohnologues, in the genome, indicated by adding *a* or *b* to the gene symbol. Approximately 26% of the protein-coding genes are ohnolgues.[34]

At least 70 % of human genes are orthologous to a zebrafish gene.[34] Therefore, using zebrafish as a model has led to a wide array of findings on the genetic causes of human disease. From resolving the polygenic background in complex diseases such as Usher syndrome, which causes loss of sight and hearing,[35] to the role of a miRNA in nonsyndromic oral clefts.[36] The popularity of zebrafish as a model organism is still growing and has outpaced more traditional model organisms for genetic research, such as C. elegans or yeast (Figure 2).

Despite the increase in popularity as a model for developmental genomics, there is no large-scale functional annotation available on the developmental dynamics of the genome.[37]

1.4.2 Dog

The dog (*Canis familiaris*) was the first domesticated animal,[38] with zooarchaeological findings indicating that the time of domestication was more than 15,000 years ago.[39] Today, dogs show the highest variability of any mammalian species, with more than 400 different breeds with distinctive behavioral and morphological traits, with respect to the latter, most prominently in skull shape, body size, and pigmentation.[40]

The adaptation of breeding standards and aggressive inbreeding for specific morphologies from the middle of the 19th century reduced the variance of the genetic pool of dogs of the same breed.[41] This led to many breed-specific genetic diseases, making dogs interesting genetic models for studying the cause and treatment of many disorders shared between dogs and humans, including cancer,[42] heart diseases,[43] neurological disorders, for example, dementia,[44] psychological disorders, for example, OCD,[45] as well as diabetes,[46] and aging.[47] To this genetic predisposition comes the fact that dogs live very close to us humans, with the effect that they are exposed to the same environmental factors as us.

The dog genome was, after the human and mouse genomes, the third mammalian genome to have a reference assembly available in 2004/2005.[48,49] The current reference contains 2.4-2.5 Gb (depending on the version), consists of 38 autosomal chromosomes and 30,000 genes, two thirds of which encode proteins.[50] Approximately 93 % of the dog genes have a human orthologue.[51] All current reference genomes are based on female dog samples, therefore, the reference and annotation for the Y-chromosome are still missing.

With the ever growing number of identified gene variants in dogs, we see that many fall into not yet annotated regions of the genome. Therefore, the need for an exact annotation of the gene models and a functional annotation of the genome increases accordingly.[52]

1.5 Gathering and annotating sequencing experiments

The advent of open science and the great increase in the amount of sequencing data, as well as in the variety of experimental designs and sequencing techniques used, led to new opportunities to address and complement research questions with already available sequencing data. A crucial aspect here is to be able to find the appropriate data first and then utilize them with respect to the underlying conducted biological experiments. Therefore, systematic description of available sequencing data, together with description of the underlying biological experiments and sample details, is a crucial prerequisite.[53,54]

Two approaches can be distinguished for collecting and annotating the sequencing experiment data.

General approach:

Sequencing databases such as Sequence Read Archive (SRA),[55] the European Nucleotide Archive (ENA)[56] and Gene Expression Omnibus (GEO)[57] collect raw or processed sequencing data, open them to the community, and provide identifiers to connect the data to scientific publications. Sequencing data are accompanied by a high-level description of the experiments, samples, and technologies used. Due to the scope of these databases and, therefore, limited requirements, the quality of the annotation shows a large variation.[58]

Project specific approach:

Metadata schemata in genome annotation projects including ENCODE,[59] ModENCODE[60] and FANTOM[61] describe experimental aspects more systematically and with a higher level of detail. This enables consistent processing of sequencing data within the projects and further allows for direct comparison between all data. At the same time, significant human resources are required for such data annotation and curation. Furthermore, the underlying technical solutions were specific to each of these projects and were not designed to be reused in other contexts.

In the following, I will give a brief overview of the annotation schemata used in SRA and ENCODE.

SRA

SRA's annotation schema has five main sections: Study, Experiment, Sample, Run, Analysis, and Submission. The Study section is the parent of all other sections and contains general information about the overall study. The section Experiment covers all technical details about the library construction and sequencing platform. The child of this section is Sample, which describes the biological aspects of the biosamples. The Run section links the sequencing files to a Sample instance. The processed files can be linked to the Study in the Analysis section. The Submission section is non-public and saves details about the annotation process and annotator. SRA uses controlled vocabularies for some fields, e.g., species or instrument names. Instances of one section, e.g. one specific biosample, can be reused in other Studies instances. See Figure 4 for the relationships between the sections and some terms used for metadata in SRA.

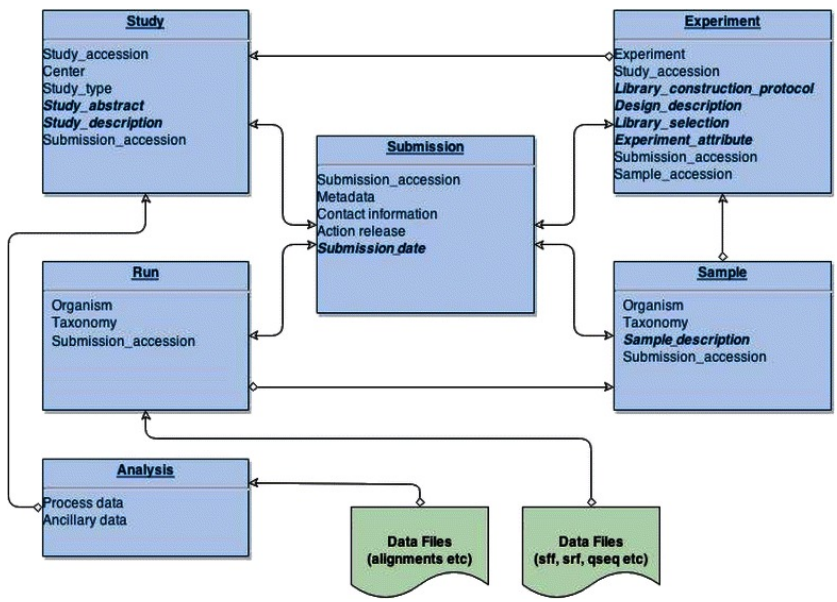


Figure 4. SRA database schema with the relationships between the tables indicated by arrows and some example fields for each table. Source:[58]

ENCODE

The ENCODE annotation schema is separated into six sections: Experiments, Biosamples, Antibodies, Libraries, Files, and Pipelines. In general, it follows the SRA annotation design. Additionally,the SRA Experiment section is divided into Experiment, Antibodies, and Libraries. There is no equivalent to the SRA Study section. A section was added that describes the computational pipelines in detail. The fields themselves are more specifically designed for ENCODE studies and therefore capture more details. Figure 5 gives a brief overview of the fields in each section.

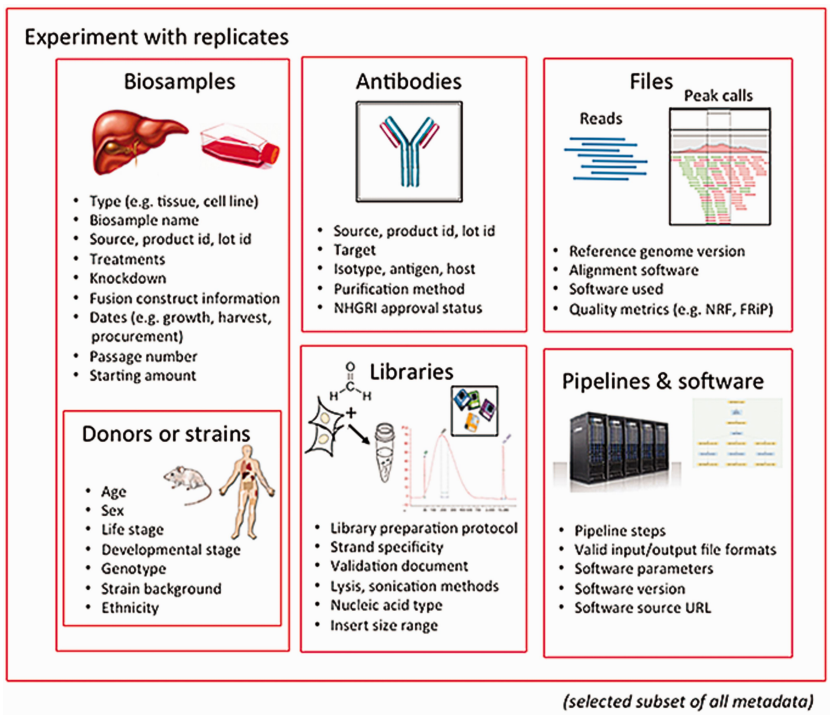


Figure 5. ENCODE metadata schema with a subset of fields in each section. Source: [62]

1.5.1 Principles of scientific data management

In 2016, principles for scientific data management were presented, stating that data should be findable, accessible, interoperable, and reproducible (FAIR).[63]

Both SRA and ENCODE follow them to some extent. The findability in SRA is, for example, hindered by the variability in quality of metadata description.[58]

2 METHODS

2.1 Collaborations and consortia

The complexity of current research questions and the variety of fields to address them require researchers to collaborate. Furthermore, more practical aspects, such as sharing the financial burden or increasing the chance to acquire funding, drive scientists to work together.

One mode of governance for these collaborations is the construction of a consortium. Consortia are "time-limited collective research endeavors, which operate under one or more contractual agreements, and typically have a formal management structure and governance structure".[64]

Two consortia contributed the most to my Ph.D. studies, DANIO-CODE (`birmingham.ac.uk/generic/danio-code/`) with results in Studies I and II, and the Dog Genome Annotation (DoGA) (`doggenomeannotation.org`) consortium for Studies III and IV. These two were different in many aspects, for example, scale, with DANIO-CODE consisting of more than 30 partners around the world, while DoGA started with 3 research groups from Sweden and Finland.

In both cases, one of the most important aspects was, as with any collaboration, communication. In the two consortia, this occurred via different media, most commonly through Slack and email, but also through regular video conference calls, in-person meetings, workshops, and conferences. In addition, secondment visits, where one or two people were embedded in the group of one of the consortium members, were used to strengthen collaborations and make them more efficient.

2.2 Gathering of Material and Data

As mentioned in the introduction of this thesis, different sources of data were used in my Ph.D. studies.

DANIO-CODE data

In the DANIO-CODE consortium, we invited zebrafish labs to submit their transcriptomics, epigenomics, and methylomics data. These data sets were produced mainly for already published studies. We facilitated the initial annotation effort by a multiday jamboree in 2016, where people on site in Liège, Belgium, as well as connected by video calls and chat, could receive support and upload their data. The jamboree helped us to give quick support about definitions of different terms, etc. It also generated a group dynamic, where people gave their full attention to the annotation effort, and in general strengthened the connection of the members to the consortium.

Furthermore, we added complementary data sets from other published studies and produced some data sets to fill open gaps in important developmental stages. We use the DANIO-CODE Data Coordination Center (DCC), which is based on *-DCC (Study I), to collect, annotate, and distribute the data. By taking a snapshot of the data at certain time points, so-called data-freezes, we version the data and allow easier collaboration on the same data sets, while still being able to update them.

DoGA data

In the DoGA project, our consortium partners in Helsinki collected samples from dogs donated to science by their owners or through zoos and hunters in the case of the samples from wolves. We collected and annotated the sequencing files on the DoGA DCC, similar to DANIO-CODE but more in-depth with respect to biological metadata; for example, histological reports are attached to some samples.

2.2.1 Combating batch effects through experimental design

The participation of many different researchers at different time points leads to a loss of controlled parameters in an experiment. This can induce non-bioligcal variation into the data, called batch effects. In order to minimize these effects during library generation in Studies III and IV, we decided to uniformly distribute the tissue of origin, as well as the specific dog of each sample across the deliveries between Helsinki and Stockholm, in the sequencing libraries, and on the flow cells. In the case of the STRT-libraries we additionally added External RNA Controls Consortium (ERCC) spike-in RNA in equal amounts to each sample,

to later normalize for them.

2.3 5'-end RNA sequencing methods

Conventional RNA-seq is able to quantify gene expression, discover transcripts, and splice isoforms. However, regular RNA sequencing has difficulties in identifying the exact TSS if a gene has alternative start sites.[65] Several approaches have been developed to enrich for the 5'-ends of the transcripts in order to better identify their start sites. I will explain two of these methods, which play a significant role in this thesis, in more detail.

2.3.1 CAGE-seq

Cap analysis of gene expression sequencing (CAGE-seq) is a method that uses the 5'cap capture of transcribed RNA to identify exact TSSs and the corresponding expression levels.

Library preparation

In order to produce CAGE libraries, cDNA is generated on extracted RNA strands using random primers. Next, biotin and later magnetic beads are attached to the 5' cap. RNAase I is added to digest loose ends and RNA with incomplete reverse transcription. The complete RNA-cDNA complex is then pulled down via the magnetic beads. The RNA strand is released and a 5' linker with a binding site for the restriction enzyme EcoP15 is attached. A complementary strand is synthesized to the cDNA. EcoP15 is added to cleave the DNA 27 bp after the 5'-end. Sequencing adapters are ligated to the resulting short fragments. The library is then amplified and sequenced on Illumina platforms (Figure 1A).[66]

Data processing

The sequenced reads are then processed with the following steps (Figure 1B):

- QC of sequencing files using FastQC[67] to examine sequencing quality, number of reads, and unexpected sequence biases.
- Trim linker, EcoP15 site, and leading G's to compensate for insertion bias.[68]
- Filter out the rRNA reads.

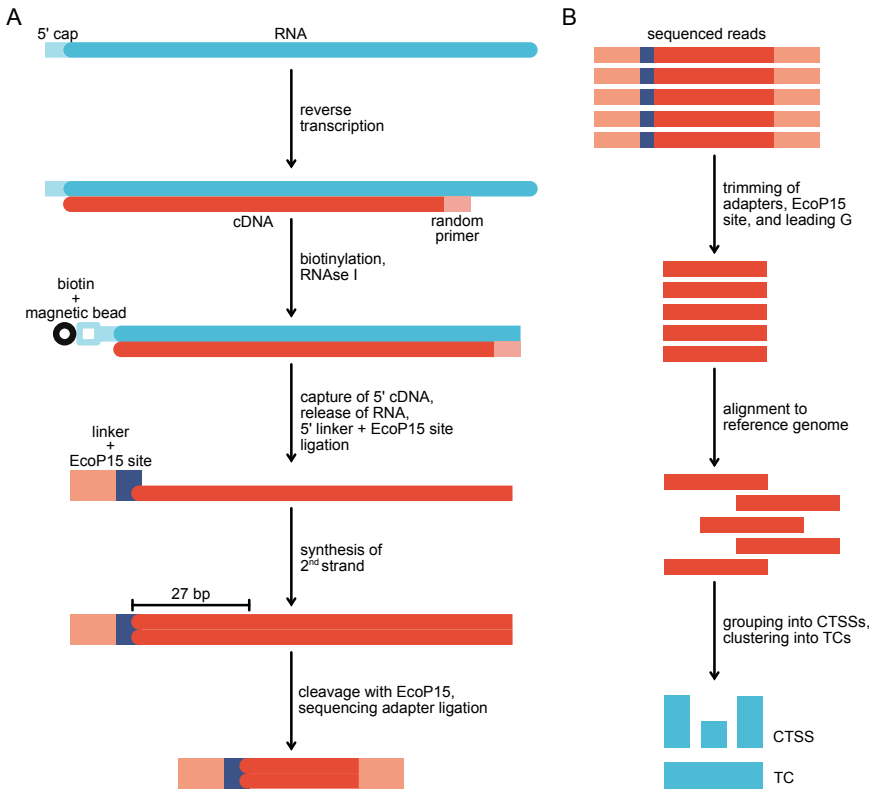


Figure 1. Visualization of steps for CAGE-seq library preparation (A) and data processing (B).

- QC of the trimmed files using FastQC to verify the successful removal of all sequence biases.
- Align to the reference genome. Due to the short CAGE-seq reads, this is commonly done using bowtie.[69]
- QC of alignment results using log files from bowtie, specifically inspecting the number of mapped reads.
- Sum up the aligned reads to CAGE tag start sites (CTSSs), which are 1 bp long regions with the number of CAGE tags mapped to them. To compare expression across samples, tags per million (TPM) normalization is applied.

- Aggregate nearby CTSSs on the same strand into a tag cluster (TC). There are several approaches to cluster CTSSs. I used the slice-reduce approach of CAGEfightR in Studies II-IV.[70] In this method, CTSSs below a certain expression threshold are dropped (slice) and the remaining CTSSs are then grouped together if they are not further apart than a predefined gap width (reduce).

- QC of CTSS and TCs by visual inspection in a genome browser, for example, in Zenbu.[71] Furthermore, if the libraries were produced by different people or sequenced on different machines, a PCA-plot of the TCs with the highest variation in expression can reveal library batch effects.

2.3.2 STRT

Single-cell tagged reverse transcription (STRT) is a library preparation method that uses Oligo-dT primer and template switching to measure mRNA expression levels at the 5'-end even for low input levels, down to single cells.

Library preparation

First, an Oligo-dT primer is biotinylated to the polyA tale of isolated RNA and cDNA is generated with 3-6 cytosines at the 3'-end of the DNA. Using a template switching oligonucleotide, the reverse transcriptase switches the template, and introduces a barcode sequence. Afterwards, the second strand is synthesized. The barcoded CDNA is amplified and magnetic beads are biotinylated to the 5'-ends. After enzymatic fragmentation, the biotinylated fragments are pulled down and can be sequenced.

Data processing

The sequenced STRT reads were processed with Picard,[72] with QC for the number of reads per sample, and removal of PCR duplicates based on unique molecular identifiers. The filtered reads are then aligned to a reference genome with HISAT2.[73] The mapped samples are checked for outliers based on mapping quality and similarity between replicates.

2.4 Data analysis

After the sequenced reads are processed, biologically relevant questions can be addressed through their analysis.

2.4.1 Promoter identification

TCs from CAGE-seq and STRT can be used to identify promoter regions. To extract reliable promoter regions from the CAGE-seq TCs, clusters with low expression and a low number of samples expressing them can be dropped. The strategy for STRT is similar, but due to the ology-dT capture strategy for STRT, the cluster can not only be found at the 5'-starts of a gene, but a certain amount of exon painting occurs as well. Especially, the 3'-ends of a gene show often a high amount of expression, which need to be taken into account when considering quantification.

2.4.2 Enhancer prediction

CAGE-seq is capable of capturing eRNA. Therefore, active enhancers can be identified by their balanced, bidirectional TC clusters, which is an even more reliable method compared to enhancer prediciton based on histone modifications.[20] For Studies II and IV, we used the enhancer detection method implemented in CAGEfightR. This approach first selects TCs that are in close proximity to each other but on opposite strands. Subsequently, the balance score of these bidirectional clusters is calculated. We used the default Bhattacharyya coefficient. This coefficient splits the bidirectional cluster from the midpoint in a given window (we used the default ± 200 bp) and by strand. The expression scores for each of the four quadrants are then used to calculate the coefficient as follows.

Definition. *Let PU be the expression on the plus strand in the upstream direction from the midpoint, MU the expression on the minus strand upstream, PD the expression in the downstream direction from the midpoint on the plus strand, and MD the expression downstream on the minus strand. The Bhattacharyya coefficient BC is then defined as*

$$BC := \sqrt{\frac{MD}{PU + MU + PD + MD} \cdot \frac{1}{2}} + \sqrt{\frac{PD}{PU + MU + PD + MD} \cdot \frac{1}{2}}$$

The Bhattacharyya coefficient is equal to 1 for a perfectly balanced enhancer signal.

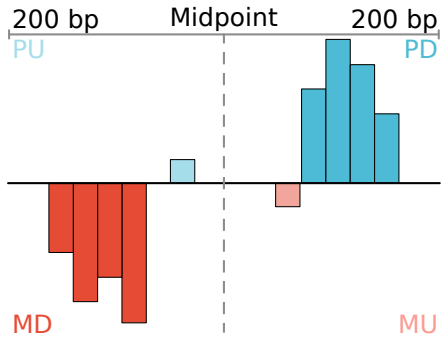


Figure 2. Example of a well balance bimodal enhancer signal according to the Bhattacharyya coefficient with same colors as in the formula.

We used the default threshold of 0.95 in studies II and IV. The bidirectional clusters are then further filtered to not overlap any identified enhancer and are only in annotated intronic or intergenic regions. An additional filter step is to require a certain number of samples to have a bidirectional signal in a cluster.

2.4.3 Enhancer-promoter interactions

CAGE-seq also allows the detection of enhancer-promoter interactions. This is due to the fact that the production of eRNA is associated with active enhancers, transcribed promoter regions with the gene product and both can be quantified using CAGE-seq. The correlation between expression patterns has been shown to correspond to the interaction between the specific enhancer and the promoter.[20]

3 RESEARCH AIMS

The aim of this thesis is to use large-scale sequencing experiments to improve our knowledge of regulatory elements in the model organisms zebrafish and dog. The specific aims of the studies are the following:

- Study I: Designing and implementing a species-specific, but flexible annotation schema to capture, explore, and export data and metadata from sequencing experiments. This platform should facilitate the collection and distribution of a wide variety of data and its metadata, while guaranteeing a high quality of it.
- Study II: Generating an atlas of regulatory elements during zebrafish development.
- Study III: Establishing a bio-bank of dog and wolf tissue samples and identifying promoter regions using STRT.
- Study IV: Creating a tissue-level catalog of promoters, enhancers, and their interaction in dogs.

4 RESULTS

4.1 Study I: *-DCC: A platform to collect, annotate, and explore a large variety of sequencing experiments

Sequencing experiments become ever more complex and so becomes the number of parameters one needs to consider when analyzing the resulting data. In Study I, we approached the challenge of capturing sequencing experiments in such a way that makes data input easy, but also guarantees reliable downstream processing and analysis of the data.

We looked at already available platforms, especially SRA, ENCODE, and mod-ENCODE and added more fields that are specific to a modal organism. We presented an annotation structure, designed to capture a sequencing experiment, from the study idea to the processed sequencing files. We separated them into five sections (Series, Biosamples, Assay, Sequencing, and Data). To guarantee a minimum quality of annotations, we required some fields (e.g. sequencing platform) to be filled out. Depending on the entered data, some fields can be conditionally required; for example, if the biosample type is set as 'tissue', one needs to also provide an anatomical term.

We took our annotation schema and implemented it into a web-platform based on a Django back-end. This allows for easy adaptation of the metadata fields and their requirements for different applications. The platform comes with two ways to add annotations and files. One way is a web form that goes stepwise through each annotation section. The other way is through a csv file, which is recommended for larger uploads. Additionally, the platform allows one to browse through the annotations, select a subset of files and download them together with the annotation. Processed files can be uploaded and linked to the original files and annotation through a web-form.

Modified versions of the annotation schema were implemented in a DCC for the DANIO-CODE consortium as well as for the DoGA consortium.

4.2 Study II: Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements

The goal of the DANIO-CODE consortium was to use publicly available zebrafish sequencing data to annotate the functional elements of the zebrafish genome. In Study II we report the first results of the effort. After we set up the *-DCC platform from Study I for zebrafish data at `danio-code.zfin.org`, we collected and annotated 1,802 sequenced samples from 38 research groups, covering 38 different developmental stages and 19 different types of assays. The types of assays span methylomics, epigenomics, and transcriptomics methods.

We identified more than 140,000 cis-regulatory elements using ChIP-seq histone marks and open regions based on ATAC-seq. Using CAGE-seq data as described in 2.4.2, we verified elements classified as enhancer.

Furthermore, we provide high-precision TSSs, alternative promoters, and promoter architecture classes during development. We also show that the regulatory elements active in the early stages cluster in regions rich in H3K27ac. The study also presents a method for transferring regulatory regions between fish and mammals regardless of sequence conservation.

To facilitate a wide usability of this resource, we also generated a UCSC trackhub, consisting of all data used in the analysis, as well as generated annotation tracks for the newly identified elements.

4.3 Study III: Dog Gene Promoterome and Tissue Expression Atlas

In Study III, we present the DoGA biobank and an STRT based promoterome. The biobank comprises over 5000 samples from 120 tissues from 16 dogs and is the most comprehensive collection of dog tissues to our knowledge.

We took a subset of 428 samples from the biobank, extracted RNA and generated STRT libraries. The 428 samples cover 88 tissues from 9 dogs. Additionally, we performed a pilot CAGE-seq run with 10 samples from 5 dogs. The sequencing files were deposited together with their metadata in the DoGA DCC. This platform was also based on *-DCC and it's implementation presented in Study I.

We performed the promoter prediction as described in 2.4.1. In total, we report 56,236 robust promoters using STRT, of which 11,366 were supported by CAGE-seq.

4.4 Study IV: Enhancer mediated gene regulation in Dog

For Study IV we generated CAGE-seq libraries from 116 samples taken from the DoGA biobank described in Study III. These samples cover 37 tissues from 10 different dogs, including embryos.

The identification of promoter regions resulted in approximately 55,000 comprehensive promoters. We also selected a subset of ca. 12,000 highly reliable robust promoters. We identified 18,500 and 5,510 alternative promoter regions, respectively. Compared to the current annotation of gene models, we find $\sim 23,000$ and $\sim 12,000$ promoter regions, which are not in the vicinity of any annotated transcripts in Ensembl.

We overlapped the promoter regions with previously published ATAC-seq and ChIP-seq for histone marks. The comprehensive set of promoter regions expressed in the same tissues as the previously published data sets has an overlap of 70% with open regions. 79% of the robust set fall into open regions. Regarding transcriptional histone marks, 89% of the comprehensive promoters have the mark for active transcription H3K27ac and 82% the promoter-mark H3K4me3. The overlap is even higher for the robust set with 93% coincide with an H3K27ac signal and 90% with H3K4me3.

Because CAGE-seq also captures eRNA and allows the identification of active enhancer regions, we used the method described in 2.4.2. Using different reliability thresholds, we report 28,565 comprehensive enhancer candidates and 5,482 robust ones.

Using the same data sets as for the promoter regions, we show that 36% and 62%, respectively, overlap with open regions. Approximately 57% and 83%, respectively, coincide with the H3K27ac signals. Furthermore, 58% of the comprehensive and 78% are supported by the enhancer mark H3K4me1.

Moreover, we identified regulatory regions enriched in each of the tissues. Testis is the tissue with the most enriched promoters and enhancers. Skin has the second highest number of enriched promoters, while the retina / optical nerve sample has the second highest number of enriched enhancers. Based on distance and

co-expression, we identify ~59,000 enhancer-promoter interactions between comprehensive sets and ~13,000 robust enhancer-promoter pairs.

Based on sequence alignment, we also identify ~400 enhancers with orthology with human enhancers. Looking at their interaction with promoters, we can link 66 of these enhancers to 103 promoters with an orthologous gene in humans. Adding the aspect of tissue enrichment, we find 8 conserved regulatory networks, meaning enhancers which are conserved in human, interacting with the same orthologuous genes, and also enriched in the same tissues.

5 DISCUSSION AND PERSPECTIVES

5.1 What is a gene again and what about promoters and enhancers?

I started this thesis with a discussion of the definition of a gene, from its abstract beginnings as some inherited factor to a region on the DNA. However, the most common term I use in my work is *tag cluster*, which depending on the context and the extension of the surrounding window can mean TSS, promoter, or enhancer. According to the definition of Alberts[15] TCs could be considered genes. How come we don't use that term then? On the one hand, as a bioinformatician, I look with data science glasses at my results, which makes it easier to use the more abstract and non-biological term TC. This is the case even if we usually name them based on their location in the genome before associating them with a specific gene ID or symbol. On the other hand, in my experience, we use the term gene only when we leave the data science part of our analysis and want to look at the biological features of certain findings, meaning that we often deal with a more traditional definition of the term.

As with any definition, the text book definition of Alberts[15] also has its limits. According to it, a gene includes the "functional unit, encompassing coding DNA sequences, noncoding regulatory DNA sequences, and introns", which would mean enhancers are part of a gene. Consequently, the promiscuity and tissue specificity of enhancer-promoter interactions make the physical dimension of the gene even more dependent on the context. In the end, even with our numerous attempts to pin down genes on the genome, Johannsen's abstract definition of a gene as something "which is of essential significance for the character of an organism" might be helpful to not be too strict with our concepts.

Furthermore, comparing the classical definitions of promoters and enhancers with recent discoveries gives a less clear picture. Promoters can amplify the transcription of distal genes[74,75] and enhancers also act as transcription initiation sites.[76] Should we discard the distinction and talk generally about cis-regulatory elements for transcription that come with a certain enhancer or promoter potential as proposed in?[17] Or are we missing subtypes or subregions of enhancers and promoters that would more clearly separate these entities?

5.2 Successful Ph.D. studies in consortia

The power of large-scale scientific collaborations and consortia also comes with costs. A version of Brook's law[77] says that the productivity of a team scales worse than linearly with the number of members. This is also true for research collaborations. A scientific corollary would be that the speed with which a publication is written does not scale linearly with the number of shared first-authors, as I have experienced in Study II. However, each of the five shared co-first authors in this study was essential throughout the study. The high number also shows the difficulty in correct attribution of contributions, especially in large-scale collaborations.

The time dilation in collaborations and consortia also makes it difficult to fit four years of Ph.D. studies in it. I was fortunate to be instrumental in the setup for the data collection, as well as seeing the fruits of it in publishing Study II. However, this, together with the pandemic, meant that it took me longer than the target of four years to be able to defend my Ph.D. thesis. One possibility would have been to finish before Study II was published. However, this would either result in not getting credited in the same manner, as a first-authorship generally comes with an assumed shared workload until the release of the publication. Alternatively, joining later in the project would have prohibited me from learning how to design a study and also from having influence on it.

5.3 Bulk vs. Single-cell sequencing

Most of the data in this thesis are bulk sequencing data, while single-cell sequencing has been one of the biggest innovations in recent years.[78,79] Bulk sequencing allows only for the measurement of the average signal of a given input, which can consist of millions of different cells in the case of ChIP-seq, for example.[80] The higher resolution of single-cell data helped, for example, identify new cell types and states,[81] and cellular dynamics.[82]

On the one hand, this is due to the small amount of available single-cell sequencing data at the beginning of Study II. Sequencing data are usually only made publicly available when the corresponding studies are accepted by a journal. This process often takes a long time, especially with newly established techniques. The delay between the development of a method and the application for larger studies reduces the number of public data sets available even further. For example, the

protocol for single-cell combinatorial indexing ATAC-seq (sciATAC-seq) was first described in 2015.[83] The study using this technique in a larger cohort of zebrafish embryos, used in Study II, was first published in June 2020 on bioRxiv,[84] while the data were published with the first revision more than a year later, in September 2021.

On the other hand, even though high-throughput single-cell RNA-seq started a decade ago, not all bulk sequencing techniques are yet available at the single-cell level. The single-cell version of the CAGE-seq protocol, C1 CAGE, was published only in 2019.[85] Similarly, a high-throughput method for analyzing histone modifications at single cell levels was only published last year with single-cell cleavage under targets and tagmentation (scCUT&Tag).[86]

Finally, from a more pragmatic point of view: Even by only considering the averages of up to several million cells, we are still able to discover new regulatory elements and mechanisms.

5.4 Reinventing the wheel, or The development of tools in a scientific context

For Study I, we developed an annotation system for sequencing experiments and implemented versions of it for Studies II-IV. Was this even necessary, when SRA and ENA are well-established platforms? Furthermore, we even based our annotation structure on theirs. There are four closely related reasons why this approach still benefits science.

First, neither the DANIO-CODE DCC nor the DoGA DCC is thought of as replacements for more general repositories for sequencing data. On the contrary, all raw reads in the DCCs are or will be made available in SRA/ENA. Our specialized DCCs will improve the findability and accessibility aspects of the FAIR principles. Although the metadata are well indexed and, therefore, in principle findable, the large variety of available data makes it difficult to download it, especially for persons not so well versed in data retrieval and the data structures of the specific data repositories. The fact that our DCCs contain only data from one species with metadata fields and names specific to this species makes it easier to find and access our data.

Furthermore, we also use DCC platforms to share data, including metadata, on the basis of data freezes between partners, which is not the recommended usage

of SRA/ENA. Working with the same metadata from the beginning also helps to find gaps in it and, in the end, strengthens the quality of the data and the annotations.

Having the data on our own platforms also allows us to link them more strongly to the derived data and findings. Showing raw data in addition to processed data enhances the discovery and reusability of the data.

Finally, our DCCs are also social spaces for specific research communities. Although it is not a medium for direct communication, it acts as a platform for future collaborations and a ”waterhole” to find new members for current or future consortia.

5.5 Ethical perspectives on my studies

In genetics, we usually look at two aspects of ethical considerations.

First, we rely heavily on biological samples, in this case, from two model organisms, where the individuals are killed to extract the material. For studies III and IV, we extracted samples only from dogs whose owners wanted to put them down for medical reasons. Only after this procedure did the pathologists start to extract the samples. This ensured that no unnecessary harm was done to the animals and was also approved by the Ethics Board of Finland. The data we used for study II were already published, which means that no additional animals were injured or came from very early stages of zebrafish development, when no cognitive function, which could experience suffering, is assumed according to EU guidelines.[87]

Second, can the research we are doing lead to increased acceptance of eugenics? Although our findings are about regulatory elements in zebrafish and dogs, the goal of our research is to use that information for humans. However, none of our findings are directly related to human phenotypes and therefore could be used to select or discriminate humans. Of course, the general risk with fundamental research is that our findings could turn out to be used to actually establish a link, but I hope that civil society, including us scientists, will always take an active stance against any notion of eugenics and intervene before any established links are applied.

A second field of ethical considerations concerns the direct environmental impact of our research. All our computational work requires electricity, which comes with a certain carbon footprint. It is difficult to estimate how much the actual

amount is. I used the UPPMAX high-performance computing cluster or my local computer for all the analysis. According to the UPPMAX operator, they are using a green energy mix and air from the server halls is used for both heating and cooling of adjacent university buildings. The energy impact of my local laptop is even lower than that of UPPMAX and comes from a similar energy mix. Therefore, we assume that, even though there is an impact on carbon output, the impact of the computations is low, and we hope that the impact of the research result compensates for the impact of the released carbon.

Up until the start of the COVID-19 pandemic, we assumed that in person gatherings were all necessary and irreplaceable aspects of research work. Due to my multi-national collaborations, I took numerous flights to meetings, secondments, and conferences during my Ph.D., with a high carbon footprint. At least with respect to meetings and conferences, a more hybrid approach with different benefits and drawbacks has been established. One result is that this change in work modes reduced the carbon output of our multinational collaborations, by only heaving video meetings instead of annual in-person meet-ups. My secondment visits to the west coast of the United States and Japan were the largest singular expenses of my CO_2 budget and generated approximately 5.7 tons of CO_2 (compared to 2.5 tons for all trips inside of Europe). These secondment stays were vital in the establishment of the DANIO-CODE DCC and it is hoped that their impact on research outweighs the induced damage.

6 CONCLUSIONS

Using the hunter-gatherer-annotator approach, my objective was to improve our knowledge of regulatory elements, especially promoter and enhancer regions in zebrafish and dogs. This will enable the research community to use these model organisms to better understand biological processes and their dysregulation in humans.

For Study I, we designed an annotation structure that enables one to capture the metadata of sequencing experiments for downstream analysis. We also implemented the concept into a web-platform, which enables easy, but controlled, entry of data, querying the data, and also data export. This structure was then used to manage the data for Studies II-IV.

In Study II, we showed the possibility of a large-scale genome annotation effort, achieved with a majority of already published data, to catalog a large number of cis-regulatory elements active during the development of zebrafish. This shows the feasibility of one of the promises of open science by reanalyzing and integrating publicly available data to generate new insights. At the same time, we also offer the data to the community to support their own research.

For Study III and IV, we went the opposite way and first created a biobank for samples from dog and wolf tissues. From a subset of these, we generated our own STRT libraries and CAGE-seq libraries. The low input requirements for STRT allowed us to catalog promoter regions in numerous tissues and subtissues. With the help of CAGE-seq, we provide additional support for the identified STRT libraries and identified active enhancer regions. Furthermore, we looked for cis-regulatory elements enriched in certain tissues and interactions between enhancers and promoters. Finally, we identified several enhancers, promoters, and their interactions, which are conserved between dogs and humans.

7 ACKNOWLEDGMENTS

The success of hunter-gatherers depends on working as a group and supporting each other. This is also true for hunter-gatherer-annotators.

First, I would like to thank my main supervisor, **Carsten Daub**, who was with me throughout all my scientific "hunts". You might have not realized from the beginning what you got into when accepting a mathematician from Austria (you know, the one with not so many free-roaming kangaroos (-;). But we soon found our shared "passion" for rigorous annotation structures. I will take many things with me from the years we had together. For sure your emphasis on direct social interactions for collaborations, which sent me off in many different directions and allowed me to experience the (work) culture on the west coast of the US and in Japan. Although I have always escaped doing any of your many sport activities with you, maybe one day your knees will also become weak and you will join me on the bike.

A warm thank you to my co-supervisor **Monte Westerfield**. Your calm and open attitude and the appreciation people surrounding you have for you, are a guiding example for me in the oftentimes over-competetive world of science.

A special thank you to **Per Uhlén**. You kickstarted my entire adventure in Sweden, by accepting me into your lab for my master thesis, based on an e-mail written in very German Swedish. You opened my eyes to Karolinska Institute as a possible home and showed me many new ways how I can use my skills.

Many thanks to **Ferenc Müller**, who, by launching and managing DANIO-CODE and the ZENCODE-ITN, created great opportunities for my Ph.D. and for all the other ITN-members/ZENKIDS. Your enthusiasm and your endurance fueled us through the long journey of Study II, which finally paid off. A special thanks also for recognizing the value of building scientific resources.

Speaking of resources, I would like to thank all members of **ZFIN**, which not only generate a highly valuable scientific support but who, as I experienced especially during my stay in Eugene, are also a bunch of very nice people and great hosts. Special thanks to **Anne Eagle** for coordinating a lot of the things and to **Ryan Martin** for being a great admin.

Thanks you to my hosts in Japan **Piero Carninci** and **Erik Arner** and to **Jesscia Severin** and **Jayson Harshbarger** for all their help. Thank you also to **Alessandro Bonetti** and **Ana Maria Suzuki**, for the excellent food and fun chats.

I also want to thank the two other cofounders of the DoGA consortia, **Hannes Lohi** and **Juha Kerre**. I didn't know at the beginning, but you would become like two additional co-supervisors to me. You taught me so many things during all our meetings and dinners, which made me a better scientist. Also thank you to all the other members of the DoGA consortium, it is a privilege to see all of your work and I hope my analyses can help you go even further. Special thanks to **Sruthi** and **Mehar** for all the help with the data handling and analysis.

Thank you, **Ben Brown** and **Marcus Stoiber**, you were invaluable for the establishment of *-DCC and helped to get everything rolling in the beginning.

Thank you **Damir**, I enjoyed all the discussions we had and without your willpower and endurance we would not have gotten Study II where it is today.

Now to another group of special people during my Ph.D., the Daub lab members. Thank you **Kadir**, for being by my side for many of the 'hunts'. Your deep knowledge about science and the network around it always amazes me. Thank you also to **Enrichetta**, you always showed me the way through my Ph.D. studies and were a very good anchor for the whole group. Thank you to **Tahmina**, I thoroughly enjoyed our time together and I hope there will be many more opportunities in the future where I can make my cheesecake for you. Thank you **Rasha**, for all the times you helped organize presents, fika, and other social emergencies (also thanks for the CAGE-seq libraries!). Thank you **Marine**, the Queen of Dominion, it was so much fun to have you in our lab and thanks to the lab videos some of it is captured for posterity. Thank you **Fiona**, it was great to have you in the lab and your contributions really helped push my Ph.D. forward. Thank you **Sabina**, you might have had some bad timing by coming in just at the end of my Ph.D. and meeting a Matthias who was rather tired (of the Ph.D. and of COVID), but working with you gave me more confidence in the data and myself, and I feel privileged to have guided you a bit on your way. Thank you **Lea**, you also came to the group at the end of my era and I can already see that we were missing someone like you in the lab in recent years. Thanks for all you shared thoughts and points of view at the office, in meetings, and at the lunch table. Special thanks to two former Daub lab members, **Michaël** and **Irene**, your work made my life much easier and brought the DCCs into the nice shape they

are now. Thank you to all the other former Daub lab members, **Amitha**, **Niyaz**, and **Kelvin**, I learned many things from you and I enjoyed our time together. Also, thank you to all the other students who I had the honor of supervising. You not only made Swedish summers less quiet and boring, you also taught me a lot along the way.

I was very lucky to have an even larger second group at the beginning of my Ph.D. with the **ZENCODE-ITN**. Going with you to all the different workshops and meetings always felt like a fun school trip (maybe with a bit more Jenga than I remember playing back in school) and it was a great support to have people on my side struggling with similar things and seeing them achieve great things. A special thanks to **Ana** for all the (too unappreciated) organizing you did. You are the reason we had great trips and helped a lot to fuse the group even more together. I also want to thank you, **Andrea**, for your infectious enthusiasm and all the great time we had together.

I want to also thank the **nf-core community**, which is a very nice harbor for Open Science ideas and just a group of very nice people. I stumbled into you by a happy little accident, but your openness made me feel immediately welcomed and appreciated. You helped me engage with a worldwide community, which was a nice distraction from my normal day, especially during the COVID-winters with little scientific and social exchange otherwise. In addition, your pipelines helped me when I needed to quickly process some data.

Thank you to all the people who I happened to meet during my years in NOVUM, NEO and at the BioNut department, you made the sometimes dull Flemingsberg a lot more fun and interesting. Special thanks to **Eva** and **Monica**, for helping me through all the bureaucracy.

Thanks **Emilia**, without your open and warm attitude (and ignoring my first surprised "no"), I would not have gotten such a good bike-to-work companion, flute partner, and friend.

Thank you **Joana**, for all your patience in the last year, for your support, and for sharing your time (and all the butter (-;) with me.

Thanks for all the Sunday evenings, **Marie**. One of the positive outcomes of the COVID-19 pandemic was that I was able (re)build this bridge to you. Thanks for always having an open ear and even a flat (when I happen to start my vacation one week too early)!

Two people helped me greatly to find my way in Sweden, so thank you for making me feel more welcome here, **Stina** and **Amanda**. I cherished all the time and messages we shared. They made my life here much easier and more interesting.

Two other people laid the foundation for my long way to this Ph.D. thesis, held me by my hand (literally, for a long time), and always helped me when I needed something. Danke **Mama** & **Papa**! I know it is not easy to have a child so far away, so I greatly appreciate all your support (and, at the end, I finally got to a medical school like you wanted (-:). Thanks also to my brothers, **Flo** and **Tom**, to **Hansi** and all the rest of my family in Austria, for always making me feel welcome and rooting, as well as widening, my perspectives!

REFERENCES

[1] Johannsen, W. *Elemente der exakten Erblichkeitslehre*. Fischer, 1909.

[2] De Vries, H. *Intracellulare pangenesis*. G. Fischer, 1889.

[3] Morgan, T. H. "The Theory of the Gene". In: *The American Naturalist* 51.609 (1917), pp. 513–544.

[4] Sturtevant, A. H. "The linear arrangement of six sex? linked factors in Drosophila, as shown by their mode of association". In: *Journal of experimental zoology* 14.1 (1913), pp. 43–59.

[5] Beadle, G. W. and Tatum, E. L. "Genetic Control of Biochemical Reactions in Neurospora". In: *Proceedings of the National Academy of Sciences of the United States of America* 27.11 (Nov. 15, 1941), pp. 499–506.

[6] Avery, O. T., MacLeod, C. M., and McCarty, M. "STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES". In: *The Journal of Experimental Medicine* 79.2 (Feb. 1, 1944), pp. 137–158.

[7] Brenner, S., Jacob, F., and Meselson, M. "An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis". In: *Nature* 190.4776 (May 1961), pp. 576–581. DOI: 10.1038/190576a0.

[8] Jacob, F. and Monod, J. "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3 (June 1, 1961), pp. 318–356. DOI: 10.1016/S0022-2836(61)80072-7.

[9] Monod, J. and Jacob, F. "General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation". In: *Cold Spring Harbor Symposia on Quantitative Biology* 26 (Jan. 1, 1961), pp. 389–401. DOI: 10.1101/SQB.1961.026.01.048.

[10] Jacob, F., Ullman, A., and Monod, J. "LE PROMOTEUR,ELEMENT GENETIQUE NECESSAIREA L'EXPRESSION D'UN OPERON". In: *CR Hebd Seances Acad Sci* 258 (1964), pp. 3125–3128.

[11] Reichardt, L. and Kaiser, A. "Control of λ repressor synthesis". In: *Proceedings of the National Academy of Sciences* 68.9 (1971), pp. 2185–2189.

[12] Hagenbüchle, O., Tosi, M., Schibler, U., Bovey, R., Wellauer, P. K., and Young, R. A. "Mouse liver and salivary gland α-amylase mRNAs differ only in 5' non-translated sequences". In: *Nature* 289.5799 (Feb. 1981), pp. 643–646. DOI: 10.1038/289643a0.

[13] Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA". In: *Cell* 12.1 (Sept. 1, 1977), pp. 1–8. DOI: 10.1016/0092-8674(77)90180-5.

[14] Berget, S. M., Moore, C., and Sharp, P. A. "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.8 (Aug. 1977), pp. 3171–3175.

[15] Alberts, B. *Molecular biology of the cell*. 2017. ISBN: 978-1-315-73536-8.

[16] Lenhard, B., Sandelin, A., and Carninci, P. "Metazoan promoters: emerging characteristics and insights into transcriptional regulation". In: *Nature Reviews Genetics* 13.4 (Apr. 2012), pp. 233–245. DOI: 10.1038/nrg3163.

[17] Andersson, R. and Sandelin, A. "Determinants of enhancer and promoter activities of regulatory elements". In: *Nature Reviews Genetics* 21.2 (Feb. 2020), pp. 71–87. DOI: 10.1038/s41576-019-0173-8.

[18] Consortium, F. et al. "A promoter-level mammalian expression atlas". In: *Nature* 507.7493 (Mar. 27, 2014), pp. 462–470. DOI: 10.1038/nature13182.

[19] Banerji, J., Rusconi, S., and Schaffner, W. "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences". In: *Cell* 27.2 (Dec. 1, 1981), pp. 299–308. DOI: 10.1016/0092-8674(81)90413-X.

[20] Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., and Sandelin, A. "An

atlas of active enhancers across human cell types and tissues". In: *Nature* 507.7493 (Mar. 2014), pp. 455–461. DOI: 10.1038/nature12787.

[21] Wong, E. S., Zheng, D., Tan, S. Z., Bower, N. I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B. M., Kikuchi, K., McGlinn, E., Francois, M., and Degnan, B. M. "Deep conservation of the enhancer regulatory code in animals". In: *Science* 370.6517 (Nov. 6, 2020), eaax8137. DOI: 10.1126/science.aax8137.

[22] Karr, J. P., Ferrie, J. J., Tjian, R., and Darzacq, X. "The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer–promoter communication". In: *Genes & Development* (Dec. 30, 2021). DOI: 10.1101/gad.349160.121.

[23] Arensbergen, J. van, Steensel, B. van, and Bussemaker, H. J. "In search of the determinants of enhancer–promoter interaction specificity". In: *Trends in Cell Biology* 24.11 (Nov. 1, 2014), pp. 695–702. DOI: 10.1016/j.tcb.2014.07.004.

[24] Heitz, E. "Das Heterochromatin der Moose". In: *I. Jahrb. Wiss. Bot* 69 (1928), pp. 762–818.

[25] Heitz, E. "Die somatische Heteropyknose bei Drosophila melanogaster und ihre genetische Bedeutung". In: *Zeitschrift für Zellforschung und Mikroskopische Anatomie* 20.1 (1933), pp. 237–287.

[26] Allfrey, V. G., Faulkner, R., and Mirsky, A. "Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 51.5 (1964), p. 786.

[27] Andersson, R. "Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model". In: *BioEssays* 37.3 (2015), pp. 314–323. DOI: 10.1002/bies.201400162.

[28] Long, H. K., Sims, D., Heger, A., Blackledge, N. P., Kutter, C., Wright, M. L., Grützner, F., Odom, D. T., Patient, R., Ponting, C. P., and Klose, R. J. "Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates". In: *eLife* 2 (Feb. 26, 2013). Ed. by A. Ferguson-Smith, e00348. DOI: 10.7554/eLife.00348.

[29] Angeloni, A. and Bogdanovic, O. "Enhancer DNA methylation: implications for gene regulation". In: *Essays in Biochemistry* 63.6 (Sept. 24, 2019), pp. 707–715. DOI: 10.1042/EBC20190030.

[30] Du, Q., Luu, P.-L., Stirzaker, C., and Clark, S. J. "Methyl-CpG-binding domain proteins: readers of the epigenome". In: *Epigenomics* 7.6 (Sept. 2015), pp. 1051–1073. DOI: 10.2217/epi.15.39.

[31] Parichy, D. M. "Advancing biology through a deeper understanding of zebrafish ecology and evolution". In: *eLife* 4 (Mar. 25, 2015), e05635. DOI: 10.7554/eLife.05635.

[32] Grunwald, D. J. and Eisen, J. S. "Headwaters of the zebrafish — emergence of a new model vertebrate". In: *Nature Reviews Genetics* 3.9 (Sept. 2002), pp. 717–724. DOI: 10.1038/nrg892.

[33] Streisinger, G., Walker, C., Dower, N., Knauber, D., and Singer, F. "Production of clones of homozygous diploid zebra fish (Brachydanio rerio)". In: *Nature* 291.5813 (May 1981), pp. 293–296. DOI: 10.1038/291293a0.

[34] Howe, K. et al. "The zebrafish reference genome sequence and its relationship to the human genome". In: *Nature* 496.7446 (Apr. 2013), pp. 498–503. DOI: 10.1038/nature12111.

[35] Eberhart, J. K., He, X., Swartz, M. E., Yan, Y.-L., Song, H., Boling, T. C., Kunerth, A. K., Walker, M. B., Kimmel, C. B., and Postlethwait, J. H. "MicroRNA Mirn140 modulates Pdgf signaling during palatogenesis". In: *Nature Genetics* 40.3 (Mar. 2008), pp. 290–298. DOI: 10.1038/ng.82.

[36] Li, L., Meng, T., Jia, Z., Zhu, G., and Shi, B. "Single nucleotide polymorphism associated with nonsyndromic cleft palate influences the processing of miR-140". In: *American Journal of Medical Genetics Part A* 152A.4 (2010), pp. 856–862. DOI: 10.1002/ajmg.a.33236.

[37] Tan, H., Onichtchouk, D., and Winata, C. "DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish". In: *Zebrafish* 13.1 (Feb. 2016), pp. 54–60. DOI: 10.1089/zeb.2015.1179.

[38] Larson, G. and Bradley, D. G. "How Much Is That in Dog Years? The Advent of Canine Population Genomics". In: *PLoS Genetics* 10.1 (Jan. 16, 2014), e1004093. DOI: 10.1371/journal.pgen.1004093.

[39] Shannon, L. M., Boyko, R. H., Castelhano, M., Corey, E., Hayward, J. J., McLean, C., White, M. E., Abi Said, M., Anita, B. A., Bondjengo, N. I., Calero, J., Galov, A., Hedimbi, M., Imam, B., Khalap, R., Lally, D., Masta, A., Oliveira, K. C., Pérez, L., Randall, J., Tam, N. M., Trujillo-Cornejo,

F. J., Valeriano, C., Sutter, N. B., Todhunter, R. J., Bustamante, C. D., and Boyko, A. R. "Genetic structure in village dogs reveals a Central Asian domestication origin". In: *Proceedings of the National Academy of Sciences* 112.44 (Nov. 3, 2015), pp. 13639–13644. DOI: 10.1073/pnas.1516215112.

[40] Ostrander, E. A., Wang, G.-D., Larson, G., vonHoldt, B. M., Davis, B. W., Jagannathan, V., Hitte, C., Wayne, R. K., and Zhang, Y.-P. "Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health". In: *National Science Review* 6.4 (July 2019), pp. 810–824. DOI: 10.1093/nsr/nwz049.

[41] Parker, H. G., Kim, L. V., Sutter, N. B., Carlson, S., Lorentzen, T. D., Malek, T. B., Johnson, G. S., DeFrance, H. B., Ostrander, E. A., and Kruglyak, L. "Genetic Structure of the Purebred Domestic Dog". In: *Science* 304.5674 (May 21, 2004), pp. 1160–1164. DOI: 10.1126/science.1097406.

[42] Schiffman, J. D. and Breen, M. "Comparative oncology: what dogs and other species can teach us about humans with cancer". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1673 (July 19, 2015), p. 20140231. DOI: 10.1098/rstb.2014.0231.

[43] Moïse, N. "Inherited arrhythmias in the dog: potential experimental models of cardiac disease". In: *Cardiovascular Research* 44.1 (Oct. 1, 1999), pp. 37–46. DOI: 10.1016/S0008-6363(99)00198-4.

[44] Studzinski, C. M., Araujo, J. A., and Milgram, N. W. "The canine model of human cognitive aging and dementia: Pharmacological validity of the model for assessment of human cognitive-enhancing drugs". In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. CANINE MODEL OF COGNITIVE AGING: FURTHER DEVELOPMENTS AND PRACTICAL APPLICATIONS 29.3 (Mar. 1, 2005), pp. 489–498. DOI: 10.1016/j.pnpbp.2004.12.014.

[45] Dodman, N. H., Ginns, E. I., Shuster, L., Moon-Fanelli, A. A., Galdzicka, M., Zheng, J., Ruhe, A. L., and Neff, M. W. "Genomic Risk for Severe Canine Compulsive Disorder, a Dog Model of Human OCD." In: *International Journal of Applied Research in Veterinary Medicine* 14.1 (2016).

[46] Catchpole, B., Adams, J. P., Holder, A. L., Short, A. D., Ollier, W. E. R., and Kennedy, L. J. "Genetics of canine diabetes mellitus: Are the diabetes susceptibility genes identified in humans involved in breed susceptibility to diabetes mellitus in dogs?" In: *The Veterinary Journal* 195.2 (Feb. 1, 2013), pp. 139–147. DOI: 10.1016/j.tvjl.2012.11.013.

[47] Hoffman, J. M., Creevy, K. E., Franks, A., O'Neill, D. G., and Promislow, D. E. L. "The companion dog as a model for human aging and mortality". In: *Aging Cell* 17.3 (2018), e12737. DOI: 10.1111/acel.12737.

[48] Giani, A. M., Gallo, G. R., Gianfranceschi, L., and Formenti, G. "Long walk to genomics: History and current approaches to genome sequencing and assembly." In: *Computational and structural biotechnology journal* 18 (2020), pp. 9–19. DOI: 10.1016/j.csbj.2019.11.002.

[49] Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., deJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., and Lander, E. S. "Genome sequence, comparative analysis and haplotype structure of the domestic dog". In: *Nature* 438.7069 (Dec. 2005), pp. 803–819. DOI: 10.1038/nature04338.

[50] Wang, C., Wallerman, O., Arendt, M.-L., Sundström, E., Karlsson, Å., Nordin, J., Mäkeläinen, S., Pielberg, G. R., Hanson, J., Ohlsson, Å., Saellström, S., Rönnberg, H., Ljungvall, I., Häggström, J., Bergström, T. F., Hedhammar, Å., Meadows, J. R. S., and Lindblad-Toh, K. "A novel canine reference genome resolves genomic architecture and uncovers transcript complexity". In: *Communications Biology* 4.1 (Feb. 10, 2021), pp. 1–11. DOI: 10.1038/s42003-021-01698-x.

[51] Goodstadt, L. and Ponting, C. P. "Phylogenetic Reconstruction of Orthology, Paralogy, and Conserved Synteny for Dog and Human". In: *PLOS Computational Biology* 2.9 (Sept. 29, 2006), e133. DOI: 10.1371/journal.pcbi.0020133.

[52] Steenbeek, F. G. van, Hytönen, M. K., Leegwater, P. A. J., and Lohi, H. "The canine era: the rise of a biomedical model." In: *Animal Genetics* 47.5 (Oct. 2016), pp. 519–527. DOI: 10.1111/age.12460.

[53] Fuller, J. C., Khoueiry, P., Dinkel, H., Forslund, K., Stamatakis, A., Barry, J., Budd, A., Soldatos, T. G., Linssen, K., and Rajput, A. M. "Biggest challenges in bioinformatics". In: *EMBO reports* 14.4 (2013), pp. 302–304.

[54] Molloy, J. C. "The open knowledge foundation: open data means better science". In: *PLoS biology* 9.12 (2011), e1001195.

[55] Kodama, Y., Shumway, M., and Leinonen, R. "The Sequence Read Archive: explosive growth of sequencing data." In: *Nucleic Acids Research* 40 (Database issue Jan. 2012), pp. D54–6. DOI: 10.1093/nar/gkr854.

[56] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., and Cochrane, G. "The European Nucleotide Archive". In: *Nucleic Acids Research* 39 (Database issue Jan. 2011), pp. D28–D31. DOI: 10.1093/nar/gkq967.

[57] Edgar, R., Domrachev, M., and Lash, A. E. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic Acids Research* 30.1 (Jan. 1, 2002), pp. 207–210. DOI: 10.1093/nar/30.1.207.

[58] Alnasir, J. and Shanahan, H. P. "Investigation into the annotation of protocol sequencing steps in the sequence read archive". In: *GigaScience* 4 (2015), p. 23. DOI: 10.1186/s13742-015-0064-7.

[59] Consortium, E. P. "The ENCODE (encyclopedia of DNA elements) project." In: *Science* 306.5696 (Oct. 22, 2004), pp. 636–640. DOI: 10.1126/science.1105136.

[60] Washington, N. L., Stinson, E., Perry, M. D., Ruzanov, P., Contrino, S., Smith, R., Zha, Z., Lyne, R., Carr, A., Lloyd, P., et al. "The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details". In: *Database* 2011 (2011).

[61] Abugessaisa, I., Shimoji, H., Sahin, S., Kondo, A., Harshbarger, J., Lizio, M., Hayashizaki, Y., Carninci, P., consortium, F., Forrest, A., Kasukawa, T., and Kawaji, H. "FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki." In: *Database: the Journal of Biological Databases and Curation* 2016 (July 9, 2016). DOI: 10.1093/database/baw105.

[62] Hong, E. L., Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., et al. "Principles of metadata organization at the ENCODE data coordination center". In: *Database* 2016 (2016).

[63] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3 (Mar. 15, 2016), p. 160018. DOI: 10.1038/sdata.2016.18.

[64] Morrison, M., Mourby, M., Gowans, H., Coy, S., and Kaye, J. "Governance of research consortia: challenges of implementing Responsible Research and Innovation within Europe". In: *Life Sciences, Society and Policy* 16 (Nov. 16, 2020), p. 13. DOI: 10.1186/s40504-020-00109-z.

[65] Adiconis, X., Haber, A. L., Simmons, S. K., Levy Moonshine, A., Ji, Z., Busby, M. A., Shi, X., Jacques, J., Lancaster, M. A., Pan, J. Q., Regev, A., and Levin, J. Z. "Comprehensive comparative analysis of 5'-end RNA-sequencing methods". In: *Nature Methods* 15.7 (July 2018), pp. 505–511. DOI: 10.1038/s41592-018-0014-2.

[66] Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. "CAGE: cap analysis of gene expression". In: *Nature methods* 3.3 (2006), pp. 211–222.

[67] Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., and Wingett, S. *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics Cambridge, United Kingdom, 2010.

[68] Kawaji, H., Lizio, M., Itoh, M., Kanamori-Katayama, M., Kaiho, A., Nishiyori-Sueki, H., Shin, J. W., Kojima-Ishiyama, M., Kawano, M., Murata, M., Ninomiya-Fukuda, N., Ishikawa-Kato, S., Nagao-Sato, S., Noma, S., Hayashizaki, Y., Forrest, A. R., and Carninci, P. "Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing". In: *Genome Research* 24.4 (Apr. 2014), pp. 708–717. DOI: 10.1101/gr.156232.113.

[69] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." In: *Genome Biology* 10.3 (Mar. 4, 2009), R25. DOI: 10.1186/gb-2009-10-3-r25.

[70] Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R., and Sandelin, A. "CAGEfightR: analysis of 5'-end data using R/Bioconductor". In: *BMC bioinformatics* 20.1 (2019), pp. 1–13.

[71] Severin, J., Lizio, M., Harshbarger, J., Kawaji, H., Daub, C. O., Hayashizaki, Y., Bertin, N., and Forrest, A. R. "Interactive visualization and analysis of large-scale sequencing datasets using ZENBU". In: *Nature biotechnology* 32.3 (2014), pp. 217–219.

[72] *Picard toolkit*. 2019.

[73] Kim, D., Langmead, B., and Salzberg, S. L. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature Methods* 12.4 (Apr. 2015), pp. 357–360. DOI: 10.1038/nmeth.3317.

[74] Dao, L. T., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., et al. "Genome-wide characterization of mammalian promoters with distal enhancer functions". In: *Nature genetics* 49.7 (2017), pp. 1073–1081.

[75] Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., et al. "A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells". In: *Nature methods* 14.6 (2017), pp. 629–635.

[76] Li, W., Notani, D., and Rosenfeld, M. G. "Enhancers as non-coding RNA transcription units: recent insights and future perspectives". In: *Nature Reviews Genetics* 17.4 (Apr. 2016), pp. 207–223. DOI: 10.1038/nrg.2016.4.

[77] Brooks Jr, F. P. *The mythical man-month: essays on software engineering*. Pearson Education, 1995.

[78] Methods, N. "Method of the year 2013". In: *Nat Methods* 11.1 (2014), pp. 1–1.

[79] Teichmann, S. and Efremova, M. "Method of the Year 2019: single-cell multimodal omics". In: *Nat. Methods* 17.1 (2020), p. 2020.

[80] Furey, T. S. "ChIP-seq and Beyond: new and improved methodologies to detect and characterize protein-DNA interactions". In: *Nature reviews. Genetics* 13.12 (Dec. 2012), pp. 840–852. DOI: 10.1038/nrg3306.

[81] Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., Van Der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., et al. "Molecular architecture of the mouse nervous system". In: *Cell* 174.4 (2018), pp. 999–1014.

[82] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., Bruggen, D. van, Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. "RNA velocity of single cells". In: *Nature* 560.7719 (Aug. 2018), pp. 494–498. DOI: 10.1038/s41586-018-0414-6.

[83] Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing". In: *Science* 348.6237 (2015), pp. 910–914.

[84] McGarvey, A. C., Kopp, W., Vučićević, D., Kempfer, R., Mattonet, K., Hirsekorn, A., Bilić, I., Trinks, A., Merks, A. M., Panáková, D., Pombo, A., Akalin, A., Junker, J. P., Stainier, D. Y. R., Garfield, D., Ohler, U., and Lacadie, S. A. *Single-cell-resolved dynamics of chromatin architecture delineate cell and regulatory states in wildtype and cloche/npas4l mutant zebrafish embryos*. bioRxiv, June 26, 2020, p. 2020.06.26.173377. DOI: 10.1101/2020.06.26.173377.

[85] Kouno, T., Moody, J., Kwon, A. T.-J., Shibayama, Y., Kato, S., Huang, Y., Böttcher, M., Motakis, E., Mendez, M., Severin, J., Luginbühl, J., Abugessaisa, I., Hasegawa, A., Takizawa, S., Arakawa, T., Furuno, M., Ramalingam, N., West, J., Suzuki, H., Kasukawa, T., Lassmann, T., Hon, C.-C., Arner, E., Carninci, P., Plessy, C., and Shin, J. W. "C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution". In: *Nature Communications* 10 (Jan. 21, 2019), p. 360. DOI: 10.1038/s41467-018-08126-5.

[86] Bartosovic, M., Kabbe, M., and Castelo-Branco, G. "Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues". In: *Nature Biotechnology* 39.7 (July 2021), pp. 825–835. DOI: 10.1038/s41587-021-00869-9.

[87] Authority (EFSA), E. F. S. "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related to the aspects of the biology and welfare of animals used for experimental and other scientific purposes". In: *EFSA Journal* 3.12 (2005), p. 292. DOI: 10.2903/j.efsa.2005.292.