ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

Master Program in Molecular Biology and Biomedicine

Master thesis

Establishment of a multi-omic strategy for the identification of enhancer-gene regulatory networks in model organisms

Εδραίωση πολυ-ωμικών μεθοδολογιών για τον προσδιορισμό ρυθμιστικών δικτύων ενισχυτών-γονιδίων στόχων σε οργανισμούς-μοντέλα

by **Myrto Mitletton**

Supervisor: Matthieu Lavigne
MSc committee: Matthieu Lavigne, Anastasios Pavlopoulos, Christos Delidakis

Gene Control Mechanisms Lab, IMBB, FORTH, Crete
in collaboration with Developmental Morphogenesis Lab, IMBB, FORTH, Crete

Heraklion 2023

Abstract

Ever since the emergence of Next Generation Sequencing techniques, -omics and multi-omics strategies have been offering amazing insights in structural and functional annotation of genomes and have boosted the development of both research and therapeutic approaches.
The first goal of this project was the construction and implementation of a pipeline for the analysis of ATAC-seq (chromatin accessibility) and RNA-seq (gene expression) data, enabling genome-wide enhancer identification and prediction of putative enhancer–gene links. The above pipeline was first established to analyze data derived from wild-type mice (*Mus musculus*) and mice with a gene knocked-out.
Once established and tested on the mouse dataset, the pipeline was used for the analysis of a developmental time-course dataset derived from the crustacean model organisms *Parhyale hawaiensis*. *Parhyale* has been put forward as an attractive experimental model to study tissue and organ morphogenesis during normal development and regeneration. I analyzed already produced mRNA-seq and Omni-ATAC-seq datasets, acquired at different stages of *Parhyale* embryogenesis: S13, S17 and S19.
My analyses identified Differentially Expressed Genes and Differentially Accessible Regions between these experimental conditions and generated hypotheses about the composition of enhancer-gene regulatory networks. I then selected three *Parhyale* genes for detailed analysis by quantitative RT-PCR and immunofluorescence: gooseberry (gsb), homothorax (hth) and lola-like (lolal), which were shown to have correlating expression and accessibility characteristics and implicated in developmental processes like segmentation, appendage formation and Hox gene regulation, respectively.
The established pipeline provides great amounts of data available for future research in mouse and *Parhyale*, and can be easily adjusted for data analysis of other model organisms and types of experiments.

Περίληψη

Από την απαρχή της ανάπτυξης των τεχνικών Αλληλούχισης Νέας Γενιάς (Next Generation Sequencing-NGS), οι–ωμικές και πολυ-ωμικές προσεγγίσεις έχουν αποδώσει πληθώρα δεδομένων σχετικά με την δομή και τη λειτουργία των γονιδιωμάτων και των στοιχείων τους και έχουν ωθήσει την ανάπτυξη τόσο ερευνητικών, όσο και θεραπευτικών πρακτικών σε ποικίλους τομείς.

Πρωταρχικός στόχος της παρούσας διπλωματικής εργασίας ήταν η δημιουργία μίας μεθοδολογίας (pipeline) για την ανάλυση δεδομένων προσβασιμότητας χρωματίνης (ATAC-seq) και έκφρασης γονιδίων (RNA-seq), η οποία θα επιτρέψει τον προσδιορισμό ενισχυτών και των γονιδίων στόχων τους σε γονιδιώματικό επίπεδο.Η μεθοδολογία αναπτύχθηκε αρχικά mRNA και ATAC-seq δεδομένα ποντικιού, που απομονώθηκαν από ποντίκια αγρίου τύπου (WT) και ποντίκια τα οποία έχουν υποστεί knock-out (KO) σε ένα γονίδιο ενδιαφέροντος.

Μετά την εφραμογή της στα παραπάνω δεδομένα, η μεθοδολογίας χρησιμοποιήθηκε για την ανάλυση δεδομένων εμβρυϊκής αναπτυξης του οργανισμού-μοντέλου *Parhyale hawaiensis*. Το καρκινοειδές *P. hawaiensis* είναι ένας ανερχόμενος οργανισμός-μοντέλο στη μελέτη της ανάπτυξης και μορφογένεσης, τόσο σε φυσιολογικές αναπτυξιακές συνθήκες, όσο και σε συνθήκες αναγέννησης ιστών. Στην παρούσα εργασία έγινε ανάλυση mRNA-seq and Omni-ATAC-seq δεδομένων από τα παρακάτω στάδια εμβρυικής ανάπτυξης του *P. hawaiensis*: S13, S17 and S19. Από τις αναλύσεις, έγινε προσδιορισμός διαφορικώς εκφραζόμενων γονιδίων και διαφορικώς προσβάσιμων χρωματινικών περιοχών μεταξύ των διαφορετικών πειραματικών συνθηκών και έγιναν προσπάθειες πρόβλεψης ρυθμιστικών δικτύων ενισχυτών-γονιδίων στόχων.

Εν συνεχεία, στον *P. hawaiensis*, επιλέχθηκαν τα παρακάτω τρία γονίδια για μία πιο λεπτομερή ανάλυση μέσω ποσοτικής PCR σε πραγματικό χρόνο (qPCR) και ανασοϊστοχημείας: gooseberry (gsb), homothorax (hth) και lola-like (lolal). Τα γονίδια αυτά εμφάνισαν θετική συσχέτιση μεταξύ Omni-ATAC-seq και RNA-seq δεδομένων και φάνηκαν να κατέχουν σημαντικό ρόλο σε ποικιλία αναπτυξιακών διαδικασιών όπως ο μεταμερισμός, η ανάπτυξη των άκρων και η ρύθμιση της έκφρασης των ομοιωτικών γονιδίων αντίστοιχα.

Η μεθοδολογία που δημιουργήθηκε απέδωσε μεγάλο αριθμό δεδομένων που μπορούν να χρησιμοποιηθούν για την ανάπτυξη ποικίλων επερχόμενων πειραμάτων στο ποντίκι και στον *P. hawaiensis*, ενώ με τις απαραιτητες προσαρμογές, μπορεί να χρησιμοποιηθεί για την ανάλυση δεδομένων και άλλων οργανισμών-μοντέλων.

Acknowledgements

I would like to deeply thank Matthieu Lavigne for giving me the opportunity to do my master thesis in his lab, for all his guidance and support and for believing in me until the end. You helped me grow both as a scientist and as a person and you sent me to explore and learn new things and acquire new skills. For that, I am very grateful.
Lots of thanks to Nektarios Belmezos for all his help during the first steps of this thesis when I did not have any prior experience with coding or the linux system.
I would like to thank Anastasios Pavlopoulos for trusting me with the analysis of the Parhyale data, for letting me do the validation experiments in his lab and for all his help with the experiments.
I thank Valia Stamataki and Niovi Rafailidou for their help with the Parhyale embryo dissections and the primer design; especially Niovi for all the long hours she spent with me for the acquisition of the confocal microscopy images of my data. I also thank Irini Karapidaki for her help with the end point PCR and John Rallis for the annotation files and his useful tips throughout the process.
I would like to thank Panayotis Verginis and Miranta Papadopoulou for trusting me with their data in the mouse and Christos Delidakis for being a member of my MSc committee.
Lots of thanks to Panagiotis Ioannidis for giving me access and helping me with the Apollo Genome Browser and the Chrysalida blasting tool.
I would like to thank all members of Lavigne and Pavlopoulos Labs, along with the members of Ntini Lab for their support and all the fun we had this past year.

But most importantly I would like to express my gratitude to my family and friends. They were there for me in my highest and lowest points, supported me when I needed it the most and believed in me all this time, even when I didn't believe in myself. None of this would be possible without your love, your hugs, your support and all the good times we had together.
Lots of love to my parents and my sister, who always believed in me, never backed down and supported me in my wildest decisions and craziest moments.
Lots of love to Nikiforos Bantounas, my most beloved person in this whole world, who has been my anchor and greatest supporter through all this time. I love you so much.

Contents

Introduction

Next Generation Sequencing Techniques

Starting from Sanger sequencing in 1977, the deciphering of nucleotide sequences has revolutionized the fields of biology and medicine. The development of Next Generation Sequencing (NGS) techniques that followed, which had the ability to read short or long nucleotides sequences from multiple molecules in parallel, in an effective and cost-efficient way, led to a massive production of data that boosted research and enriched clinical applications. NGS provided us with new prognostic, diagnostic and therapeutic tools, which enabled the identification of genes and mutations responsible for a variety of diseases, the evolution of precision medicine -especially in cancer-, the use of SNPs for population genetics and for the study of hereditary diseases etc. *(Qin D. 2019)*. In research, Whole Genome Sequencing of model organisms and pathogens led to the uplifting of the fields of OMICS as it helped with research in chromatin structure (Genomics) and epigenetics (Epigenomics), gene expression and regulation (Transcriptomics), chromatin-protein and protein-protein interactions (Proteomics), metabolism (Metabolomics) etc. Many different tools and techniques have evolved since the development of NGS for the examination of the different questions. For this project, RNA-seq and ATAC-seq will be analyzed.

Total RNA Sequencing (RNA-seq) is an assay for the assessment of the transcriptome, that is the total number of sequences derived from mature mRNAs and transient non A-tailed RNAs -such as eRNAs, or nRNAs. During this process, the RNA of cells or tissues is isolated and turned into a cDNA library (Fig. 1A). The fragments are then sequenced and the reads are aligned to the reference genome or transcriptome. We use either the total RNA of a tissue (bulk RNA sequencing), or some of the RNA molecules can be filtered out, depending on the aim of the experiment. A common filtering includes the isolation of mRNA for the construction of the library with the use of oligo$(dT)_{18}$ primers. Those primers bind only to the poly-A tails, which are unique to processed complete mRNA molecules.

RNA-seq allows the identification of the genes that are expressed in a specific tissue or condition and the quantification of their expression levels. The gene expression is very easily quantified with the calculation of the number of reads that are mapped to each gene or coding region. The assay allows the assessment of a large number of genes, even entire transcriptomes, simultaneously and it is highly sensitive for both low and high levels of gene expression. As the reads are sequenced and mapped to the reference genome the process is capable of providing single-base resolution *(Wang Z. et al. 2009)*.

Apart from gene expression levels, RNA-seq can provide information on isoforms and alternative splicing mechanisms, since all different exon combinations of a specific cell identity are present on the total RNA of the tissue. The process is also important for the examination of the production of other coding regions -like small or non-coding RNAs-, which offer insights in multiple levels of cell function, like cell homeostasis (rRNA), transcription regulation (eRNAs), cell defense (miRNAs) etc. *(Stark R. et al. 2019)*.

Apart from studying the cells or tissues in a single specific status, we can also compare quantitative expression data between two or more experimental conditions. That allows for the examination and comparison of conditions that occur naturally in organisms -

like during embryonic and organ development - or that are a result of experimental manipulation -like the knockout of genes, chemical manipulation or other type of treatments. Several tools have been established that allow the comparative analysis of RNA-seq data between multiple conditions. That way, we can enrich our understanding of cell function, as well as study factors that cause cellular processes to be disturbed and lead to diseases or other malfunctions *(Stark R. et al. 2019)*.

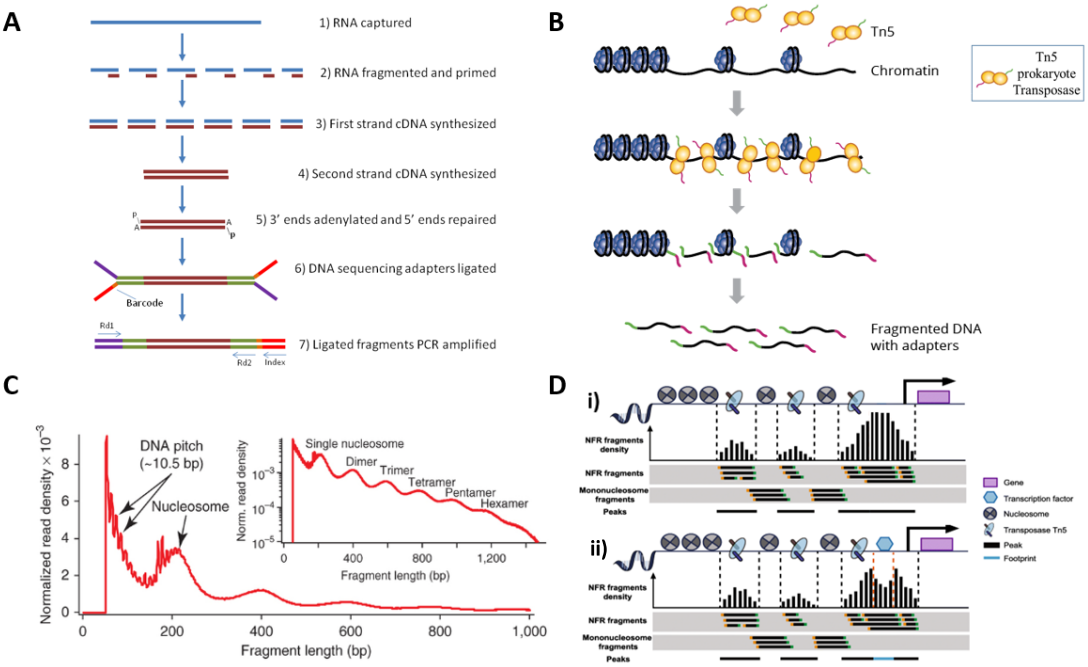


Figure 1. (A) Summary of RNA sequencing. The RNA molecules are fragmented and adaptors are ligated to the fragment's ends. (B) Summary of ATAC sequencing. A hyperactive Tn5 dimer loaded with sequencing adaptors cuts accessible genomic sites and ligates the adaptors at the fragment's blunt ends. (C) A typical fragment size histogram for ATAC-seq data. We expect a large peak at ~100bp that corresponds to inter-nucleosomal regions and subsequent smaller peaks at ~200, 400, 600bp for mono-, di-, tri-nucleosomal regions respectively. (D) i) when chromatin is accessible Tn5 produces fragments that form a peak ii) when a protein binds to the DNA the site is partially protected, less fragments are created and a unique peak formation called 'footprint' is produced.

The Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) is an assay used for the assessment of chromatin accessibility across the genome. It utilizes Tn5, a prokaryotic transposase of a class II transposon that uses the "cut and paste mechanism". In ATAC-seq, Tn5 has been modified to become hyperactive and is loaded with sequencing adaptors. During the process, Tn5 dimerizes, cuts accessible genomic sites and ligates the adaptors at the fragment's blunt ends *(Buenrostro JD. et al., 2013)*. After the repair of the blunt ends, the fragments end up with short repeats of 9bp at both ends *(Berg 1989)* (Fig. 1B). That 9bp nucleotide sequence varies across different organisms *(Zhang H. et al. 2021)*, but a consensus target site is A-GNTYWRANC-T, (where N=all 4 bases, Y=T or C, W = A or T, and R =A or G) *(Goryshin I.Y. et al. 1998)*. The cleavage sites of Tn5 depend on the chromatin landscape, that is, the presence of nucleosomes and proteins *(Li Z. et al. 2019)*. Regions of the genome that are tightly compacted (high nucleosome concentrations) are less accessible due to steric hindrance and are therefore less likely to be cut. As a result, Tn5 targets primarily open chromatin sites *(Buenrostro JD. et al., 2013)*. The library produced by the accessible chromatin reads is then sequenced and aligned to the reference genome. Genomic regions enriched with reads are called peaks and are identified as open chromatin regions.

In each cell state, a specific set of genes is accessible to the transcription machinery and it is thus expressed. In those cells, coding regions are only nucleosome-free when the

RNA polymerase is present at the site of transcription, but some of their total regulatory regions remain in an open state *(Klemm S. L. et al. 2019)*. As a result, ATAC-seq can provide information mostly for the identification of cis-regulatory elements, not the genes that are expressed and is widely used for enhancer landscaping and for the determination of accessibility changes between different experimental conditions (e.g. developmental stages, healthy and disease conditions etc.). Apart from that, it provides information concerning the positioning of nucleosomes and higher levels of compaction of the DNA. More specifically, the fragment size can be indicative of the number of nucleosomes nearby as longer fragment sizes are produced by nucleosome-richer regions. In regions not covered bynucleosomes, Tn5 produces reads as small as less than 100bp , but in the presence of mono-, di- and tri- nucleosomes the produced fragments can vary in length ~200, 400 and 600bp respectively (as a results of ~146bp being packed around each nucleosome) (Fig. 1C). Because ATAC-seq does not require size selection during library preparation, nucleosome landscapes can be determined. However, the process is not as effective as with data from MNase-seq due to the decreased coverage observed outside the open chromatin regions *(Yan F. et al., 2020)*.
At a DNA-protein interaction level, ATACseq is used to impute patterns of transcription factors' (TF) binding, a process called "footprinting". When a TF is bound to an open chromatin region of the DNA, that site is partially protected from Tn5 by the TF and is not nicked - at least not as much as it would be if the site was not bound by a protein. As a result, the binding sites when a TF is bound to them have a unique peak formation, often called "footprint" (and so the process is called "footprinting") (Fig. 1D) *(Buenrostro JD. et al., 2013)*. TF footprinting requires high resolution ATAC-seq data with read depth in the range of 200 M reads per replicate for mammalian genomes.

ATACseq has several advantages when compared to other assays used for the identification of regulatory regions or the epigenetic landscapes (e.g. DNAse-seq, FAIRE-seq etc). Firstly, the whole procedure is completed in only two steps. That results in the reduction of experimental time to several hours (instead of days) and the decrease of error probability. Furthermore, the process can be carried out effectively with 5.000 nuclei, a relatively small number when considering that other processes require a minimum of 1 million cells (FAIRE) or 50 million cells (DNAse -seq) as input. In fact, the process can be done even on as little as 500 nuclei but sensitivity decreases noticeably *(Buenrostro JD. et al., 2013)*. Additionally, ATACseq is suitable for paired-end sequencing and single-cell analyses, which gives us the opportunity to study open chromatin and gene regulation profiles for each particular cell in a sample *(Buenrostro JD et al., 2015)*.

As ATAC-seq provides such information-rich results, it is not difficult to imagine that there have been efforts for the improvement of the process. Omni-ATAC-seq includes additional protocol steps, such as the use of detergents (NP40, Tween-20 and digitonin) and PBS, which aim to remove mitochondria and improve signal-to-noise ratio. Mitochondria lack chromatin packaging and their DNA is widely accessible. As a result, Tn5 produces numerous reads with high coverage that decrease the overall signal-to-noise ratio. The additional steps remove mitochondria and increase cell permeabilization resulting in a higher chromatin/mitochondrial reads percentages. The

extracted data are therefore of better quality because they have decreased background and lead to improved signal-to-noise ratios. The decreased background offers higher numbers of greater confidence peaks (true peaks) and provides the opportunity to use the technique in cells and tissues that have a lot of noise (e.g. snap frozen cells, human keratinocytes etc.). Finally yet importantly, less total reads (due to the lack of mitochondrial reads) means lower sequencing costs *(Corces M. et al. 2017)*. Although these improvement steps provide a great number of benefits, the original ATAC-seq process is still widely used.

Typically, NGS technologies can sequence up to 300bp before the per base quality drops significantly *(Pervez M.T. et al. 2022)*. Although the sequencing of maximum 300bp of one end of a fragment (single-end sequencing) is generally enough to find their location on the genome with accuracy, it also results in some loss of information. Additionally, in some genomic regions that have duplications or are rich with repetitive elements there is a chance that a read of 300bp would be mapped to more than one positions on the reference genome. An alternative method is paired-end sequencing. In paired-end sequencing, a fragment is sequenced from both sides and the produced reads consist of a read pair. The reads are then aligned at the reference genome. As the distance between the reads is known, they are more precisely aligned because the probability of both reads being able to align to more than one position is lower. Paired-end sequencing is therefore more informative, facilitates alignment in regions with repetitive elements and allows the detection of insertions and deletions in a genome sequence *(Illumina Inc 2017)*.

The model organism *Parhyale hawaiensis*

Parhyale hawaiensis is a marine amphipod crustacean of the class of malacostraca that has only recently become a very promising model organism. Malacostracan crustaceans live all around the world, in both marine and freshwater environments and include groups of species of high economical and nutritional importance such as crabs, prawns, shrimps and lobsters *(Kao D. et al. 2016)*. Parhyale itself is found at circumtropical shallow intertidal coastlines around the world, in rocky substrates and its popular habitats include bays, estuaries and mangroves *(Paris M et al. 2022)*.

Parhyale has multiple characteristics that make it an attractiveanimal model. It is small in size and it is easy and cheap to grow in the lab. The animal has a relatively short life cycle and it can be easily cultivated in artificial seawater, at an ideal temperature of 26°C. It can be fed with common fish food like fish flakes, kelp powder, carrots, and pellets formulated for feeding shrimp. Additional requirements include the existence of a proper substrate that can consist of aragonite rock or crushed coral *(Paris M et al. 2022)*. Parhyale's genome is sequenced and annotated and multiple techniques have been established for its study and manipulation. Those include gene editing and transgenesis technologies -such as RNAi, the Minos system, morpholino, CRISPR etc.- and staining and in situ hybridization techniques. The availability of a variety of genetic markers and drivers, along with the partial transparency of Parhyale's body has made it possible to perform live imaging as well as and cell lineages analyses *(Paris M et al. 2022)*.

Parhyale is widely used for the study of embryonic development -especially limb formation and regeneration-, cellulose digestion, ecotoxicity and evolutionary and comparative biology.
Parhyale has become a very good model for developmental biology for several reasons. Its development is direct, its embryogenesis lasts ~10 days (Fig. 2A) and the females can produce embryos every two weeks. After fertilization, the eggs are found in a ventral brood pouch externally of the females' bodies, from where they can be very easily isolated. The eggs can be grown outside of the pouch in Filtered Artificial Sea Water with Antibiotics (FASWA) at 26℃. A female can produce up to 25 eggs, which are synchronized and are large enough to allow for microinjections *(Browne et al. 2005, Gerberding et al. 2002)*. The embryo's developmental stages have been characterized in detail according to morphological criteria by Browne et al. 2005. Early cleavage is holoblastic (total), and at the 8-cell stage each cell is determined to give rise to a specific germ layer (Fig. 2A), allowing for experimental manipulation of early cells *(Gerberding et al. 2002, Kao et al. 2016)*. Parhyale is also capable of regenerating its appendages and can maintain that ability throughout its lifetime. Hence, it has become a very important model for limb development and regeneration that can help shed some light to the mechanisms of limb formation and their regulation.

Parhyale can also digest lignocellulose (wood), the most abundant raw material on earth. Lignocellulose digestion is very limited in Metazoans and is restricted to very few terrestrial species (e.g termites, ruminants and beetles) and a handful of marine species of the multicrustacea clade (e.g. *Limnoria quadripunctata* (isopod) and *Chelura terebrans* (amphipod) etc.). Unlike the above insects, that lack some enzymes necessary for wood digestion and rely on specified intestinal microbiota for the completion of the process, Parhyale -and other multicrustaceans- expresses its own glycosyl hydrolase enzymes of the GH7 family required for the hydrolytical digestion of cellulose *(Kao D. et al. 2016)*. Thus, studying Parhyale's metabolism can help us understand how lignocellulose is broken down and lead to advances in biofuel production and exploitation.
The animal has also been used in the field of ecotoxicology. Because of its distribution to tropical regions, its large abundance and the ease of cultivating it in the lab, it is becoming an important model for the measurement of both short- and long-term toxicity to tropical coastal ecosystems. These data, combined with results from genotoxicity and immunotoxicity assays, can help us draw conclusions on the repercussions of pollution and examine the way contaminants affect the survival, growth and reproduction of marine species and communities *(dos Santos et al. 2022)*.
Parhyale is a member of the Pancrustacean family, a group that includes Insects and Crustaceans. Since it has become a model organism with a sequenced genome it is very useful for comparative studies and evolutionary analyses. Having a member of a group that is paraphyletic to insects can be very useful for studying the emergence of patterns and adaptations among insects and crustaceans. Additionally, Parhyale can help decipher the relationships between the members of the Pancrustacean family, which are yet to be determined *(Kao D. et al. 2016)*.

Although *Parhyale hawaiensis* became a promising animal model at the early 2000s, its genome was sequenced several years later, in 2016, by Kao, Lai, Stamataki et al. Parhyale's genome comprises of 23 chromosome pairs and has an estimated size of 3.6 Gb, making it one of the largest genomes among the arthropods. It is characterized by a great number of repetitive elements, high heterozygosity and an increased gene size due to an expansion in intron length *(Kao et al., 2016)*. After its initial assembly, Parhyale's genome has been reassessed and improved, but because of its large size and the presence of repetitive sequences, its assembly remains a challenging task that is not yet completed. As a result, the latest version of Parhyale's genome (genomeV5.fa) consists of 278.189 contigs, with lengths ranging from 200 to 75.825.039 bp (Fig 2B). Parhyale's genome annotation was built from previously assembled transcriptomic data, gene homology and ab initio predictions in 2016 *(Kao et al. 2016)*. However, due to the genome assembly being incomplete, some genes were split between contigs and the corresponding gene predictions were also incomplete. In its latest version, the annotation has improved but it still has inaccuracies, such as scaffolds with wrong orientation and unresolved polymorphic repeats *(Paris M et al. 2022)*.

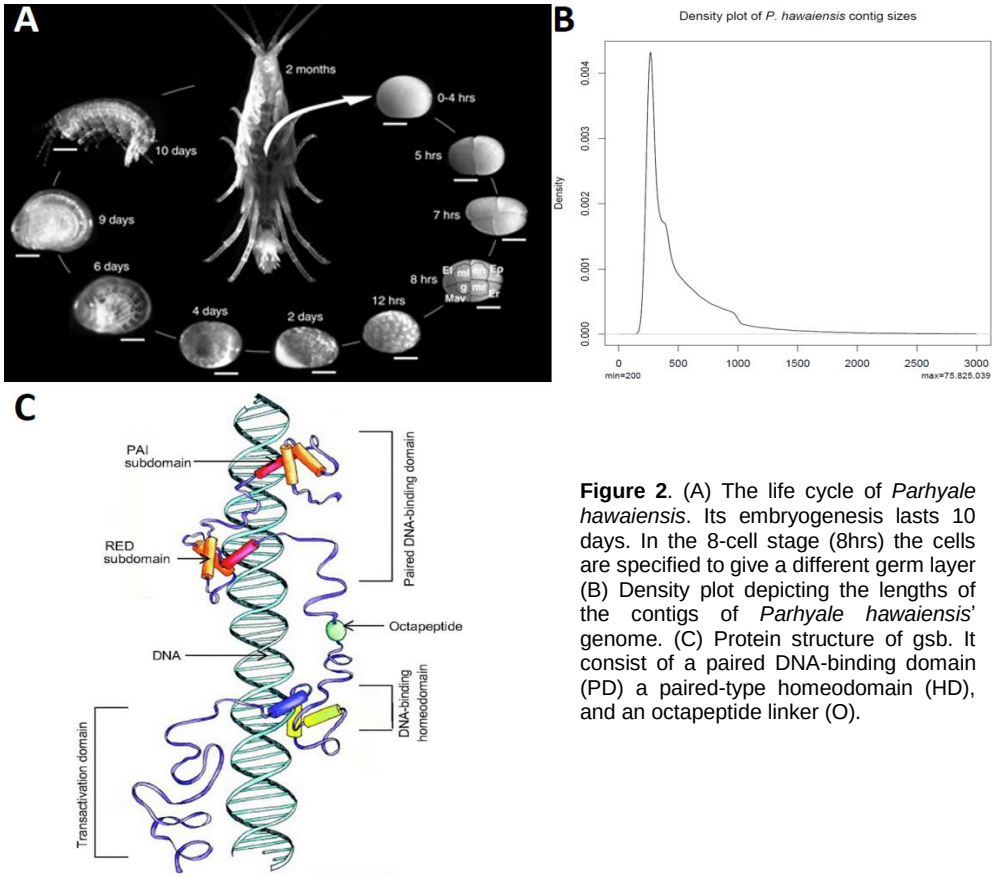


Figure 2. (A) The life cycle of *Parhyale hawaiensis*. Its embryogenesis lasts 10 days. In the 8-cell stage (8hrs) the cells are specified to give a different germ layer (B) Density plot depicting the lengths of the contigs of *Parhyale hawaiensis*' genome. (C) Protein structure of gsb. It consist of a paired DNA-binding domain (PD) a paired-type homeodomain (HD), and an octapeptide linker (O).

Genes of interest

Gooseberry (gsb) is a gene of the Pax 3/7 gene family (also known as Pax group III gene family). Pax 3/7 genes are transcription factors, highly conserved across Metazoans that are expressed during embryonic development. They consist of a 128 amino-acid paired-domain (PD) and a paired-type homeodomain (HD), linked together by an octapeptide (O) (Fig 2C). Other members of the family include paired (prd) and

gooseberry-neuro (gsbn) in Drosophila and Pax3 and Pax7 in vertebrates (humans, mice, zebra fish etc.) *(Thompson B. et al. 2021)*.
In Drosophila, gsb is a segment polarity gene that is first detected at the end of cellularization and is expressed at the posterior of each parasegment in 14 stripes (one-segment periodicity). After germ band extension, the gsb's signal in stripes 4-14 is limited to the neuroectoderm, where it activates gsbn *(Davis G.K. et al. 2001)*. prd, gsb and gsbn were derived from the same ancestral gene as a result of two duplication events. The first event gave rise to prd and the ancestor of gsb and gsbn. That ancestor was then submitted to a second duplication event to produce gsb and gsbn *(Balczarek et al. 1997)*. Supporting this hypothesis, prd and gsb roles appear to be interchangeable in Drosophila's embryogenesis. Specifically, when the prd coding region is placed under the control of gsb cis-regulatory elements it rescues gsb mutant effects and vice-versa *(Li and Noll 1994, Xue and Noll 1996)*.
In arthropods other than Drosophila, gsb has only been studied as part of the Pax 3/7 gene family and not individually. More specifically, Davis G.K. et al. 2005 made the two monoclonal antibodies (Abs) DP311 and DP312, that cross react to prd, gsb and gsbn in Drosophila, to examine the role of Pax 3/7 genes in arthropod segmentation. Both Abs have a core epitope of 8-amino acids (PD(V/I)YTREE) that recognizes a large portion of helix 2 of the HD domain. The HD domain is not restricted to the members of the Pax3/7 family so the Abs did not specifically bind to them. However, because the expression of the non-Pax3/7 HD bearing proteins did not have a stripe pattern their signal did not interfere with the segmentation analysis. Their work showed that although there is some variety across the inspected arthropods (members of insects, chelicerates, crustaceans and myriapods), Pax3/7 are expressed in segmental stripes. gsb or other Pax3/7 family genes have not yet been studied in Parhyale, but considering that the Pax3/7 family members are conserved -not only among arthropods but also in crustaceans- it is very likely that they would play a role in segmentation and they will have some kind of stripe pattern.

Homothorax (hth) is a homeodomain transcription factor that has an α-helix with which it interacts with extradentical (exd). hth mediates exd's nuclear localization and together they form a cofactors complex *(Gramates L.S. et al. 2022)*. It has been shown that hth requires exd for many of its functions. More particularly, in Drosophila, when exd function is eliminated, the hth genotype appears to be lost *(Rieckhof et al. 1997)*. hth is important for multiple processes in fly development like appendage development and patterning and nervous system and eye morphogenesis *(Gramates L.S. et al. 2022)*.
hth has been studied in a variety of arthropods of all subphylums, where it was found to be expressed in the appendages and occasionally in the body wall through the proximal femur. The gene exists mostly in one copy, but two paralogs have been found in spiders *(Bruce H. S. 2017)*. Interestingly, although the gene is located in the legs of many arthropods, its expression pattern varies. In Drosophila, exd is expressed throughout the legs but hth is located in the proximal leg podomeres. In Parhyale, the same pattern occurs but hth expression is extended to the next podomere in the biramous appendages (uropods and pleopods). In contrast, the opposite seems to be the case in chelicerates and millipedes where hth is located all over the leg and exd is restricted. Therefore, the pattern was at some point of the arthropod evolution reversed *(Prpic N. M. et al. 2008)*.

Specifically for Parhyale, it was shown that hth is expressed in the head lobes (S12-S17) and expands towards the rest of the embryo as the animal develops. At S18, it is expressed in all head appendages apart from Mn and in ring-like patterns at the base of each thoracic limb. Eventually, it is located in the lateral body wall and the proximal parts of the limbs *(Bruce H. S. 2017)*.

Longitudinals lacking-like (or lola-like or lolal or batman or ban), is a BTB domain (or POZ domain) protein and a member of the Trithorax gene Group (TrG).
TrG members are mostly known for their role in histone modifications and chromatin remodeling, which result in the loosening of chromatin structures. Additionally, they are involved in transcription, either by being part of the transcription machinery or by interacting with its members. Because of their contribution to rendering chromatin more accessible, TrG are believed to be involved in transcription activation, but some of the proteins also have repressing functions *(Kingston R. E., Tamkun J. W., 2014)*.
In Drosophila, lolal is a maternal factor that is carried to the nucleus after the maternal-to-zygotic transition. It has been shown to serve a role in larval lymph gland hematopoiesis, salivary gland morphogenesis and, most importantly, the regulation of Hox genes *(Gramates L.S. et al. 2022)*. lolal is responsible for the repression of the expression of Scr and Ubx, but there have been indications that it might be involved in the Hox gene activation *(The UniProt Consortium, 2023)*.
lolal does not have a DNA binding domain, so its interactions with chromatin are most likely indirect, through its interaction with other proteins *(Brody T. 1999)*. Thus, the gene's dual role could be attributed to its interaction with Trithorax-like (or trl or GAGA factor), a TrG gene that also has both activation and repression roles in Hox genes. More specifically, it has been suggested that lolal is recruited to DNA by trl through the heterodimerization between the BTB/POZ domains of the two proteins. Together, they are both necessary for the maintenance of Scr and Ubx in a repressive state *(Faucheux M. et al. 2003)*.
lolal has not been studied extensively in arthropods. Phylogenetic studies have revealed that it is only present in Pancrustaceans and it has not been found in genomes of vertebrates or echinodermata. Other members of the lolal group are present in different arthropods subphyla and they are thought to have derived from a series of duplications during the arthropod evolution. More specifically, Bab is the oldest and is found in all subphyla. It generated Ttk (present in insects and chelicerates but lost in crustaceans). Ttk first generated lolal -found in Pancrustaceans- and later gave rise to Mmd4, which is only present in insects. Bab later generated Psq, also found in Pancrustaceans and Mmd4 generated Lola, which is also only present in insects *(Quijano J. C. et al. 2016)*.

Establishing the Pipeline

Aim of the project

The first aim of this project was the construction of a pipeline for the integrative analysis of chromatin accessibility (ATACseq) and gene expression (RNAseq) data for the identification of putative enhancers and the prediction of enhancer-gene regulatory networks (eGRNs). Consequently, the goal was to use this pipeline for the analysis of mRNA-seq and ATAC-seq datasets from different stages of *Parhyale hawaiensis*' embryonic development, in order to identify putative enhancer–gene links that play major roles in Parhyale's embryogenesis.

The pipeline was established through the analysis of a set of ATACseq and RNAseq data in the mouse *Mus musculus*, derived from a collaborative laboratory. Both types of data were produced from wild type (WT) mice and mice with a knock-out (KO) of a gene, in triplicates. The libraries were sequenced on the NextSeq500 platform (Illumina) and paired-end reads of 150bp were produced.

The goal of the pipeline was to process the data and perform a correlation analysis in order to determine differences in gene expression across the two conditions, as well as to identify regulatory networks that may contribute to those differences.

Quality control and trimming

Quality control (QC) is performed on the raw sequenced data (fastq files) in order to ensure that they are of appropriate quality and that no technical problems occurred during the library preparation procedure or the sequencing.

For the examination of the data, the FastQC tool *(Andrews S. 2010)* performs a comparison of the raw data with a 'normal' sample -which is considered random, diverse and follows a normal distribution- to establish whether the data deviate from the expected. In its report, the tool provides an evaluation of several parameters related to the quality of the data and the process, in addition to some information on the basic statistics of the data (e.g. number of sequences, length of reads, GC content etc.). The sequencing quality can be evaluated by multiple factors like the amounts of bases added by the sequencer with not sufficient confidence (number of Ns in the sequenced reads), the existence of enrichment bias -which is identified through the calculation of the amounts of duplicated sequences- and the per base quality score. For the latter, the tool retrieves the quality score or Phred score of each base from the fastq file, which expresses the probability of that base being wrong. A Phred score =20 indicates that the probability of incorrect base incorporation during sequencing is 1%. A base with Phred score <20 is considered of poor quality *(Andrews S. 2010)* (Fig. 3A).

Other important factors, such as biases that occurred during library preparation can be identified through the percentage of the bases in the reads. Last but not least, FastQC can provide information on potential contaminants or poor complexity in the data, through the calculation of GC content in the sequences and of the number of overrepresented sequences in the samples. More specifically, if the mean GC content does not follow a normal distribution (e.g. by having additional peaks etc.), it either indicates that more than one types of DNA are found in the sample and it is contaminated or that there are overrepresented sequences with a particular GC content spiking out of the distribution (Fig. 2B). If overrepresented sequences are identified it

may indicate that they come from a source different from that of the samples or that some specific sequence of the sample genome has been selected and amplified preferentially. In this case, sequences are subsequently run through databases of known contaminants for the determination of the contamination source, if any *(Andrews S., 2010)*.

Although in general deviance from the expected samples means low data quality, depending on the data's origin some biases might be expected. For example, in ATACseq data the Per Base Sequence Content (base percentage) *should* deviate from the expected due to the Tn5 bias in the 5'end of the read. Although it is not an absolute requirement, Tn5 has a preference in the target site for the motif A-GNTYWRANC-T, (where N=all 4 bases, Y=T or C, W=A or T, and R=A or G), which is reflected in the data *(Goryshin I.Y., 1998)* (Fig. 2C).

The ensemble of the low-quality data like low quality bases, reads with very small or large lengths (that do not provide any information or increase signal-to-noise-ratio), along with the sequences matching with sequencing adapters normally used in NGS facilities are removed from the data during the Trimming step. This step aims at the cleaning of the reads for a more efficient alignment to the reference genome.
For the QC and the trimming, I used the Trim Galore tool *(Krueger F. et al., 2021)*, a Perl wrapper around FastQC and the trimming tool Cutadapt, thus, combining two steps.

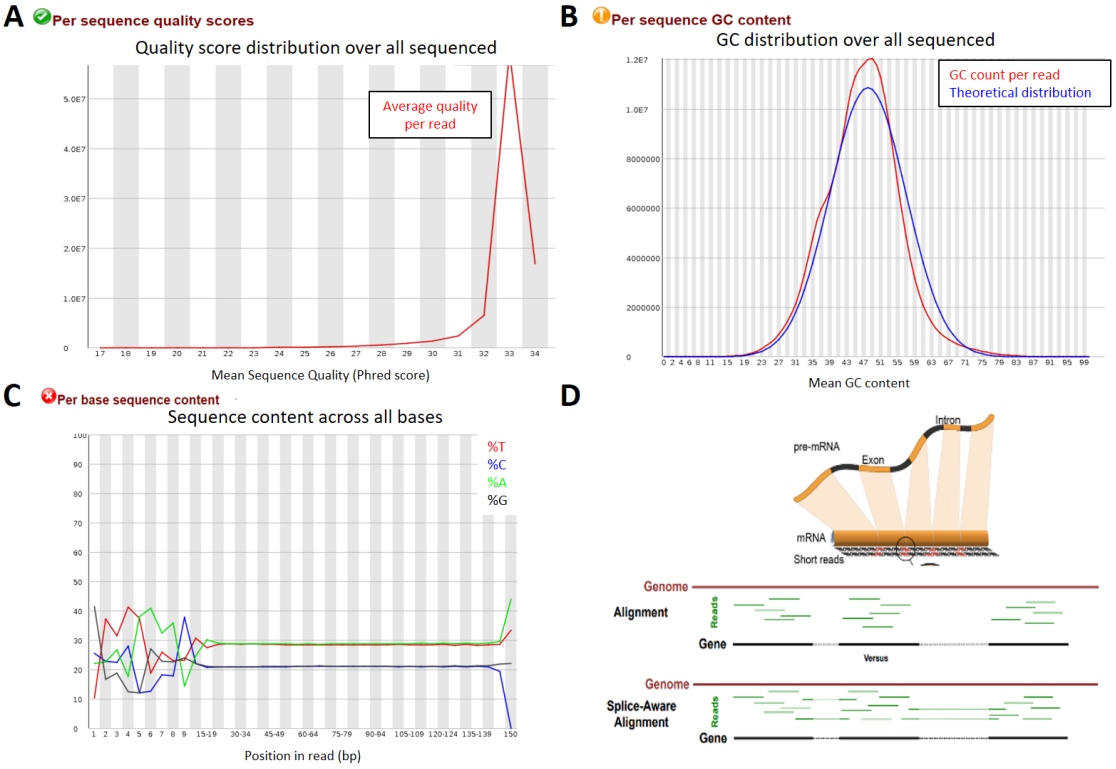


Figure 3. (A) Phred score distribution plot as part of a FastQC report. The Phred score expresses the probability of a base having been incorporated incorrectly during sequencing. A base with Phred score <20 is considered of poor quality. (B) GC% distribution plot as part of a FastQC report. The sample's GC curve is compared to a theoretical curve form a hypothetical sample that follows the normal distribution. A second peak could be indicative of sample contamination or the presence of overrepresented sequences. (C) Base content plot as part of a FastQC report. The base percentages are expected to be homogenous across the sample. Deviation from the expected base percentages is indicative of biases. The Tn5 preference motif is evident in the edge of each fragment. (D) Depiction of the difference between a splice-aware and a splice-unaware aligner. A read that comprises of an exon-exon junction is split and aligned in its corresponding exons by a splice-aware aligner but is discarded completely by a splice-unaware one.

Mapping

The next step is the mapping or alignment of the reads to the reference genome, that is, finding the corresponding part of the reference genome for each read.

There are two types of alignment strategies: the alignment-based method and the alignment–free method. In the first method, the algorithm looks for the correspondence of sequences in a base-to-base level, where the residues have to be identical and in the same order. Each base is classified as either a match or a mismatch, after having taken into consideration possible insertions or deletions (gaps) *(Zielezinski A. et al., 2017)*. Although this method is very computationally expensive, alignment-based tools are still widely used for a variety of reasons. They have been available longer and, as a result, most tools for downstream analysis and visualization purposes were made for their outputs. Additionally, the base resolution that they provide is very important in some subsequent processes like motif analyses and TF binding sites identification.

On the other hand, the alignment-free algorithm does not match individual residues, but quantifies the similarity and dissimilarity of the two sequences. Both sequences are divided in k-mers. The more k-mers a read shares with the reference sequence, the greater their similarity and the higher the chance it comes from that sequence. This process is also called "pseudoalignment" because no per base alignment is produced throughout the process. Because of the latter, the task is less expensive computationally and the overall process is performed faster but as effectively *(Zielezinski A. et al., 2017)*.

With the alignment-based method, reads can be mapped to either the genome or the transcriptome, while alignment-free methods map to transcriptome only. Alignment to transcriptome allows for gene expression detection in isoform level, because the transcriptome includes all different transcripts that can be produced by a coding region (if they are well characterized in the organism). It also allows for more accurate quantification of reads from RNA-seq data, as it is easier to count reads that comprise of an exon-exon junction etc.

On the other hand, mapping to genome allows for the integration of information about introns and exon-intron junctions in the analysis and can be performed by splice-aware or spice-unaware aligners. The splice-aware aligners take into account the introns' positions on each gene. As a result, if a read comprises of an exon-exon junction it is split and aligned in both exons *(Yi L., et al. 2018)*. On the other hand, in splice-unaware aligners such a read would be discarded from the alignment (Fig. 2D). Splice-aware aligners allow the identification of novel isoforms that are not included in the transcriptome. Finally, in the case of data not derived from RNAseq most reads are expected to be aligned to non-coding regions and mapping to genome is the only valid aligner choice. A summary of alignment tools are on table 1.

		To genome	To transcriptome
Alignment-based	splice aware	STAR	
		HISAT2	
	splice unaware	Bowtie2	Bowtie2
		BWA	BWA
Alignment-free			Salmon
			Kallisto

Table 1. A table summarizing the alignment tools.

The ATAC-seq data provide information for mostly non-coding regions and splice awareness is not essential for proper read mapping. Therefore, I used the splice-unaware aligner BWA *(Li H. 2013)* to map the reads to the genome. The RNA-seq data need a splice-aware aligner to ensure that intron position is taken into account, so HISAT2 *(Kim D. et al. 2019)* was used. Both sets of data were aligned to the mm10 version of the mouse genome. The mapping results are on tables 2 and 3 below:

ATAC Mapping Results	Number of reads	Mean Coverage	%GC	mean read length
WT1	489,121,088 (93.45% mapped, 6.55% unmapped)	12.4297	44.87%	111.1
WT2	525,846,210 (98.04% mapped, 1.96% unmapped)	14.2928	45.16%	111.09
WT3	294,535,192 (96.77% mapped, 3.23% unmapped)	7.8468	45.18%	111.09
KO1	360,415,798 (95.77% mapped, 4.23% unmapped)	8.6926	44.41%	111.11
KO2	286,017,394 (96.41% mapped, 3.59% unmapped)	6.9557	46.37%	111.08
KO3	474,998,810 (95.95% mapped, 4.05% unmapped)	11.0653	44.57%	111.06

Table 2. A summary of the mapping results of the mouse ATAC-seq data.

RNA Mapping Results	Number of reads	Mean Coverage	%GC	mean read length
WT1	45,815,870 (90.21% mapped, 9.79% unmapped)	16.8983	51%	118.35
WT2	39,402,306 (59.42%mapped, 40.58% unmapped)	10.1131	51.71%	115.52
WT3	62,037,258 (68.05% mapped, 31.95% unmapped)	15.1621	50.91%	118.1
KO1	43,236,628 (93.41% mapped, 6.59%unmapped)	15.3639	51.09%	118.12
KO2	35,484,702 (71.9% mapped, 28.1% unmapped)	9.6522	51.26%	116.6
KO3	28,617,494 (86.76% mapped, 13.24% unmapped)	9.9976	51.52%	118.86

Table 3. A summary of the mapping results of the mouse RNA-seq data.

In the RNA-seq data, the ratio of mapped/unmapped reads was a bit low in some of the samples, but the overall coverage was fine. Both datasets had similar mean read lengths, with the one from RNAseq being the most variable (ATAC-seq mean read length = 111.07, RNA-seq mean read length = 117.59). The read lengths are smaller than originally expected from the library (l=150bp) because of the trimming of the adaptors and the low quality bases from the previous step. For a more detailed overview, the fragment size plots are depicted in figure 4.

In paired-end RNA-seq data we expect the fragment sizes to be greater than twice the times of the read length *(Jaksik R. et al. 2021)*. In this case, longer than 2x115=230bp (Fig. 4A). We observed that the majority of fragment sizes in our RNA-seq data are > 230bp

as expected. But it is also normal to have fragment sizes smaller than 230bp in this graph, due to overlap of the paired read in smaller fragments..

In an ATAC-seq dataset we expect a large peaks at <100bp that represents the regions of open chromatin, a peak ~200bp for mono-nucleosomal regions and other smaller peaks for multi-nucleosomal regions *(Buenrostro J. et al. 2013)*. In our data, we only observe the peak for the open chromatin regions, which suggests that less nucleosomal regions than expected were captured by Tn5 (Fig. 4B).

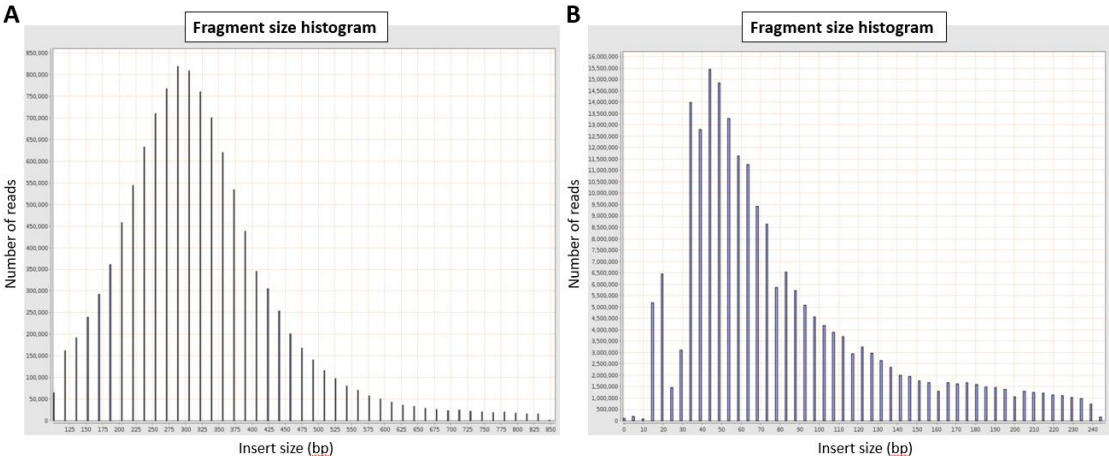


Figure 4. (A) RNA-seq fragment size histogram. Ideally, the fragments are expected to be larger than 2 x 115bp (read length). (B) ATAC-seq fragment size histogram. We observe only inter-nucleosomal fragments.

Peak calling

Peak calling is a method for the identification of genomic regions enriched with aligned reads. In this thesis, this step is only performed for the ATACseq data in order to characterize the open chromatin regions on the genome. A popular peak calling tool is macs2 *(Zhang Y. et al. 2008)*. After filtering out duplicate reads, macs2 detects genomic regions with overall read coverage greater than the expected background. Those are the peaks. The expected background is calculated by the following equation (Eq. 1):

$$Expected\ background = \frac{number\ of\ reads\ \ *\ \ read\ length}{mappable\ genome\ size}$$

Equation 1. Calculation of the expected background by macs2 during the peak calling process.

The mappable genome size is always smaller than the actual genome size because some genomic regions do not produce any reads.
Subsequently, the tool utilizes Poisson distribution for the calculation of a p-value for each peak, in order to examine the potential existence of local biases in read background levels. The p-value is corrected with the FDR method to give the q-values. The default cutoff for a peak to be statistically significant is q-value < 0.05 *(Zhang Y. et al. 2008)*.

p-value correction is essential due to the multiple testing problem according to which, the greater the number of independent statistical tests, the greater the chance of type I errors (greater false positive rate). A common correction method is that of Benjamini

and Hochberg (BH) -also known as False Discovery Rate (FDR) method -that accounts for the number of tests run *(Benjamini Y., Hochberg Y. 1995)*.
The peaks found by the peak calling process are shown in column 1 on table 4.

Visualization of Aligned reads and Peaks

For the visualization of our mapped reads and open chromatin regions, the data are uploaded to Genome Browsers. Genome Browsers include genome sequences and annotations with a graphical interface and navigation tools, thus providing the user with an interactive way for genomic data visualization. Most Genome Browsers are web-based, which makes them easily accessible, and are built on high performance servers, which can support computationally expensive tasks in a large scale more easily. They are divided in two groups according to the number of species they provide information for: Multiple-species browsers than include more than one species and thus offer comparative analyses across species and species-specific browses that offer more detailed annotation information on a unique species *(Wang J. et al. 2013)*.
For the visualization of the data, I used the UCSC genome browser *(Kent W.J. et al. 2002)*, a web-based multi-species browser. This software offers genome display in all scales (ranging from per base level to chromosome level resolution) and rich annotation information that includes: gene and transcript info, protein information, regulatory region coordinates, conservation across other organisms, SNPs, CpG islands and other repeats info etc. The gene annotations come from multiple repositories such as ENCODE, RefSeq (NCBI), GENCODE etc. *(Kent W.J. et al. 2002)*.
Although BAM files (the output file format of the Alignment process) can be visualized directly on UCSC, the bigWig format is more suitable for mapped reads visualization. In bigWig the data appear continuous, not as individual reads, so the final output is graph-like. Each base gets a floating-point number, which is normalized for the effective genome size. Those floating point numbers are depicted on the y-axis, while on the x-axis are the chromosome coordinates. Also, only the portions of the file that are needed for visualization are transferred to the browser, which makes bigWigs files faster to display *(Kent W.J. et al. 2010)*.

	Old_peaks	New_peaks	Final peak percentage
WT1	102,637	29,272	28.5%
WT2	96,192	33,628	35%
WT3	53,224	18,201	34.2%
KO1	80,191	29,797	37.2%
KO2	109,007	50,256	46.1%
KO3	72,604	18,217	25.1%

Table 4. Numbers of peaks and final peak percentages before and after filtering out the peaks with q-value <8.

For the conversion of BAM files to bigwig files bamCoverage *(Ramírez F. et al. 2016)*, a deepTools tool was used, with RPKM (Reads Per Kilobase per Million) normalization of the floating point numbers and a bin size of 10. As a rule of thumb, smaller bin size offers higher resolution but results in larger files that need more computational strength *(Ramírez F. et al. 2016)*. For the upload of the peak coordinates for peak visualization on UCSC the narrowPeak files (the output file format of the Peak calling process) suffice.

When we uploaded the narrowPeak data on UCSC, we noticed something was off. Although all samples are RPKM normalized and the same threshold was applied for all samples in macs2 (default: q-value < 0.05), some read enrichments that were expected to be peaks were in fact not. An example is shown on figure 5 where in sample WT1 a read enrichment with floating point number =9 is considered a peak, whereas in sample KO3 a read enrichment with floating point number =10 is not. Therefore, the threshold had to be changed. The `-log(q-values)` of the read enrichments varied from 1.3 to 13,643.5, with the majority of them being from 1.3 to 20, peaking at around 2.5 (Fig 6A). We decided to consider peaks only the read enrichments with `-log(q-values) > 8` (Eq. 2).

$$q\_value < 0.05 \quad \rightarrow \quad -log(q\_value) > 1.3$$

$$-log\,(q_{value}) > 8 \quad \rightarrow \quad q\_value < 10\text{^}(-8)$$

Equation 2. Calculation of the starting and final threshold of the q-values of the peaks.

A figure of the browser after the new threshold is depicted on figure 5. The results with the final numbers of peaks are shown on table 4.

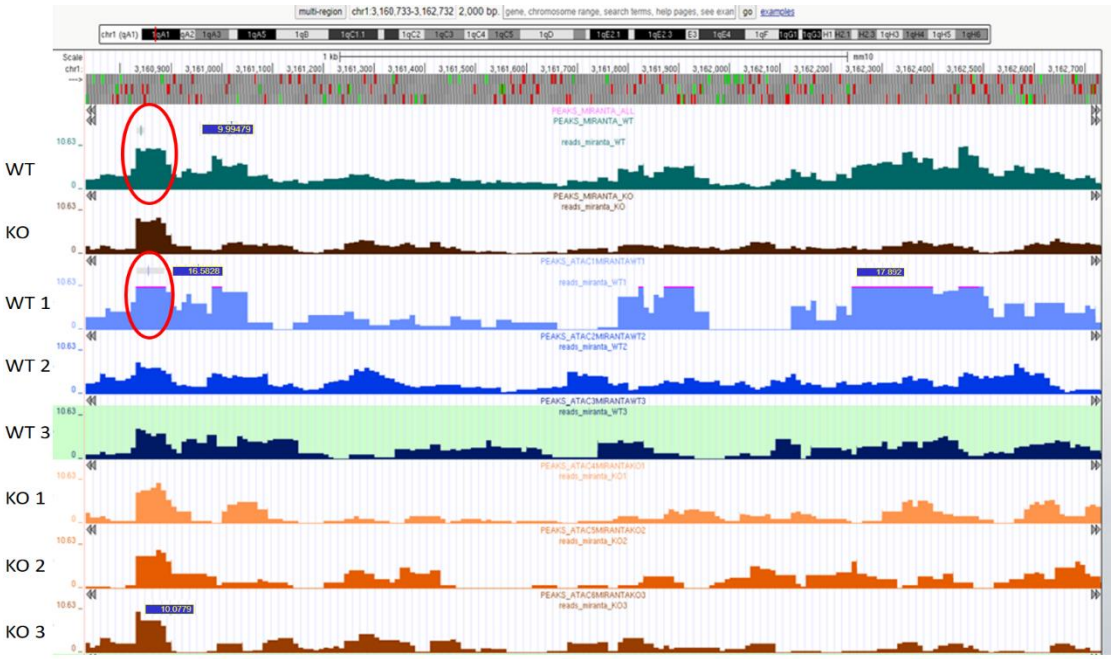


Figure 5. A snapshot of the mouse data from the UCSC genome browser. Each line is the RPKM normalized read abundance of a specific sample. The read abundances with rectangles on top (inside the red circle) are the peaks. Some reads abundances should be peaks but are not. For example, in WT1 a read enrichment with floating point number =9 is considered a peak, whereas in KO3 a read enrichment with floating point number =10 is not. Similarly, in WT1 a read enrichment with floating point number =16 is considered a peak, whereas a read enrichment with floating point number =17 is not. Therefore, the threshold needs changing.

Read quantification for Differential Analyses

In order to pinpoint the changes in gene expression and in regions of open chromatin between WT and KO mice, we need to quantify the expression of each gene and the accessibility of each regulatory region and compare them across the two conditions. Those processes are called Differential Expression and Differential Accessibility

Analysis respectively. To do this we need to count the amount of reads that come from each gene and open region respectively and perform a statistical analysis for their comparison across the two conditions.
For the quantification of the reads we used featureCounts *(Liao Y. et al., 2014)*, a tool that counts reads aligned to specific genomic features (genes, exons, promoter, gene bodies, genomic bins and chromosomal locations). These features are used as reference positions where the tool will count the number of aligned reads and provide a report with statistical information, such as percentages of successfully and unsuccessfully assigned reads etc. The tool also identifies whether a read overlaps (for at least one base) with more than one feature. We used the option - O which counts the reads in all the features they may overlap with *(Liao Y. et al., 2014)*. This is particularly helpful in ATAC-seq data where a genomic region could be regulating more than one coding region, so its reads should be counted for all of them.
The genomic features are given to featureCounts as input in a separate file. In the case of the RNA-seq data the genomic features are the gene ids (or exons, gene regions etc.), derived from the organism's annotation file. In the case of the ATAC-seq data, peaks are not always on regions of the genome that are annotated or shared between experiments. As a result, no standard input file can be given to the tool. The reads have to be counted in every possible peak position, so the file has to be created from scratch according to the peak data of every different ATAC-seq experiment. To identify all possible peak positions we need to merge all the ATAC bam (aligned) files from all conditions and perform the peak calling process in the merged file. The resulting peaks are all the possible peaks. Before the file is used as input in featureCounts it needs to be converted from the .narrowPeak format (macs2-peak calling output) to the .gff format (featureCounts input). The merging of the bam files is done with Samtools *(Danecek P. et al. 2021)* and the conversion to gff with the bed2gtf tool *(Pfurio 2014)*.

Differential Analyses

For the differential expression and accessibility analyses either of the following tools were used: DESeq2 and EdgeR. As specified in the work of Gontarz P. et al. 2020, those tools, along with the tool limma, can successfully identify about 92.7% of the actual differentially accessible regions. Additionally, DESeq2 has better specificity and EdgeR has a better sensitivity. They both account for batch effects and have incorporated visualization tools *(Gontarz P. et al. 2020)*. That is why they were chosen over limma.
DESeq2 *(Love M.I. et al. 2014)* performs in a one-step process the normalization of the counts, the dispersion estimation and the statistical tests for the calculation of significant differentially accessible regions (DAR) and differentially expressed genes (DEG).
Differential analysis methods rely on the fact that read abundance reflects how much a gene is expressed, or a region is open. Although this is true, differences in read numbers can also be attributed to the library's depth and size and to the gene's length. Because the same genes are compared between conditions, read length does not pose an issue. However, differences in library depths between duplicates can lead to different read abundances because the greater the depth the more reads are produced from that library. Library size also affects the analysis because if in a sample, a great percentage of our

library consists of very few genes that are very highly expressed, the rest of the genes will appear falsely under-sampled. As a result, normalization has to be performed prior to the analysis in order to ensure that the differences in read abundance mirror biological differences *(Chen Y. et al 2016)*.
For the calculation of the normalization factors, the trimmed median of means (TMM) method is used. In this method, a percentage of the highest and lowest values are removed from the data set before mean calculation. The dispersion (sum of biological and technical variance) is estimated with the empirical Bayesian shrinkage (EBS) method, where the variance of a gene is adjusted (shrank) after the consideration of the total variance of the data set, as each gene's expression is not entirely independent of the other's (they are explanatory). Finally, DESeq2 uses negative binomial generalized linear models and Wald statistics for the identification of the statistically significant differences between the samples. The default p-adjusted threshold (FDR method) used is 0.1 and the default log Fold Change threshold is 0. DESeq2 also performs filtering of low counts (minimum amount of counts = 10), which are known to interfere with some statistical processes while not providing any significant information on differential expression or accessibility *(Love M.I. et al. 2014)*.

EdgeR *(Robinson M. D. et al. 2010)* also uses TMM for the normalization of the counts. For the dispersion estimation it uses EBS. The rest of the steps depend on whether there are two or multiple experimental conditions: In experiments with two conditions, it uses the quantile-adjusted conditional maximum likelihood (qCML) method for dispersion estimation and the exact test for the statistical testing. In experiments with more than two conditions it uses the Cox-Reid profile-adjusted likelihood (CR) method for dispersion estimation and either the Quasi-likelihood negative binomial method or the likelihood ratio test for the statistical testing. The low count filtering is done manually by the user and the default p-value correction method is FDR.
Because the EdgeR analysis is a multiple step process I decided to use SARtools *(Varet H. et al., 2016)*, an R package/wrapper that provides the ability to perform the EdgeR analysis, along with the visualization processes in one step.

We used DESeq for the analysis of the ATACseq data and SARtools/EdgeR for the analysis of the RNAseq data. In DESeq, by default, open regions are considered statistically significant between two conditions when they have adjusted p-values < 0.1. In our data's case however, that threshold is too strict and not enough DARs are obtained. We decided to change the threshold to p-value < 0.05. That way we obtained a total of 3743 DARs, out of which 1187 were upregulated (log_2 Fold Change (log_2FC) >0) and 2556 were downregulated (log_2FC <0). Those results are depicted in the volcano plot in figure 6B.

In SARtools/EdgeR, we first set the threshold for genes to be differentially expressed as adjusted p-values < 0.05. That threshold was again too strict and not enough DEGs were obtained. We decided to change the threshold to p-value < 0.05. We obtained a total of 820 DEGs, from which 457 were upregulated ($log_2FC>0$) and 363 were downregulated ($log_2FC<0$). Those results are depicted in the volcano plot in figure 7A.

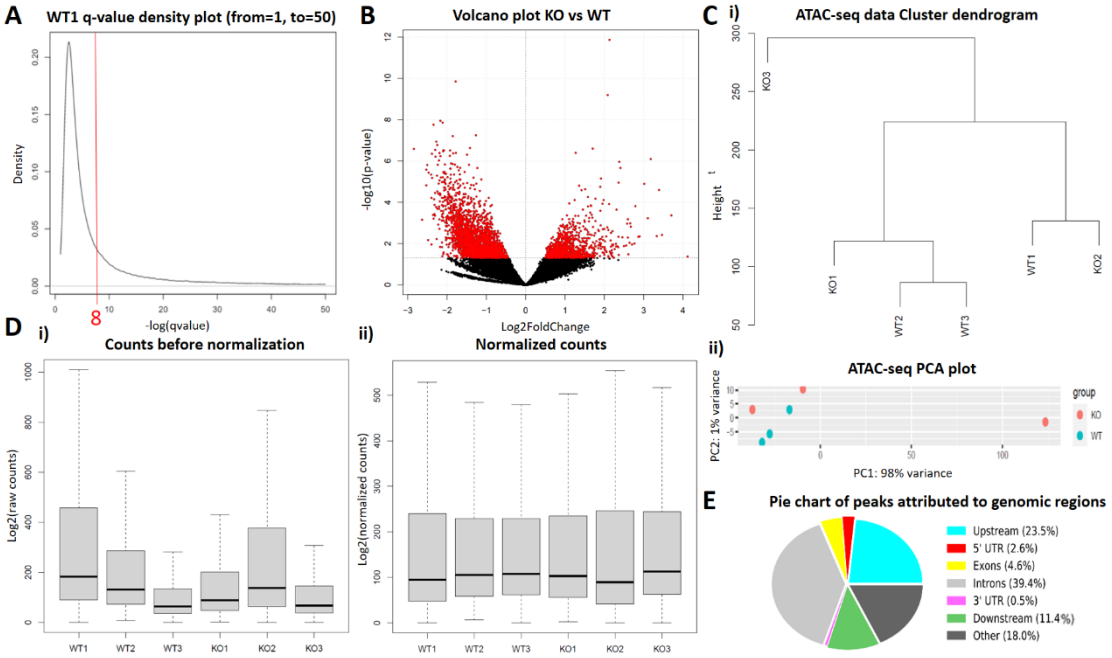


Figure 6. (A) Density plot of the –log(q-values) of the macs2 derived peaks. Only peaks with –log(q-values) > 8 were kept. (B) Volcano plot of the DESeq2 results for the KOvsWT comparison of the ATAC-seq data. Red dots correspond to the 3743 DARs that were identified with a p-value <0.05, out of which 2556 were downregulated (log2FC<0) and 1187 were upwnregulated (log2FC>0). (C) i) Cluster dendrogram of the ATAC-seq data, calculated with the Euclidean distance method. ii) PCA plot of the ATAC-seq data. The data do not cluster according to their experimental setup. (D) Boxplots of the ATAC-seq counts before and after TMM normalization with DESeq2. (E) A pie chart with the percentages of ATAC-seq DARs that were attributed to various genomic features. Out of the total 3743 DARs, 3068 were attributed to genomic regions and kept for the downstream analysis.

As shown in the PCA plots and the dendrograms from both the ATACseq (Fig. 6C) and the RNAseq datasets (Fig. 7B), the data do not cluster according to their experimental setup. This could be either due to the fact that the knock-out of a single gene is not sufficient to differentiate the data or due to errors during the experimental setup. Counts before and after normalization are shown in figures 6D and 7C for the ATAC-seq data and RNA-seq data respectively.

Attribution of peaks to genes

The next step after the identification of regulatory regions (DARs) is to find the genes that these regions regulate. At first, this task appears to be trivial and the first thought is to attribute the peak to the nearest gene. However, a region can have more than one genes in its proximity or it can be the regulatory element of more than one gene. Additionally, it is well known that enhancers can be proximal or distal, that is, they could be controlling a coding region that is nearby or several Kb away. There have even been several reports that a distal enhancer can control genes that are up to 3Mb away. As a result, peak attribution to genes can be a very challenging task and multiple tools have been developed to address this problem.

PAVIS2 *(Huang W. et al. 2013)* is a peak annotation tool that utilizes UCSC genomic annotation data to annotate peaks to genomic regions and provides visualization reports. The tool attributes peaks to gene features like exons, introns and UTRs, as well as upstream and downstream regions, the length of which is chosen as input by the user.

We set a range of 30Kb upstream and 10Kb downstream of the genes for the attribution of our peaks to genomic features by PAVIS2. The majority of our peaks are located in intronic regions (~39.5%), a significant amount (~35%) was found within the appointed range but outside of genomic features and about 18% of the peaks were not appointed. The overall results of our peak annotation process are depicted in figure 6E. We then focused our analysis only on the peaks that were attributed within the +30Kb/-10Kb range, that is 3068 peaks (the rest 18%, 673 peaks were discarded).

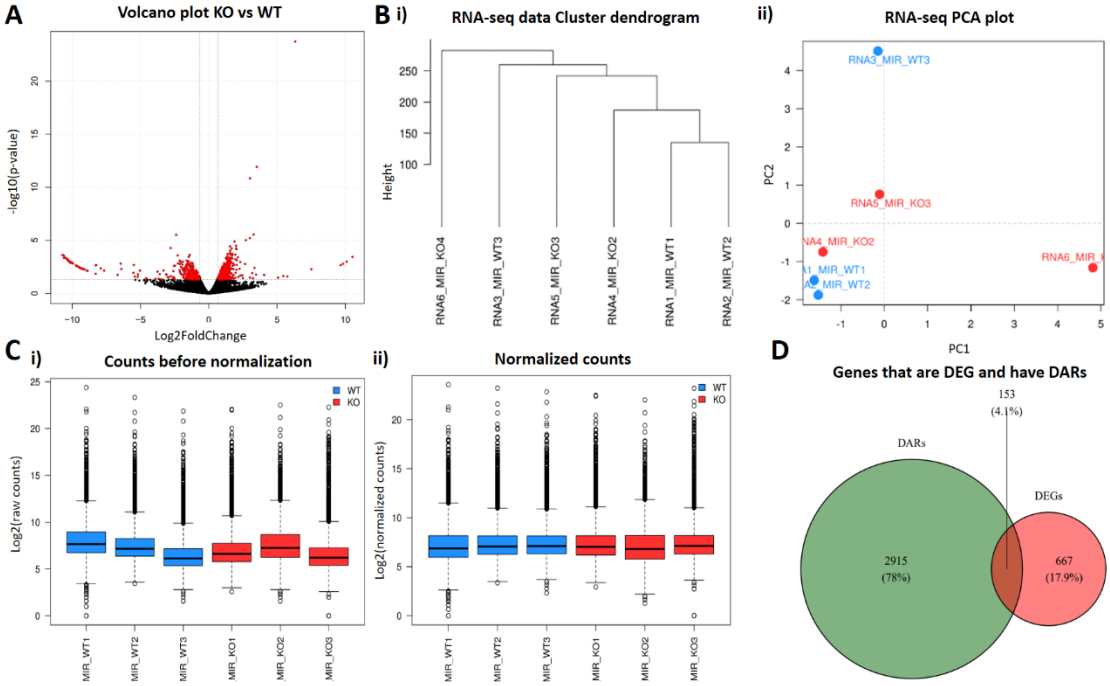


Figure 7. (A) Volcano plot of the EdgeR results for the KO vs WT comparison of the RNA-seq dataset. Red dots correspond to the 820 DEGs that were identified with a p-value <0.05, out of which 457 were upregulated (log2FC)>0 and 323 were downregulated (log2FC<0). (B) i) Cluster dendrogram of the RNA-seq data, calculated with the Euclidean distance method. ii) PCA plot of the RNA-seq data. The data do not cluster according to their experimental setup. (C) Boxplots of the RNA-seq counts before and after TMM normalization with EdgeR. (D) Venn diagram of the genes that are DEGs and have DARs. The common genes (DEGs with peaks that are DARs) are 153.

Correlation analysis

The final step is to investigate whether the ATACseq and the RNAseq results correlate with each other. We isolated the genes who have both attributed DARs and are differentially expressed (n=153) (Fig. 7D) and examined whether the $\log_2$FC of their peaks correlate with the $\log_2$FC of their differential expression.

The correlation analysis was done with the Spearman test in R using the ggstatsplot package *(Patil I. 2021)*. The Spearman test was chosen because it is a suitable correlation method when the data do not follow a Gaussian (normal) distribution. In this analysis, the p-value defines whether the data correlate and the correlation coefficient ρ shows the strength and direction of the correlation. For the data to be correlating the p-value has to be <0.1. The smaller that number, the stronger the evidence that the datasets correlate. For a correlation to be strong the |ρ| value has to be close to 1. A negative ρ shows negative correlation and a positive ρ a positive correlation.

The Spearman analysis showed strong evidence (p-value=5.708e^{-12}) of a substantial positive correlation (ρ=0.45) (Fig. 8A). We note that some genes and DARs are anti-correlated as they have opposite signs in both modalities. It would be interesting to determine if the genes are regulated by transcriptional repressors or activators to understand why these regions and genes do not change in the same direction upon KO. The correlation was even higher when I filtered for genes with either positive only or negative only log_2FC in both conditions (n=118) (p-value=2.2e^{-16}, ρ=0.71) (Fig. 8B). A heatmap with the correlating ATAC-seq and RNA-seq logFC of those 118 genes is shown in figure 9.

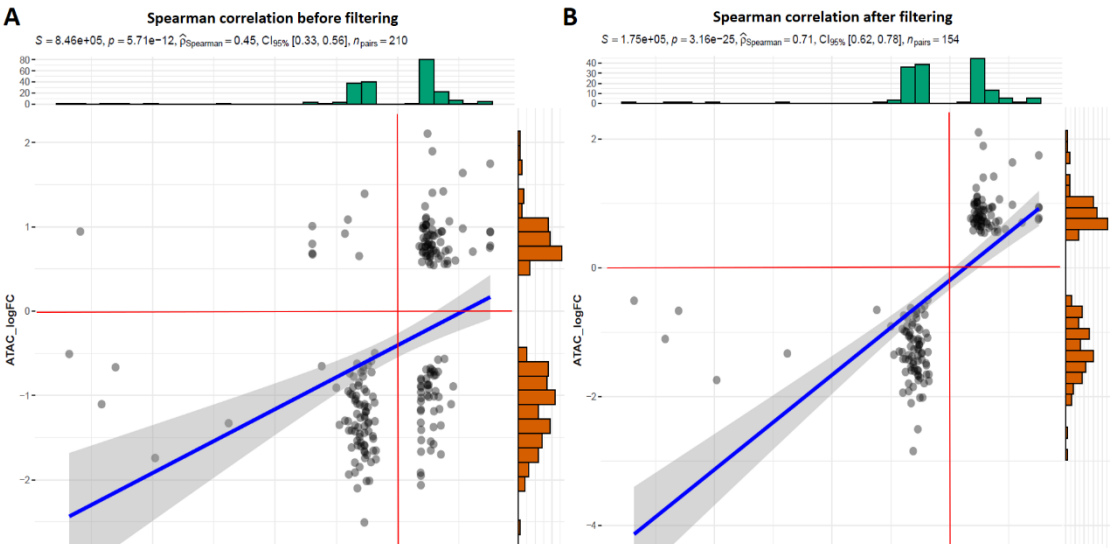


Figure 8. Spearman correlation results. (A) Spearman correlation plot before filtering. (B) Spearman correlation plot before filtering for genes that have only positive or only negative ATAC-seq and RNA-seq results. The p-value decreases and the Spearman coefficient increases after filtering. Thus, the data correlate more strongly.

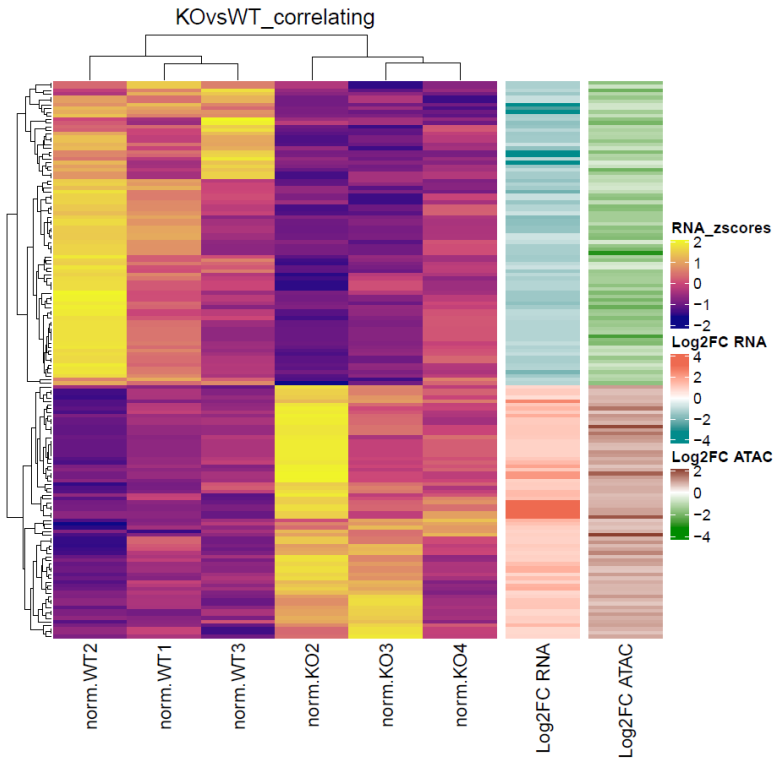


Figure 9. Heatmap of the genes with correlating ATAC-seq and RNA-seq logFC.

The main plot depicts the z-scores of the 118 genes that were derived from the Spearman correlation analysis after filtering *(p-value= 2.2e^{-16}, ρ= 0.71)*.

Cluster dendrograms show the similarities of the genes (left dendrogram) and the conditions (top dendrogram).

Additional heatmaps of the RNA-seq logFC values and the ATAC-seq logFC values of those genes are found on the right.

The correlation analysis concludes the construction of the pipeline. An overview of the pipeline is shown in figure 10.

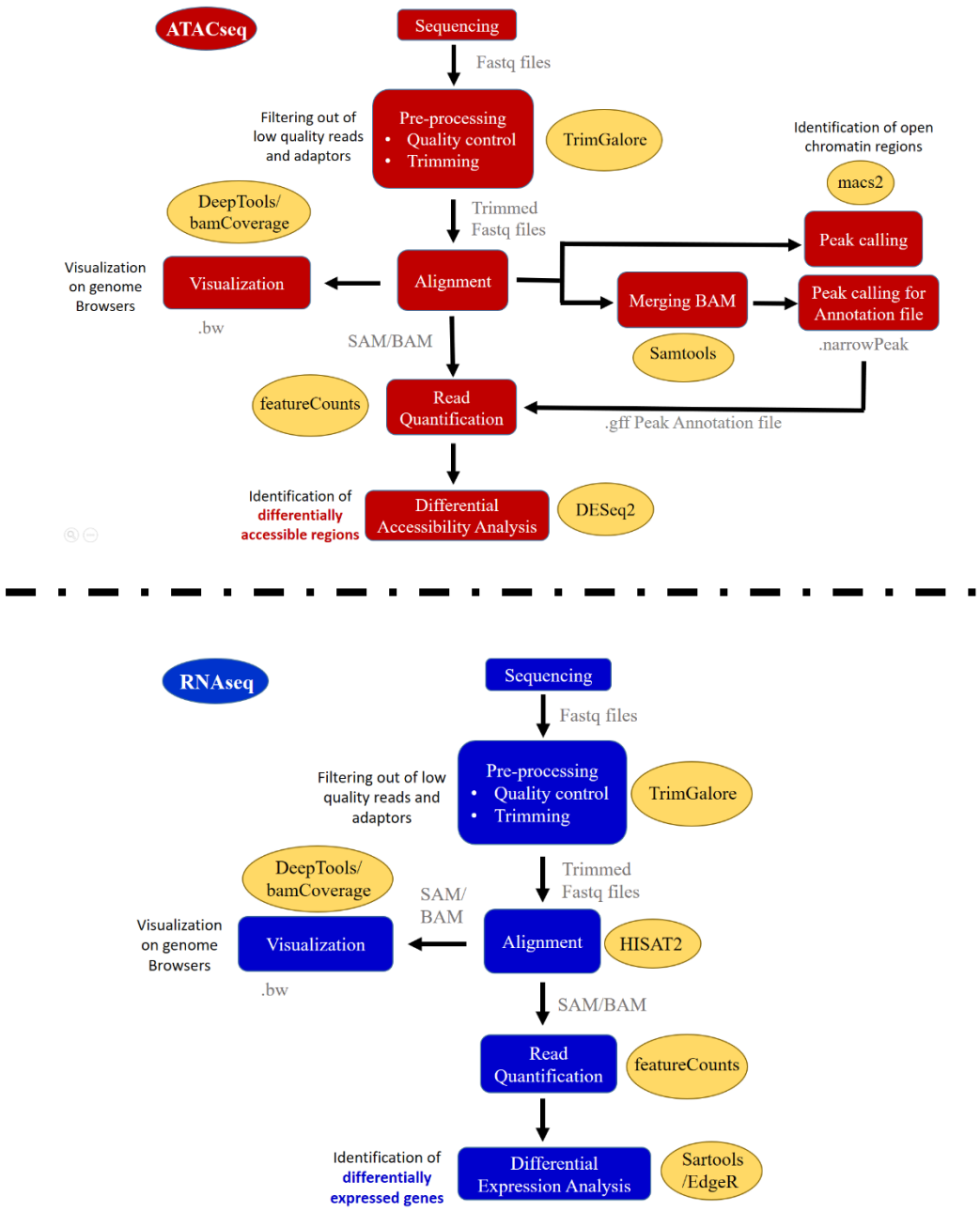


Figure 10. An overview of the established pipeline. Top: The steps for the ATAC-seq dataset. Bottom: The steps for the RNA-seq dataset. In yellow: the tools used. In gray: the files formats that are needed as input or are derived as output form each tool.

Applying the pipeline in *Parhyale hawaiensis*

The nature of the project

After creating the pipeline in the mouse dataset, I went on and applied it to datasets from the model organism *Parhyale hawaiensis*. The Omni-ATACseq data were previously produced by John Rallis in Pavlopoulos Lab and are duplicate libraries of paired-end reads (150bp), from stages S13, S17 and S19 of *Parhyale hawaiensis* embryonic development. The libraries were sequenced on the NextSeq500 platform (Illumina). The RNAseq data were retrieved as raw reads (fastq files) from the work of Sun D. et al, 2022, (available at SRA: Bioproject PRJNA765726) and are triplicate libraries of paired-end reads (150bp), from the developmental stages S13 and S19. All the developmental stages are classified according to the Browne et al. (2005) staging guide.

Applying the pipeline

On the first QC (before Trimming) that was performed for the ATAC-seq data, I noticed a great amount of reads with lengths smaller than the expected 150bp, with a large peak around 30-35bp (Fig. 11). Small reads are indicative of an over digestion by Tn5 during library preparation. These reads can be removed as they add noise to the dataset and might temper with the statistics or other analyses. As a result, we decided to filter out the reads with length <50bp. Although the expected read sizes are 150bp, having reads at ~100bp is normal. Those reads are produced from fragments shorter than 300bp, which are produced by intra-nucleosomal regions where Tn5 binds frequently and are overlapping with their pairs. I decided to choose the 50bp threshold in order to avoid losing useful information that derived from Tn5 and not from library over digestion.
Reads with length <50bp, along with the Nextera adaptors and other low quality characteristics (default Phred threshold: 20) were removed with TrimGalore in the Trimming step. Approximately 27%, 33% and 24% of the total reads were removed in stages S13, S17 and S19 respectively.

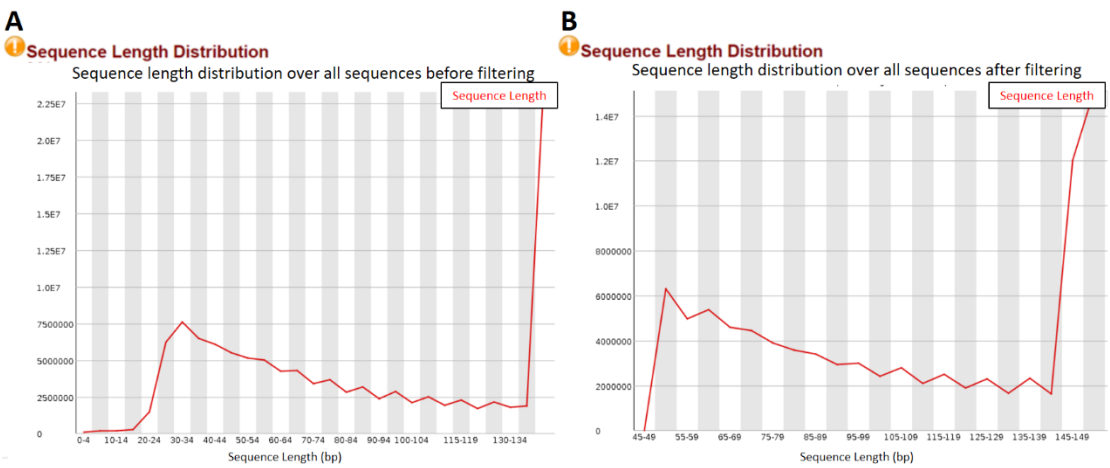


Figure 11. Sequence Length Distribution plots form the FastQC report of the ATAC-seq dataset in Parhyale. (A) Sequence Length Distribution plots before the filtering. A large peak is evident at ~35bp that then decreases slowly till the 150bp region (B) Sequence Length Distribution plots after the filtering of reads with length <50bp. Reads with lengths smaller than the expected 150bp are still present in the data.

For the mapping to the genome (version 5.0: genomeV5.fa) I used the splice-unaware aligner Bowtie2 *(Langmead B. et al. 2012)* for the ATAC-seq reads and the splice-aware aligner HISAT2 for the RNA-seq reads. The pseudoaligner Kallisto *(Bray N. L. et al. 2016)* was used for the mapping to the Transcriptome and for the quantification of the reads for the differential expression analysis.

In the RNA-seq data the mapped/unmapped ratio is a bit lower than in the ATAC-seq data, but the mean coverage is significantly higher. The mean read lengths in the ATAC-seq data are noticeably lower due to the filtering out of the reads with length <50bp (Tables 5 and 6).

In the fragment size histograms of the ATAC-seq data, S19_2 does not have the expected pattern. In S13_2 and S19_1 the second peaks are small and all apart from S19_2 are slightly moved to the left (Fig. 12). The lack of the expected patterns is probably due to library construction errors.
In the RNA-seq data, the plots show a peak around 300bp as expected, but in samples S13_2 and S19_3 the peaks slightly drift to the left (Fig. 13).

ATAC Mapping Results	Number of reads	Mean Coverage	%GC	mean read length
S13_1	164,043,634 (87.87% mapped, 12.13% unmapped)	5.6297	42.71%	107.56
S13_2	174,272,596 (87.25% mapped, 12.75% unmapped)	5.5815	43.27%	101.45
S17_1	131,129,694 (88.86% mapped, 11.14% unmapped)	4.7122	42.46%	111.17
S17_2	156,158,354 (89.11% mapped, 10.89% unmapped)	5.1761	43.03%	102.34
S19_1	134,302,470 (85.61% mapped, 14.39% unmapped)	4.3906	42.04%	106.04
S19_2	178,915,872 (82.97% mapped, 17.03% unmapped)	5.6865	42.46%	105.74

Table 5. A summary of the mapping results of the Parhyale ATAC-seq data.

RNA Mapping Results	Number of reads	Mean Coverage	%GC	mean read length
S13_1	107,549,084 (65.17% mapped, 34.83% unmapped)	62.2119	46.1%	152.73
S13_2	64,810,264 (65.62% mapped, 34.38% unmapped)	33.152	42.61%	151.25
S13_3	84,166,312 (65.58% mapped, 34.42% unmapped)	51.7813	46.45%	152.6
S19_1	76,610,844 (68.11% mapped, 31.89% unmapped)	46.7239	46.42%	152.62
S19_2	91,216,840 (68.11% mapped, 31.89% unmapped)	56.0699	46.54%	152.66
S19_3	72,996,084 (69.69% mapped, 30.31% unmapped)	50.2515	45.48%	152.66

Table 6. A summary of the mapping results of the Parhyale RNA-seq data.

macs2 was used for the peak calling with the BH correction method and the default q-value cutoff < 0.05 *(Zhang Y. et al. 2008)*.The peaks found by the peak calling process are depicted on table 7.

The results were visualized in the Apollo Genome Browser *(Dunn et al. 2019)*. Apollo is a web-based genome browser that gives users the ability to upload their own reference genomes, transcriptomes, annotation files. That feature makes it a very helpful tool for the use of NGS data in less popular model organisms like Parhyale. It has a user-

friendly interactive interface that allows easy genome navigation and editing of common tracks by multiple users *(Dunn et al. 2019).*

	Peaks
S13_1	327.471
S13_2	360.773
S17_1	262.389
S17_2	326.244
S19_1	263.684
S19_2	295.913

◄**Table 7**. Numbers of peaks derived by macs2 with a default q-value < 0.05 during the peak calling process.

►**Table 8.** The overall loss of information due to the filtering out of contigs with peaks with chromStart=0.

Filtering out:
0.75% of the contigs
0.03% of all the genome,
0.42% of the peaks (3321 peaks)
0.29% of total genes (158 genes)

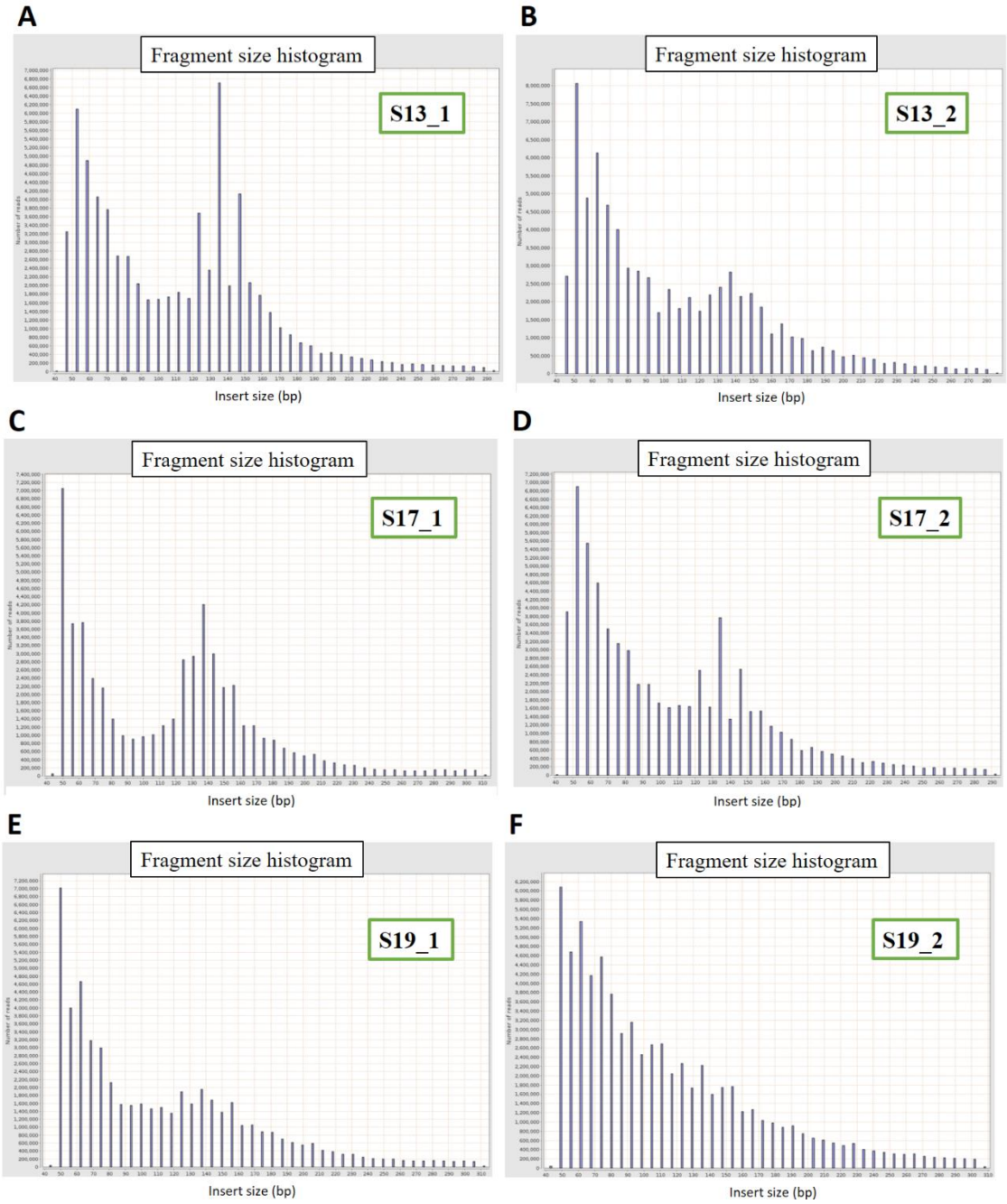


Figure 12. Fragment size histograms of the ATAC-seq results. In S13_1, S13_2, S17_1, S17_2 and S19_1 the peaks are drifted to the left. S19_2 does not have the expected peaked pattern.

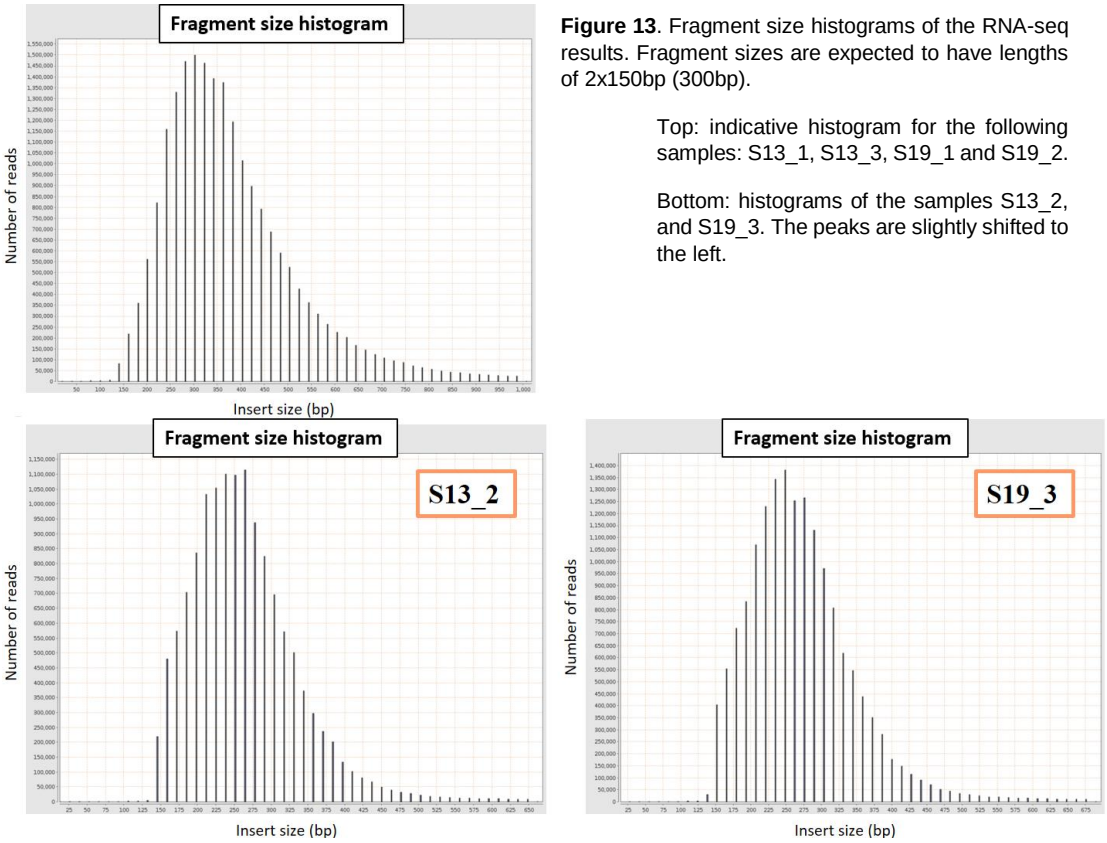


Figure 13. Fragment size histograms of the RNA-seq results. Fragment sizes are expected to have lengths of 2x150bp (300bp).

Top: indicative histogram for the following samples: S13_1, S13_3, S19_1 and S19_2.

Bottom: histograms of the samples S13_2, and S19_3. The peaks are slightly shifted to the left.

The chromStart problem

For the quantification of the ATAC-seq reads I wanted to use featureCounts but, during the narrowPeak to gff conversion of the genomic features file (featureCounts needs gff as input, as mentioned earlier), a problem emerged.

The chromStart and chromEnd positions (columns existing in both files) are the starting and ending positions, respectively, of a peak on a chromosome. In the narrowPeak format, the first base in a chromosome is numbered as 0 and in the gff format as 1 *(Kent W. J. et al. 2002)*. During the conversion from narrowPeak to gff the peak coordinates are retained. As a result, if the narrowPeak file includes peaks in the beginning of the chromosome (chromStart=0), the gff file ends up with 0s instead of 1s, which is not acceptable.

An easy solution to this problem is the replacement of the 0s by 1s in the gff, but the problem runs deeper than the surface. Because of the nature of Parhyale's genome - which is not yet fully assembled but remains in contigs- having a peak at the start of a chromosome means one of two things: a) either the peak is located at the EXACT start of the chromosome or b) the peak is located in the borderline of two contigs. Taking into consideration that the number of peaks with chromStart=0 are 3266, option b is very likely. Having a peak in a region shared between two contigs means that it will appear twice in the data. Therefore, those excess peaks have to be removed.

The contigs with peaks with chromStart=0 were isolated and a great majority of them had lengths that varied from 200 to 2000bp (Fig 14A). In order to avoid having the same peak twice I decided to remove those contigs from the analysis. Their length is small, so only a small percentage from the total genome would be discarded.

Additionally, because of the large amounts of repetitive sequences in Parhyale's genome, some contigs have overlapping parts and some regions appear in the genome assembly more than once. Therefore, there is a chance that the overall loss of genomic information is less than anticipated.

By removing the contigs with length <2000bp I filtered out 0.7523662% of the contigs (2093 contigs), 0.03369975% of all the genome, 0.4197209% of the peaks (3321 peaks) and 0.2885739% of the total genes (158 genes) (Table 8). It is worth noting that by removing the contigs with length <2000bp a) I do not remove all peaks with chromStart=0 -because all such peaks in contigs >2000bp are still left in the data- and b) by removing those contigs I remove not only the peaks with chromStart=0, but also the rest of the peaks that are located on them. For the remaining peaks with chromStart=0 the 0s were replaced by 1s in order to continue with the pipeline.

Differential analyses

We used DESeq2 for the analysis of both datasets. Again, the default threshold with adjusted p-values < 0.1 was too strict and not enough DARs were obtained. For the ATAC-seq, we used p-value<0.05 as threshold and identified 5758 DARs between S13 and S17 (2927 upregulated, 2831 downregulated - logFC threshold: 0) and 13260 DARs between S13 and S19 (6601 upregulated, 6659 downregulated - logFC threshold: 0) (Fig 14B, C).
For the RNA-seq, the default threshold was not strict enough and we used adjusted p-value<0.01 as threshold. We identified 1597 DEGs, from which 774 were upregulated ($log_2FC>0$) and 824 were downregulated ($log_2FC<0$). Those results are depicted in the volcano plot on figure 15A.

As shown in the PCA plots and the dendrograms from the ATAC-seq dataset, although the S19 points cluster together, the other two developmental stages are not clustered correctly (Fig. 14D). This could be either due to the fact that there are not many differences between the stages because they are very close to each other (no of hours), or it could be due to errors during the experimental setup.
The RNA-seq data cluster well (Fig. 15B). Counts after normalization are shown in figures 14E and 15C for the ATAC-seq data and RNA-seq data respectively.

The problem of attributing peaks to genes

For the attribution of peaks to genes, I could not use PAVIS2 because it does not have Parhyale's genome integrated in its database and it does not allow for manual uploading of genome and annotation files. What's more, peak attribution tools are built to use annotation files from known databases like NCBI or UCSC, which do not include that kind of information for Parhyale. Therefore, I had to perform the peak attribution manually.

As mentioned previously, an enhancer can be proximal or distal, but since I did not have the information or tools to look for distal enhancers, I decided to focus uniquely on proximal ones. It is safer to suggest that a regulatory region controls a nearby gene

when they are as close to each other as possible. This is why I attributed the peaks to their nearest genes only if they were up to 1Kb away. More specifically, I examined whether a peak overlaps at least 50% with a) the coding region of a gene (to check whether the peak is within transcript limits) b) 1Kb upstream of the transcript (from *the transcript's start position -1000 bases* to *the start of the transcript*) c) 1Kb downstream of the transcript (from *the transcript's end* to *the transcript's end position +1000 bases*). If a region (peak) had >= 50% overlap with more than 1 of those regions (coding, upstream or downstream region), or with >1 genes it was attributed to all those regions and genes.

Obviously, this method has several problems:

A. Some genes may be located in the +/- 1000 bases positions of other genes. In this case the peak is attributed to both genes, which is not necessarily biologically true.
B. If a peak starts before the transcript start position and ends after it, it is considered both a peak on the coding region and upstream of the gene (Fig. 16A).
C. The information about the regulation of distal elements is completely eliminated from the analysis.
D. False positive rate is great because some peaks might be attributed to genes they do not regulate.

The accuracy of the results remains to be validated by experiments for each enhancer-gene pair, or by a correlation of our results with HiC data.

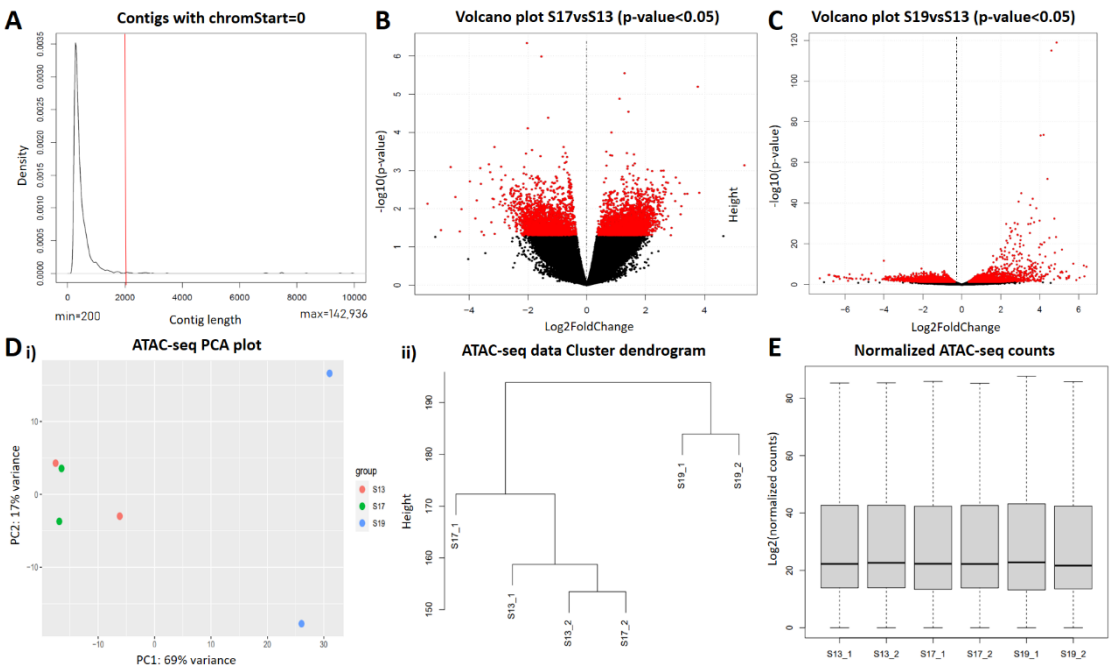


Figure 14. (A) Density plot of the lengths of contigs with peaks with chromStart=0. The majority of those contigs have lengths varying from 200-2000bp. Out of all the contigs with chromStart=0 peaks only the ones with lengths >2000bp were kept for the downstream analysis. (B) Volcano plot of the DESeq2 results for the S17vsS13 comparison of the ATAC-seq data. Red dots correspond to the 5758 DARs that were identified with a p-value <0.05, out of which 2927 were upregulated (log2FC>0) and 283 were 1 downregulated (log2FC<0). (C) Volcano plot of the DESeq2 results for the S19vsS13 comparison of the ATAC-seq data. Red dots correspond to the 13260 DARs that were identified with a p-value <0.05, out of which 6601 were upregulated (log2FC>0) and 6659 were downregulated (log2FC<0). (D) i) PCA plot of the ATAC-seq data. ii) Cluster dendrogram of the ATAC-seq data, calculated with the Euclidean distance method. The S13 and S17 data do not cluster according to their experimental setup. (E) Boxplot of the ATAC-seq counts after TMM normalization with DESeq2.

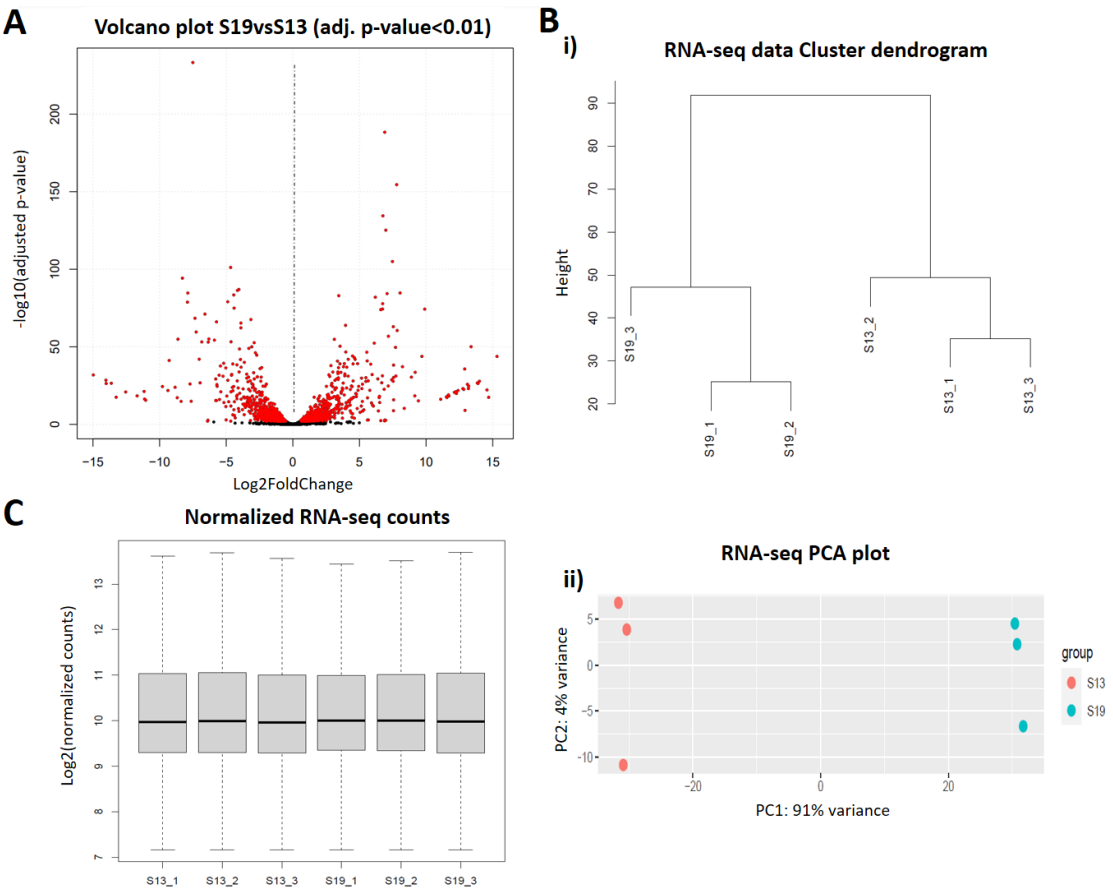


Figure 15. (A) Volcano plot of the DESeq2 results for the S19vsS13 comparison of the RNA-seq data. Red dots correspond to the 1597 DARs that were identified with an adjusted p-value <0.01, out of which 774 were upregulated (log2FC>0) and 824 were downregulated (log2FC<0). (B) i) Cluster dendrogram of the RNA-seq data, calculated with the Euclidean distance method. ii) PCA plot of the RNA-seq data. The data do cluster according to their experimental setup. (C) Boxplot of the RNA-seq counts after TMM normalization with DESeq2.

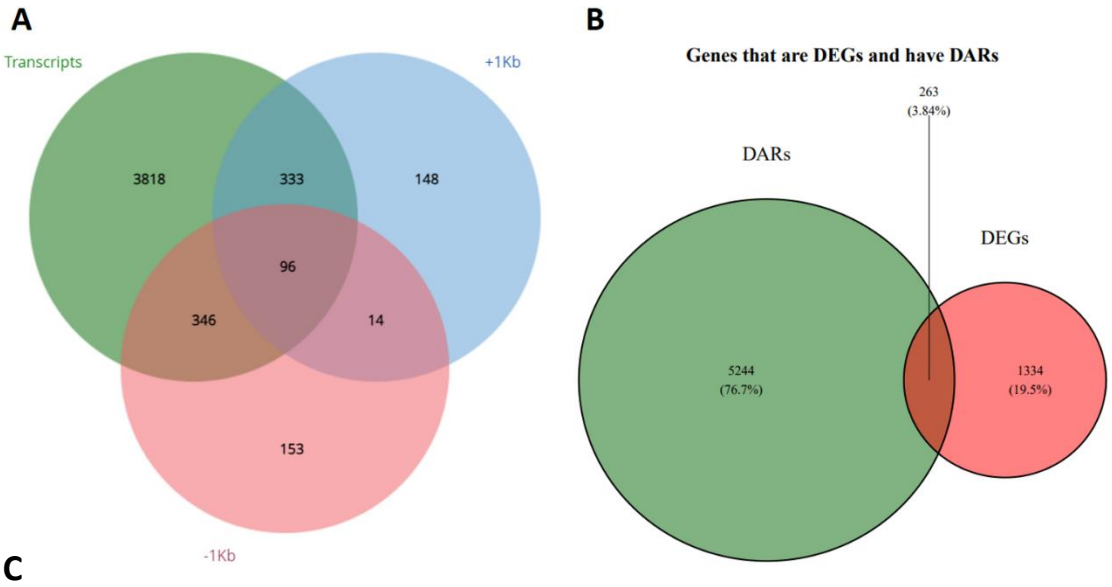
Of the 13260 DARs between S13 and S19, only 4908 (~37%) were attributed to coding regions or +/-1kb away. Those peaks were found to control a total of 6230 different transcripts in an overall number of 9311 eGRNs (because some peaks were attributed to >1 transcripts and a gene can have >1 peaks) (Fig. 16A).

The problem of the transcript ids

A setback with Parhyale's annotation is that it does not include common gene names, but the gene and transcript ids are in the form of mikado ids or MSTRG ids (Fig. 16C). For the attribution of mikado and MSTRGs ids to common gene names, John Rallis (in Pavlopoulos Lab) constructed a new annotation file. He used the sequences of each transcript, BLASTed them against the proteins of the Uniprot database and isolated the ones that were the best matches (along with some additional information concerning those proteins, like their database code, some basic functions of each protein etc.). The problems that occurred are the following:

1. only the sequences for the mikado ids were available, so all MSTRG genes are lost during the attribution of transcript ids to common gene names.
2. some mikado ids did not match any gene from any species.
3. some mikado ids match the same genes.

As a result, in RNA-seq, only 1215 of the 1597 differentially expressed transcripts were attributed to common genes. In the ATAC-seq data, out of the 6230 unique transcript ids with peaks, only 2769 genes were obtained. 3987 peaks out of 4.908 were kept (because the rest regulate MSTRGs) and the overall eGRN are 5507 of 9311 (Table 9).



C

phaw_50.283875b	Mikado loci	exon	1364946	1367524	gene_id "mikado.phaw_50.283875bG68"	transcript_id "mikado.phaw_50.283875bG68.1"	exon number "1"
phaw_50.283875b	StringTie	exon	1518815	1520733	gene_id "MSTRG.45772"	transcript_id "MSTRG.45772.1"	exon number "1"
phaw_50.283875b	StringTie	exon	1404886	1405198	gene_id "MSTRG.45771"	transcript_id "mikado.phaw_50.283986bG70.1"	exon number "1"

Figure 16. (A) Venn diagram of the peaks that were attributed to genomic regions. Those regions are either the coding regions of transcripts or regions +/-1Kb from the transcript's borders. If a peak starts before the transcript start position and ends after it, it is considered both a peak on the coding region and upstream of the gene. Therefore, some peaks are attributed to more than one regions of the same transcript and so some circles are partially overlaping. (B) Venn diagram of the genes that are DEGs and have DARs. The common genes (DEGs with peaks that are DARs) are 263. (C) Snapshot of the Parhyale hawaiensis annotation file. The gene and transcript ids are in the form of Mikado and MSTRG ids and it has no information of the common gene names.

	initial transcript ids	final common genes	% final
ATAC-seq	6230	3987	63,99679
RNA-seq	1597	1217	76,20539

Table 9. Numbers of transcript ids derived from the analysis and final numbers of genes that were kept after their attribution to common gene names.

Correlation of ATAC-seq and RNA-seq results

The correlation of the two datasets was performed in a transcript id level in order to avoid all problems related to their attribution to common gene names. I used the 1597 DEGs from the RNA-seq analysis and the 5507 different eGRNs from the ATAC-seq results. Out of them 263 of the ids were common (Fig. 16B). I examined whether the logFC values of the DEGs (RNA-seq) correlate with the logFC of the peaks of the eGRNs (ATAC-seq) with the Spearman test in R with the ggstatsplot package. The analysis showed strong evidence ($p = 3.65e^{-3}$) of a very low positive correlation. ($\rho_{Spearman} = 0.16$) between the two datasets (Fig. 17A) (n=263). A filtering of the data for either only positive or only negative log_2FC from both datasets showed a moderately

positive correlation ($\rho_{Spearman}$ = 0.66) with even more statistical significance (p = $6.06e^{-26}$) (n=200 eGRNs, corresponding to g=153 different genes) (Fig. 17B).

This analysis showed that 153 genes are differentially expressed and are additionally related to open chromatin regions (positive correlation) that were found differentially accessible between the stages S13 and S19. The eGRNs are 200 because some genes have >1 peaks in their proximity. Those genes can be viewed in the heatmap (Fig. 18).

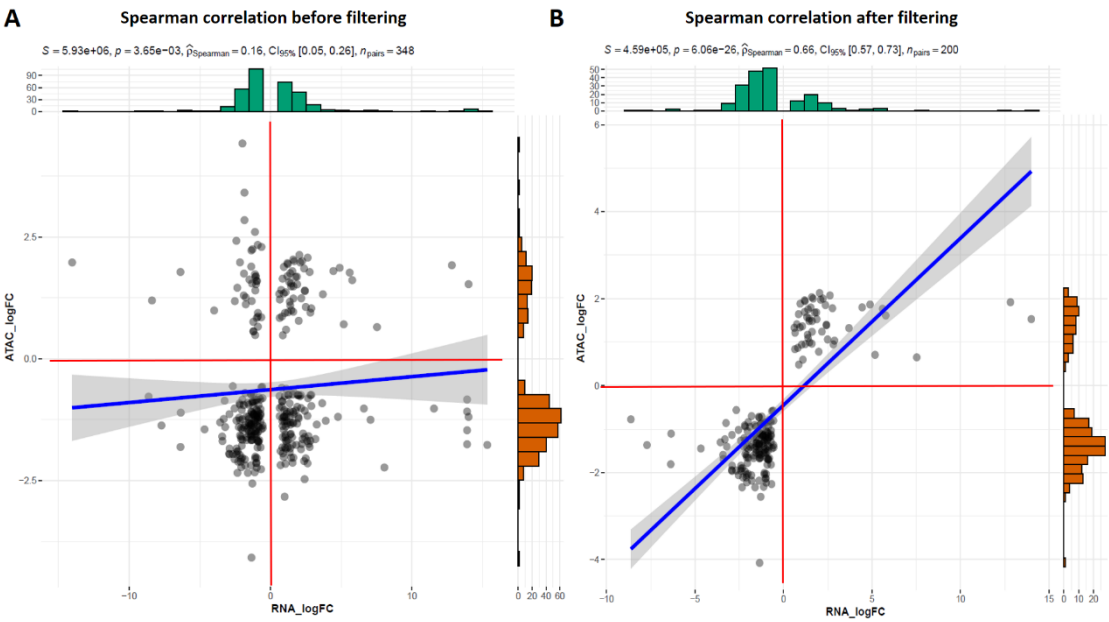


Figure 17. Spearman correlation results for Parhyale hawaiensis. (A) Spearman correlation plot before filtering. (B) Spearman correlation plot after filtering for genes that have only positive or only negative ATAC-seq and RNA-seq results. The p-value decreases and the Spearman coefficient increases after filtering. Thus, the data correlate more strongly.

Enrichment analysis

Metascape *(Zhou Y. et al. 2019)* is a gene meta-analysis web portal that performs Gene Ontology (GO) and Enrichment analyses. The tool allows for the comparison of a group of input genes with similar genes from different organisms like *Mus musculus*, *Drosophila melanogaster*, *Homo sapiens* etc. and provides information concerning the gene's biochemical pathways and the primary role of the proteins in various biological processes. It also performs an interactome analysis and contains information on functional protein structures. The tool's output is in the form of lists, in order for the results to be easily integrated to downstream analysis and multiple plots for visualization and easier interpretations of the outcomes *(Zhou Y. et al. 2019)*.

We used the 153 genes that were found from the correlation analysis after the final filtering as input to metascape. The analysis showed enrichment in multiple developmental processes like morphogenesis, regionalization and pattern formation, a result expected as the data come from different stages of Parhyale's development. The second category of processes that appeared enriched are metabolic processes, with the negative regulation of nitrogen compounds scoring second in the top 20 clusters (Fig.19).

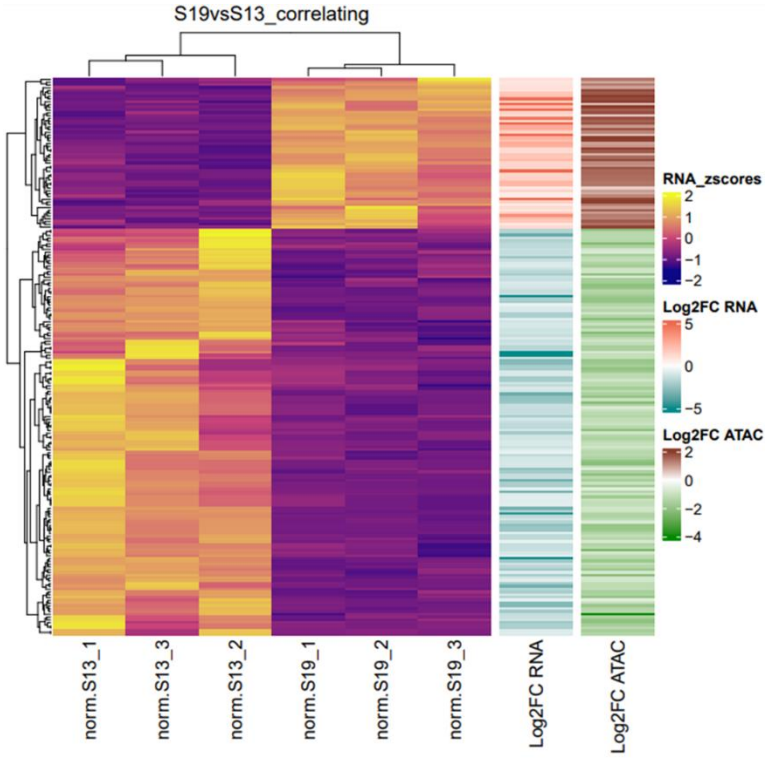


Figure 18. Heatmap of the genes with correlating ATAC-seq and RNA-seq logFC.

The main plot depicts the z-scores of the 153 genes that were derived from the Spearman correlation analysis after filtering *($p = 6.06e^{-26}$, $\rho = 0.66$)*.

Cluster dendrograms show the similarities of the genes (left dendrogram) and the conditions (top dendrogram).

Additional heatmaps of the RNA-seq logFC values and the ATAC-seq logFC values of those genes are found on the right.

Of all the genes we found, we decided to examine further genes involved in developmental pattering and morphogenesis, including gooseberry (gsb), lola-like (lolal) and homothorax (hth). All three genes showed enrichment to organ morphogenesis and regionalization and only the last two appeared enriched for processes concerning tube development. Additionally, they had adequately high normalized counts which means they are expressed high enough for us to be able to examine them. The analysis and correlation gsb's and lolal's expressions and the accessibility of their regulatory regions were shown to decrease from S13 to S19 (gsb: RNA-seq logFC = -2.2, ATAC-seq logFC=-1.2) (lolal: RNA-seq logFC = -0.8, ATAC-seq logFC=-0.9) and hth's were shown to increase (RNA-seq logFC = 1.4, ATAC-seq logFC=1.1 and 1.8).

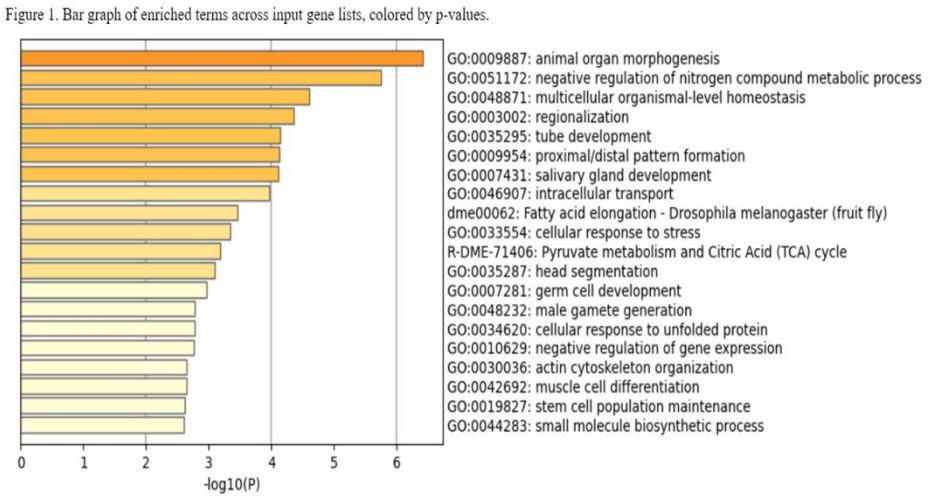


Figure 19. Results of the Gene Ontology and Enrichment analysis.

The top 20 clusters showed enrichment in developmental and metabolic processes.

Experimental Validation of the Results in *Parhyale hawaiensis*

The first step of the validation process is the validation of the RNA-seq results. This can be done by quantifying the RNA levels from the stages of interest with quantitative Real-Time PCR (q-RT-PCR) using a relative quantification approach.
To identify relatively unchanging genes suitable for q-RT-PCR normalization, I isolated those that were not marked as DEGs (p-adj>0.01), did not show much difference (-0.2<logFC<0.2), had low variation (-0.01<Variation Coefficient <0.01) and had normalized counts > 2800, so that their expression would be more easily detected. Out of the 18 genes that resulted from the filtering, I used SRP72 (mikado.phaw_50.283872cG118.1), a signal recognition particle subunit, as an internal control gene for normalization.

For the q-RT-PCR, approximately 100 Parhyale embryos of stages S13, S17 and S19 were collected and 10 µl of total RNA with concentrations 666.2ng/µl, 362.5 ng/µl, 633 ng/µl were isolated respectively. The samples were treated with DNAse and 2.9µg of RNA from each sample were used for cDNA synthesis. I used oligo$(dT)_{18}$ primers in order to examine only poly(A) mRNAs.

Due to the fact that Parhyale's genome annotation is not complete and the transcriptome does not yet include all isomorphs, exons and exon-intron junctions, I decided to design the primers so that the amplicon lies within a single exon. They would also have to be away from the UTRs because they are highly polymorphic and primer specificity would not be guaranteed. The primers were designed on Primer 3 *(Koressaar T. and Remm M. 2017)*, with lengths 18-20bp and annealing temperatures (Tm) ranging between 57-60°C. The amplicon sizes varied from 158-187bp. I made 1 primer set for hth, lola and the internal control and 2 primer sets for gsb.
After blasting the primers against the genome, I realized that the majority of them did not bind uniquely to the sites of interest, but it did not seem like the ectopic binding sites of each primer pair would give a different product. In order to make sure that the primer sets produce only one PCR product and to pinpoint the ideal Tm for the reaction, I performed a gradient end-point PCR prior to the q-RT-PCR.

The results of the PCR are depicted in the figure 20. All the genes have only 1 product at the expected heights. 55°C were not effective for the reaction. We decided to do the q-RT-PCR at 60°C, as it produced enough product and is closer to the primer's Tm. We used the primers for exon 3 for gsb as they appeared to produce more product.

For the q-RT-PCR, the reactions for the different stages and the internal control were performed in triplicates with 5µl of 1:20 diluted cDNA. The standard curves were done with 5µl of 1:10 diluted cDNA for 6 different dilutions (1, 1/4, 1/16, 1/64, 1/256, 1/1024) in duplicates and the results were derived with the ΔΔCt method. The reactions of the internal control and gsb had high efficiencies (105.3, 107.2 respectively), but the ones of lolal and hth were slightly lower, at 88.2 and 85.7 respectively (Fig. 21). The gsb and hth q-RT-PCR results validated the RNA-seq results, with the ΔΔCt deviating from the logFC (expected) by less than 1.0, that is one q-RT-PCR reaction cycle (Tables

10-12). On the other hand, lolal was showed to increase from S13 to S19, while the RNA-seq data showed a decrease between the two conditions. In addition, the q-RT-PCR results showed a larger difference (ΔΔCt=2.2), ~2.4 times higher than the absolute value of the expected RNA-seq results (logFC= -0.85) (Table 11). Although these results seem puzzling, they are in accordance with an independent set of RNA-seq results produced by John Rallis in Pavlopoulos Lab. This independent set is also in accordance with the gsb and the hth q-RT-PCR and RNA-seq results. Whether this divergence between q-RT-PCR and RNA-seq results is an isolated event remains to be determined.

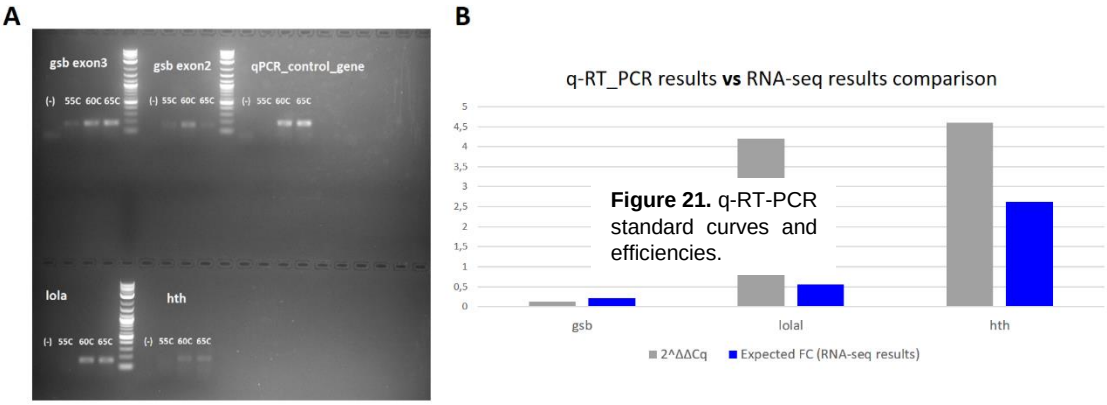


Figure 21. q-RT-PCR standard curves and efficiencies.

Figure 20. (A) End point PCR results. The PCR was performed at three different annealing temperatures: 55, 60 and 65℃. (B) q-RT-PCR results. gsb and hth gave the expected results but lolal showed an increase in its expression instead of the expected decrease.

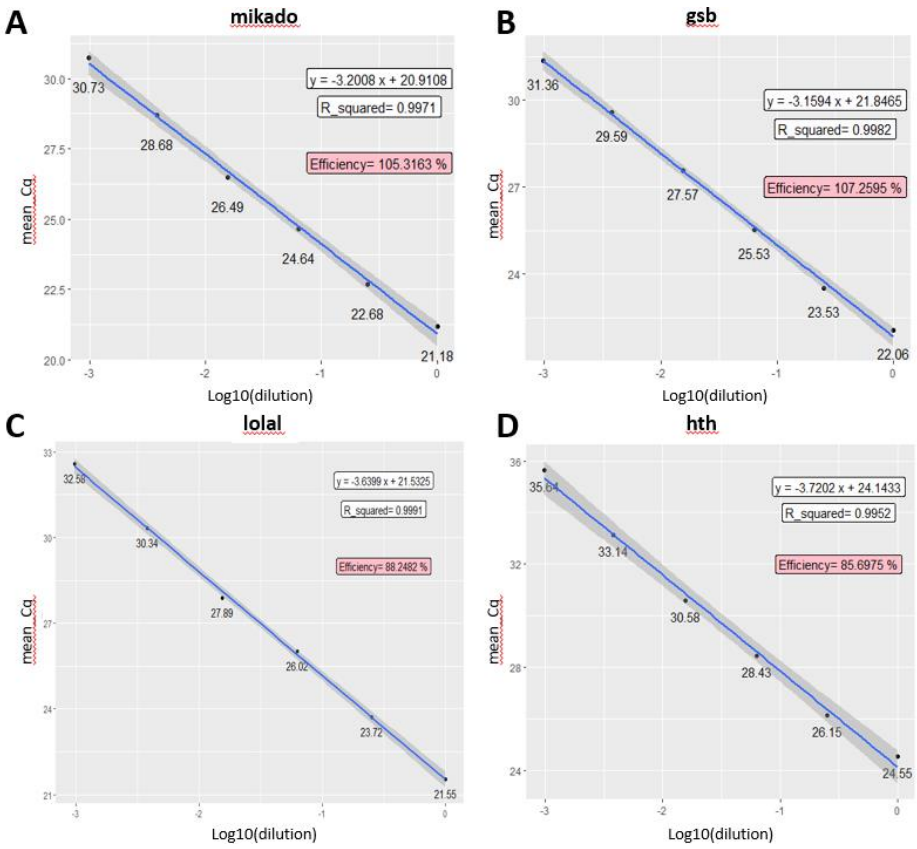


Figure 21. q-RT_PCR standard curves and efficiencies.

gsb	ΔΔCq	Expected
S13-S19	-3,08	-2,2
S13-S17	-1,58	
S17-S19	-1,50	

lolal	ΔΔCq	Expected
S13-S19	2,07	-0,85
S13-S17	1,20	
S17-S19	0,88	

hth	ΔΔCq	Expected
S13-S19	2,20	1,39
S13-S17	0,80	
S17-S19	1,39	

Tables 10-12. q-RT-PCR results for the three genes of interest. gsb and hth gave the expected results (decrease and increase respectively), but lolal showed an increase in its expression instead of the expected decrease.

I decided to then focus on gsb. gsb has been shown to be implicated in arthropod segmentation, a process that is of interest in Pavlopoulos Lab. Additionally, there were already available reagents for it in the lab, so gsb was a very good candidate for further experimentation.

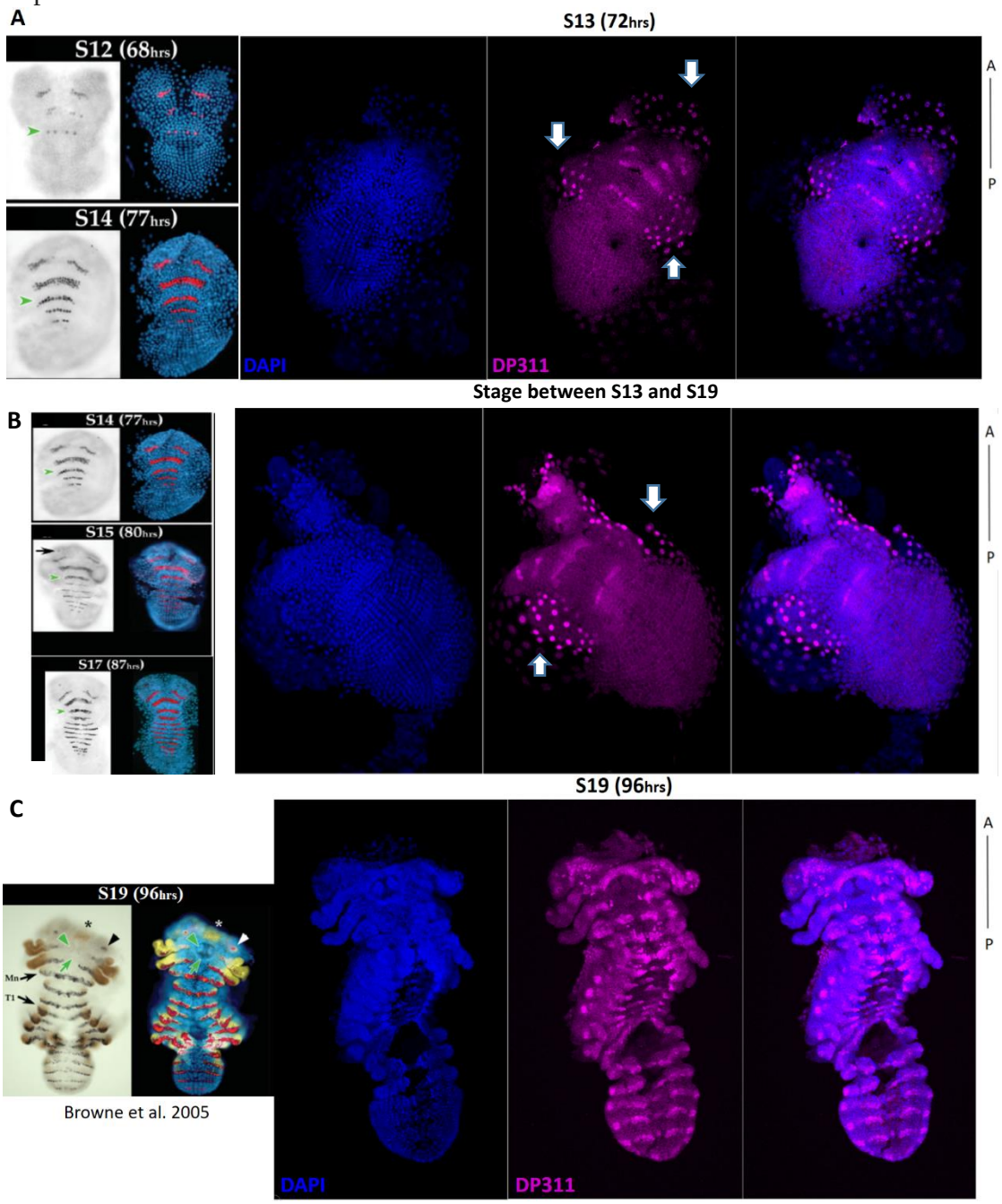


Figure 22. Immunostainings of S13 and S19 Parhyale embryos with the DP311 Ab. To the left the expected segmentation pattern as showed by Browne et al 2005. We observe a segment-like pattern in both developmental stages. The stages before S19 have an additional ectopic signal (white arrow)

The next step was to identify the location and pattern of gsb's expression. We decided to examine gsb as part of the Pax group family using a cross-species reactive antibody for gsb and other Pax3/7 proteins *(Davis et al. 2005)* and check whether it has the same strip-like pattern as in other arthropods. S13, S19 and embryos from intermediate stages were dissected, fixed and stained with the DP-311 primary mouse monoclonal Ab and a secondary anti-mouse-Alexa647 Ab. Immunostainings showed indeed a metameric stripe-like pattern in Parhyale germbands, which matched the pattern of the segments of the animal at that stage according to Browne et al. 2005 (Fig. 22). Immunostainings also showed an ectopic signal in the midgut (white arrows). Signal outside of the segments has also been observed to other arthropods and it is more likely due to cross reactivity of the DP311 to the HD domains of non-Pax3/7 proteins *(Davis et al. 2005, Davis et al. 2001)*.

Discussion

To summarize, combining gene expression data with chromatin accessibility data is very important and can provide many insights to the identification of eGRNs. For that purpose, a pipeline that analyzes both types of those datasets and performs a correlation analysis was established.

Although the pipeline was established in mouse data, it can be easily adjusted for the analysis of datasets from a variety of model organisms and experimental setups.
The application of the pipeline to different organisms has some prerequisites:

1. The genome has to be sequenced. Having a full genome assembly is not required but the genomic sequences and the total length of the genome are needed.
2. An annotation file containing at least the transcribed regions is necessary. Additional information on regulatory regions and whether the transcribed regions produce protein coding or non protein coding RNAs etc. is not required but can produce more information rich results. Also, information concerning enhancer locations and lncRNA genomic regions can help with the attribution of peaks to controlled genes and with making eGRN predictions with more confidence.

In general, it is very helpful to have the above information available in genome browsers, specified databases and/or integrated within the databases of tools. Having genomic and annotation information on browsers offers better visualization options. The whole process is automated, thus requiring less steps, and it provides detailed information not only for genes and other coding regions, but also on protein, regulatory regions and even conservation information. Using annotation files from main databases like Ensemble and NCBI allows for greater tool variety for multiple steps of the pipeline. Having more tools available can lead to more informative results that cover multiple biological processes. This is especially helpful for the attribution of regulatory regions to their controlling genes.
However, in case the above information is not available in that format there are alternative options. Several genome browsers allow the manual upload of genome and annotation files and are either available online or can be locally installed on the user's computer (e.g. IGV, NCBI Genome Workbench etc.). Also, all the steps of the pipeline can be performed with local annotation and genome files, but some procedures would have to be performed by the analyst himself, which usually requires some levels of experience, a deeper understanding of the process and is usually more time-consuming.

Concerning the different experimental setups, the backbone of this pipeline would remain the same, but the user would need to take into account a variety of parameters, like the origin and quality of the data, the different biological conditions, the number of replicates and the available tools related to that model organism. If the data are of poor quality, additional filtering steps might prove to be necessary in order to avoid compromising the signal-to-noise ratio and to ensure the credibility of the results. Great numbers of experimental conditions and different numbers of replicates might require more complex comparisons. The more the experimental conditions the more informative the whole experiment and the more the replicates the better statistical confidence of the results, especially in the differential analyses. The differential analysis tools that were used in this pipeline (DESeq2 and EdgeR) are capable of

performing comparisons and accounting for multiple experimental setups, but additional steps have to take place for the examination of their results.

The application of the pipeline to the Parhyale system turned out to be challenging. Although much progress has been made in recent years, Parhyale's genome and annotation are still lacking. As a result, some of the processes of the pipeline were hindered and large amounts of information were lost in several steps of both the ATAC-seq and the RNA-seq data analysis.

In the data that I worked with, considerable amounts of information were lost in the following steps due to the limited genomic information:

1. After the peak calling step, the contigs with length <2000bp were filtered out. That resulted in the loss of genomic information but also a loss of peaks. That is because the removed peaks were not only the ones with chromStart=0, but also the rest of the peaks on those contigs.
2. Parhyale's annotation does not include information concerning regulatory regions. During the step of the peak attribution to the nearest gene, only ~37% of the peaks were attributed, leaving the rest 63% out of the downstream analysis. However, the set threshold was very strict and can be easily changed in order to include more peaks in the downstream analysis.
3. During the attribution of transcript ids to common gene names only a small percentage of genes is kept. That is because no common gene information was available for the MSTRG transcript ids and for a small percentage of the mikado transcript ids. Also, some of the latter are matched with the same common gene name.

Additional information was lost due to the nature of our data:

1. In the quality control and trimming steps, the reads with length <50bp were removed to avoid having high background levels that would temper with the statistical analyses.
2. In the correlation step, of the initial 263 transcript ids, only 200 turned out to be correlating for both datasets.

In spite of the overall loss of data, the analysis produced numerous results that provide many useful insights concerning eGRNs and the mechanisms that mediate the transition from S13 to S19 in Parhyale's embryogenesis. The q-RT-PCR experiment, although performed in a limited number of genes, is an initial validation step that suggests that the RNA-seq results are most probably in agreement with the biological system. Of course, further experiments are required in order to establish the credibility of the RNA-seq results, especially when considering the fact that only 2 out of the 3 tested genes gave the expected outcome. It is yet to be determined whether the case of lolal is an isolated event or whether there is something fundamentally wrong with the RNA-seq data.

Further experiments are also needed in order to ensure the validity of ATAC-seq results with enhancer knock-in and knock-out experiments.

In order to validate that the peaks correspond indeed to enhancer regions we can perform knock-in experiments in 1-cell stage (S1) embryos. That would require the enhancer's sequence and a reporter gene along with its promoter. The construct can be integrated either with use of the minos transposon system or the CRISPR method, which are both established in Parhyale (Paris M et al. 2022). If that is indeed an enhancer which is active in those developmental stages, the reporter gene would be expressed in the cells that express the target gene of that enhancer.
In order to validate our eGRNs -that the enhancer regulates the predicted genes- we could perform an enhancer knock-out (or mutation) experiment and measure the difference in the expression of the target gene. Additionally, we could mutate enhancer region and check effects on predicted target genes by RT-qPCR. Of course eGRNs can have immense complexity. The expression of a gene can be regulated by more than one factors, which can be responsible for regulating the spatial and temporal expression patterns, the expression's intensity etc. But still, the above experiments would be a good first step to ensure the integrity of our ATAC-seq results.

Future Perspectives

For the pipeline

Although the established pipeline is a very useful starting tool for the analysis of ATAC-seq and RNA-seq data, additional improvements can be made in order to achieve more advanced integrative analyses. A first step would be to change the focus from enhancers to repressors and identify repressor-gene regulatory networks (rGRN). Changes in the accessibility of repressor binding sites between two conditions, due to changes in open regions borders possibly via nucleosome addition/removal or sliding, leads to differences in target gene expression. We would expect nucleosome shift/gain over an open region to mask repressor TFBS to lead to increased gene expression (increased RNA-seq logFC) and a decrease in the ATAC-seq over this region (decreased ATAC-seq logFC). The opposite would be expected from the removal of a nucleosome (Fig. 23A).
Nucleosomes positioning/fuzziness can have a regulatory role and act as a switch for gene expression (Lavigne M. D. et al. 2015). TFs can bind to target sites (TFBS) more frequently if those sites are covered with nucleosomes with lower affinity. If nucleosome remodeling occurs and exposes a repressor TF binding site, we would expect a decrease in gene expression of the target gene (decreased RNA-seq logFC) and ATAC peak could either be shifted or widened (increased ATAC-seq logFC) (Fig. 23B).
We could examine the downstream effects of both types of events and make predictions of rGRNs with the use of anti-correlating ATAC-seq with RNA-seq data.

When ATAC-seq data have high resolution, TF footprinting can be performed. As it was previously mentioned, when TFs are bound to the DNA during the ATAC-seq data production they give a very unique peak formation, called "footprint". By identifying the sequence of the TF's binding site we can make predictions for its recognition motif. By cross-referencing that binding motif to a TF binding site database (e.g. Jaspar

database *(Mathelier A. et al. 2013)*) we can identify the TF that was bound to the open chromatin region when we performed the ATAC-seq experiments *(Gusmao E. G. et al. 2014)*. Additionally, if that open region is part of an eGRN that was identified in the pipeline we can identify the TFs that regulate those eGRNs. If the DEG of the eGRN is also a TF then we can make presumptions on TF regulatory networks that might or might not be characteristic of the cell's identity *(Li Z. et al. 2019)*.
Motif analysis relies on the analysis of the sequences in the TF's footprint. However, this process is not that straightforward because those sequences can be affected by the Tn5 bias. There are tools that are used for bias correction and have been shown to increase the effectiveness of motif calling significantly. One such tool is TOBIAS *(Bentsen et al. 2020)*, that can perform both the bias correction and the motif analysis. The tool has been shown to reduce the false negative motifs (identification of greater numbers of TF binding sites) after bias correction. It can also perform differential TF analysis between conditions, as well as TF clustering and network construction if the experimental setup allows it *(Bentsen et al. 2020)*. The integration of TOBIAS in the pipeline would incorporate information concerning the proteins that play a role in gene regulation and provide a more complete understanding of the fine-tuning of our eGRN system.

Another future goal is to expand the capabilities of the pipeline in order to encompass additional datasets, apart from ATAC-seq and RNA-seq. Techniques used for the examination of chromatin or chromatin-protein interactions -like ChIP-seq, DNAse seq, CUT&TAG, CUT&RUN, FAIRE-seq, etc.- produce data whose analysis requires steps that are identical (or very similar) to the ones of the ATAC-seq data. The majority of those techniques give rise to data that produce peaks (peak calling), which are subsequently processed in a similar way as ATAC-seq derived peaks (comparison of common peaks between conditions etc.) (Fig. 23B) *(Tsompana M., Buck M. J. 2014)*. Therefore, the ATAC-seq data analysis pipeline could be adjusted for the processing of those data. Also, it would be very helpful to incorporate the results of the above techniques with ATAC-seq and RNA-seq results to get more detailed information for our eGRNs and to have a more complete overview of our biological system.

In 2020, Carullo et al. proposed that predictions concerning eGRNs can be done with the use of enhancer RNA (eRNA) data. It is known that during transcription, enhancer elements are transcribed bidirectionally by RNA pol II, which produces eRNAs that are not subsequently modified (by splicing, adding of poly-A tails etc.). eRNA production occurs prior to mRNA synthesis of target genes *(Arner E. et al. 2015)* and it is correlated with the presence of enhancer-promoter loops in target sites *(Li W. et al. 2016)*. Therefore, a way to enhance the reliability of the identification of eGRN is the use of eRNAs. For that, bulk instead of mRNA-seq data would be required because eRNAs are discarded prior to sequencing in the latter. We would also need an extra step in the pipeline that would include the correlation between DARs (ATAC-seq data) and site of bidirectional transcription (eRNA information). Seqmonk *(Popp C. et al. 2010)* is a tool that helps with the visualization of large amounts of NGS data and can be used for the identification of eRNAs and their correlation with the ATACseq data. The integration of eRNA data in our pipeline can give more precise eGRN predictions with greater confidence.

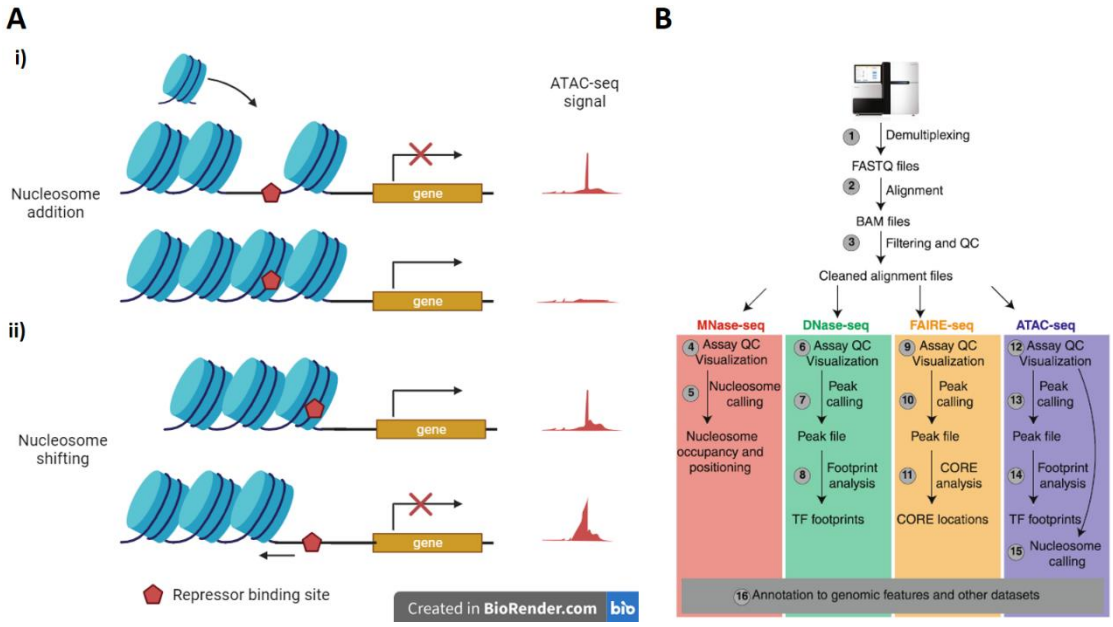


Figure 23. (A) The effects of nucleosome remodeling in repressor binding sites in gene expression. (i) The addition of a nucleosome in a repressor binding site leads to gene expression activation and a decrease in accessibility (ATAC-seq peak is deminished). (ii) The shift of a nucleosome away from a repressor binding site leads to gene expression inactivation and a increase in accessibility (ATAC-seq peak is widened). (B) Necessary steps for the analysis of MNAse-seq, DNAse-seq, FAIRE-seq and ATAC-seq data. The first 3 steps are common for all techniques and peak calling is performed in all but MNAse-seq. Footprinting analysis and nucleosome calling are common between ATAC-seq and DNAse-seq and ATAC-seq and MNAse-seq respectively. Therefore, an ATAC-seq data analysis pipeline can be easily adjusted for the analysis of the rest of the types of data.

For the Parhyale dataset

Concerning the *Parhyale hawaiensis* dataset, the next step after the validation of the results would be to reset the threshold for the peak attribution to genes. Taking into account a distance of +/-1Kb was a very strict threshold for the identification of enhancer-gene relationships, in order to minimize the false positive results rate. This resulted in only ~37% of the peaks being attributed. Expanding the threshold to longer distances would be a way to take advantage of the additional peak information that is otherwise lost. Of course we need to keep in mind that the greater the distance, the higher the chance of wrong attributions. Nevertheless, with proper validation experiments and as more and more information arises concerning regulatory elements in Parhyale, our results will get more reliable.

Additionally, the sequences of MSTRG transcript ids are now available. By repeating the initial process, that is by BLASTing the MSTRG sequences against the proteins of the Uniprot database, we could enrich the existing annotation file. That way we would avoid the additional loss of information due to the fact that no common gene names correspond to MSTRG ids, increase the results from the analysis and produce more eGRNs.

The validation of eGRN can also be done computationally with the use of HiC data. HiC is a process that allows the quantification of chromatin interactions inside the nucleus, in order to determine the spatial organization of the DNA. With this technique,

we can identify the chromatin's structure, which can have derived either randomly, or due to enhancer-promoter interactions (chromatin loops). Additionally, we can discover regions of open and closed chromatin (A/B compartments) and Topologically Associating Domains (TADs). TADs are chromatin regions with increased interaction compared to other adjacent regions and have been suggested to play a role in the regulation of gene expression *(Gong H. et al. 2021)*. By combining the ATAC-seq derived eGRN with the HiC information of chromatin loops we can further ensure the correctness of the peak attribution to their regulating genes.

Materials and Methods

The pipeline

Mus musculus ATAC-seq

Quality control and Trimming were performed with TrimGalore (v. 0.6.4_dev) *(Krueger F. et al., 2021)* with default settings (Phred score threshold: 20) for paired-end data. Alignment was done with BWA (v. 0.7.15-r1142-dirty) *(Li H. 2013)* and peak calling with macs2 (v. 2.1.1.20160309) *(Zhang Y. et al. 2008)* with default settings. The filtering of the peaks was done manually in R. For the construction of the gff file necessary for featureCounts, the bam files were merged with Samtools (v. 1.3.1 (using htslib 1.3.2)) *(Danecek P. et al. 2021)*, peaks were called with same settings as previously and the narrowPeak files were converted to gff with bed2gtf *(Pfurio 2014)*. Read quantification was performed with featureCounts (v2.0.3) *(Liao Y. et al., 2014)* for paired-end data, with the options *–countReadPairs* for the counting at the level of fragments instead of reads and *-O* for the assignment of reads to all their overlapping meta-features in case they overlap with more than one feature. Differential accessibility analysis was done with DESeq2 (v. 1.28.1) *(Love M.I. et al. 2014)* in R. The attribution of peaks to genes was done with PAVIS2 for *p-value 0.05, 30k upstream, 10k downstream*.

Mus musculus RNA-seq

Quality control and Trimming were performed with TrimGalore (v. 0.6.4_dev) *(Krueger F. et al., 2021)* with default settings (Phred score threshold: 20) for paired-end data. Alignment was done with HISAT2 (v. 2.1.0) *(Kim D. et al. 2019)* with default settings. Read quantification was performed with featureCounts (v2.0.3) *(Liao Y. et al., 2014)* for paired-end data, with the options *–countReadPairs* for the counting at the level of fragments instead of reads. Differential accessibility analysis was done with EdgeR/Sartools (v. 1.5.0) *(Varet H. et al., 2016)*.

Parhyale hawaiensis ATAC-seq

Quality control and Trimming were performed with TrimGalore (v. 0.6.4_dev) *(Krueger F. et al., 2021)* with *--length 50* to filter out reads with length <50bp and Phred score threshold: 20 (default) for paired-end data. Alignment was done with Bowtie2 (v. 2.3.5.1) *(Langmead B. et al. 2012)* and peak calling with macs2 (v. 2.2.7.1) *(Zhang Y. et al. 2008)* with *--call-summits --nomodel -f BAMPE --buffer-size 10000*. For the construction of the gff file necessary for featureCounts, the bam files were merged with Samtools (v. 1.10) *(Danecek P. et al. 2021)* and peaks were called with same settings as previously. The filtering of the contigs with

peaks at their start position (chromStart=0) and length <2000 was done manually in R. The narrowPeak files were then converted to gff with bed2gtf *(Pfurio 2014)*. Read quantification was performed with featureCounts (v. v2.0.0) *(Liao Y. et al., 2014)* for paired-end data, with the option *-O* for the assignment of reads to all their overlapping meta-features in case they overlap with more than one feature. Differential accessibility analysis was done with DESeq2 (v. 1.38.3) *(Love M.I. et al. 2014)* in R. The attribution of peaks to genes was done manually in bash and the attribution of transcript ids to common gene names was done with a python script made by Rallis in Pavlopoulos Lab.

Parhyale hawaiensis RNA-seq

Quality control and Trimming were performed with TrimGalore (v. 0.6.4_dev) *(Krueger F. et al., 2021)* with default settings (Phred score threshold: 20) for paired-end data. Alignment was done with HISAT2 (v. 2.1.0) *(Kim D. et al. 2019)* with default settings. Read quantification was performed with Kallisto (v. 0.46.1) *(Bray N. L. et al. 2016)* for paired-end data, with the options *--bias -b 40 --single-overhang*. Differential accessibility analysis was done with DESeq2 (v. 1.38.3) *(Love M.I. et al. 2014)* in R.

Visualization and correlation analysis of *Parhyale hawaiensis*

Sam to bam conversion and bam sorting and indexing was done with Samtools (1.10) *(Danecek P. et al. 2021)*. Conversion to bigwig for visualization purposes was done with bamCoverage (v. 3.5.1) *(Ramírez F. et al. 2016)* with *--binSize 10 --normalizeUsingRPKM --maxFragmentLength 0* and the tracks were uploaded on UCSC *(Kent W.J. et al. 2002)* and Apollo *(Dunn et al. 2019)*. The correlation analysis was performed in R with the ggstatsplot package *(Patil I. 2021)*.

RNA-seq data validation – Gene expression quantification

Embryo collection and freezing

Parhyale embryos were staged using the Browne et al. (2005) staging guide. The embryos were collected from the developmental stages S1 to S5 and were raised in FASWA at 26°C, till they reached the desired stages. They were isolated at S13 (c_1=170 cells), S17 (c_2=112) and S19 (c_3=127). Because S19 is a very long stage (~8h: 96h-111h after lay) we chose to isolate embryos from the middle of the stage at ~104h. The embryos were pulled down by spinning at 1000 rpm for 1sec and excess FASWA was removed. The embryos were flash-frozen in dry ice with 10-20ml ethanol (or rarely, in liquid nitrogen). The frozen embryos were stored at -80°C.

RNA Isolation

Embryos were thawed and 300µl Trizol Reagent *(Thermo Fischer-Ambion, Catalog number: 15596026)* were used for the lysis of all the embryos (c_{total}=409 cells). Trizol Reagent includes chaotropic agents to denature proteins and RNAses for the protection of the RNA from degradation. The embryos were subsequently lysed with a pestle (2 min) and by vortexing for 1min. 5µg LPA *(Sigma Aldrich, Product number: 56575)* were used as a nucleic acid carrier and 60µl of choloform *(Merck Catalog number: 1024451000)* was used as a phase separation reagent followed by vortexing for 30sec. The samples were centrifuged at 13.000 rpm for 5 min at 4°C, in order for their components to be separated in 3 layers. The aqueous (top) phase included the RNA, which was transferred to a new tube. 176µl of isopropanol *(2-propanol, Merck, Catalog*

number: 1096342511) were added and the samples were put at -20°C, for 2h for the precipitation of the RNA. The samples were centrifuged at 13.000 rpm for 30 min at 4°C and the supernatant was discarded. The RNA pellet was washed with 500μl of 70% ethanol *(VWR, avantor, Catalog number: 20821.365)* and centrifuged at 13.000 rpm for 5min at 4°C. After the removal of the ethanol, the samples were air dried and resuspended in nanopure water *(Thermo Fischer-Invitrogen, Catalog number: 10977035)* to a final volume of 10μl. The concentration of the RNA was determined in Nanodrop (1μl) *(ND-1000 Spectrophotometer)*. The samples were stored at -80°C.

cDNA synthesis

2.9μg of RNA were used for each sample. The samples were thawed and treated with DNAse for the degradation of any possible DNA molecules in the sample. The buffer for the cDNA synthesis with oligo$(dT)_{18}$ primers was added in the same tube and the samples were incubated at 25°C for 10min, followed by 30min at 50°C (At 15min the samples were spun down and incubation continued). The reactions were terminated by heating the samples at 85°C for 5min. The samples were stored at -80°C.

Primer design

The primers were designed within 1 exon only. The exonic regions were isolated on Apollo *(Dunn et al. 2019)*. The exons were blasted against the genome in chrysallida *(Priyam A. et al.2019)*, in order to identify regions with polymorphisms and avoid making primers complementary to those sites. The primers were designed on Primer 3 *(Koressaar T. and Remm M. 2017)*, with the following settings (Table 13):

	Min	**Opt**	**Max**
Primer Size	18	20	23
Primer Tm	58	60	62
Product Tm	-1000000	0	1000000
Primer %GC	40	50	60
Product size ranges	150		200

Tables 13. Primer3 settings for the design of the primers.

The primers were then blasted again against the genome to verify whether they would bind to >1 sites. Most of them were found to bind to several target site. The final primers along with their characteristics and the number of target sites that they are complementary with are shown on table 14.

Gradient end-point PCR

The reaction was done for 5 sets of primers: gsb-exon3, gsb-exon2 mikado.phaw_50.283872cG118.1, lolal and hth (Table 14) , for 30 cycles, at 3 different temperatures: 55°C, 60°C, 65°C. I used 1μl of each cDNA sample to make a cDNA pool of a total volume of 30μl. PCR was performed with Taq 2X Master Mix (M0270) in a reaction volume of 20μl. 1μl of template was added to each reaction from the cDNA

pool. The thermocycling conditions are on the table 15 below. The BIO-RAD C1000 Touch Thermal Cycler (*Product number: 1851148*) was used.

	length	**Tm**	**%GC**	**product size**	**sequence**	**Number of target sites**
gsb-exon3	20	60.46	60	180	CCATTCGACCCAGCAGTACC	x1
	20	60.95	60		GTGGTACGAGGGCTGGTTTG	x5
gsb-exon2	20	60.32	55	158	AGTGACGTGCCAGCAACTAG	x5
	20	57.57	58		CTCGAGCTTCCAAGTGAGG	x2
mikado.phaw_50.283872cG118.1	20	60.04	55	187	TTGAACAGCTCCGGGACATC	x1
	20	60.11	55		ATCATGAACTTGGGGCCGAG	x2
lolal	20	59.31	55	163	AGGGATGGATTCGGATGAGC	x2
	20	59.31	55		TGGAGGAGAGTTGGAGTTGC	x2
hth	18	58.03	56	178	ACAATGGTTCTGCGACGG	x1
	20	59.74	60		GGTTGGTGGGATAGTGGACC	x1

Tables 14. Final primers designed on Primer3 and their characteristics.

Step	**Temperature**	**Time**
Initial Denaturation	95°C	30sec
Denaturation (x30)	95°C	30sec
Annealing (x30)	55, 60, 65°C	30sec
Extension (x30)	68°C	15sec
Final Extension	68°C	5min
Hold	10°C	∞

Tables 15. Overview of the thermocycler parameters fot the end-point PCR.

Quantitative Real Time PCR

The reactions were done for 4 sets of primers: gsb-exon3, lolal, hth and mikado.phaw_50.283872cG118.1, for 40 cycles. The standard curves were done in duplicates and the conditions and negative controls were tested in triplicates. q-RT-PCR was performed with EnzyQuest 2x qPCR Master Mix Green kit, w/o ROXTM (Cat No: RN014S) in a reaction volume of 20µl. For the standard curves, I used 10µl

of each cDNA sample to make a cDNA pool of a total volume of 300µl. 5µl of template were added from the cDNA pool to each reaction. For the conditions, I used all the remaining cDNA (S13=9µl, S17=7.7µl, S19=5.6µl) that was diluted to 1:20. 5µl of template were added to each reaction. The thermocycling conditions are on the table 16 below. The efficiencies of the reactions were calculated with the linear regression method in R. The quantification was done in excel with the ΔΔCt method (Equation 3). The BIO-RAD C1000 Touch Thermal Cycler CFX96 Real Time System (*Product number: 1845096*) was used.

$$\Delta Cq_{S13/S19} = Cq_{mean.of.gene} - Cq_{mean.of.internal.conrol}$$
$$\Delta\Delta Cq_{S19vsS13} = -(\Delta Cq_{S19} - \Delta Cq_{S13})$$

Equation 3. Calculation of ΔCq and ΔΔCq for the analysis of the q-RT-PCR results.

Step	**Temperature**	**Time**
Initial Denaturation	95°C	15min
Denaturation (x40)	95°C	30sec
Annealing (x40)	60°C	30sec
Extension (x40)	68°C	15sec
Hold	10°C	∞

Tables 16. Overview of the thermocycler parameters fot the q-RT-PCR.

Stainings

Embryos from developmental stages S13, S17 and S19 were dissected and fixed in 4% PFA for 30min. The dissected embryos were left on PBS on ice for synchronization. The embryos were washed x3 with PT and treated with methanol (met) in Room Temperature (RT). The methanol treatment consist of a dehydration and a re-hydration step. In the dehydration step the embryos are successively treated with met in quantities 25%, 50%, 75% and finally with 2 washes of 100%. They can then be stored at -20°C. They can then be rehydrated with successive washes of met in concentrations 75%, 50%, 25% and 2 washes of PT in RT. The embryos are then incubated in PBT for 15min and in PBT-5%NGS for 30min in RT. The embryos are treated overnight (O/N) with the primary Ab m-DP311, 1:20 diluted in PBT at 4°C, in agitation. They are then washed in PT x3 for 10min and x4 for 30min in RT. They are again incubated in PBT for 15min and in PBT-5%NGS for 30min in RT. The embryos are incubated with the secondary Ab anti-mouse-Alexa647, 1:500 diluted in PBT for 2h in RT. They are washed with PT x3 for 10min and x3 for 30min and incubated in DAPI 1:500 diluted in PT for 30min. They are finally treated with 50% and 70% glycerol until they reach the bottom of the glass well. The samples are then mounted with DABCO mix.

PBS: 137mM NaCl, 2.7mM KCl, 10mM Na2HPO4, 2mM KH2PO4
PT: PBS + Tween?
PBT: PT + 0.1%Triton, 1%BSA
DABCO mix: 90% glycerol + PBS + 5%DABCO

References

- Andrews S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- Arner E. et al. (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*. 347(6225):1010-4. doi: 10.1126/science.1259418.

- Balczarek K. A., Lai Z. C., Kumar S. (1997) Evolution and functional diversification of the paired box (Pax) DNA- binding domains. *Molecular Biology and Evolution* 14, 8, 829-842.

- Benjamini Y., Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57, 1, 289-300

- Berg D. E. (1989) in Berg D. E. & Howe M. M., eds. Mobile DNA (Am. Soc. Microbiol., Washington, DC), pp. 109–162.
- Bentsen M., Goymann P., Schultheis H. et al. (2020) ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* 11, 4267. https://doi.org/10.1038/s41467-020-18035-1

- Bray N. L., Pimentel H., Melsted P., Pachter L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 34(5):525-7. https://doi.org/10.1038/nbt.3519

- Brody T. (1999) The Interactive Fly: gene networks, development and the Internet. *Trends in Genetics* 15 (8): 333-4. 10431196

- Browne W. E., Price A. L., Gerberding M., Patel, N. H. (2005) Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiensis*. *Genesis* 42, 124-149. doi:10.1002/gene.20145

- Bruce H. S. (2017) Expression and function of leg gap genes in the amphipod crustacean, *Parhyale hawaiensis*. *University of California, Berkeley*.

- Buenrostro J., Giresi P., Zaba L. et al. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 1213–1218. https://doi.org/10.1038/nmeth.2688

- Buenrostro J., Wu B., Chang HY.., Greenleaf W.J. (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol.* Jan 5;109:21.29.1-21.29.9. doi: 10.1002/0471142727.mb2129s109.

- Carullo N. V. N., Phillips Iii R. A., Simon R. C., Soto S. A. R., Hinds J. E., Salisbury A. J., Revanna J. S., Bunner K. D., Ianov L., Sultan F. A., Savell K. E., Gersbach C. A., Day J. J. (2020) Enhancer RNAs predict enhancer-gene regulatory links and are critical for enhancer function in neuronal systems. *Nucleic Acids Res*. 48(17):9550-9570. doi: 10.1093/nar/gkaa671.

- Chen Y., Lun A.A.T., Smyth G.K. (2016) From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5, 1438. doi:10.12688/f1000research.8987.2.

- Corces M., Trevino A., Hamilton E. et al. (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* 14, 959–962 . https://doi.org/10.1038/nmeth.4396

- Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M., Li H. (2021) Twelve years of SAMtools and BCFtools, *GigaScience*, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

- Davis G.K., D'Alessio J.A., Patel N.H. (2005) Pax3/7 genes reveal conservation and divergence in the arthropod segmentation hierarchy. *Dev Biol*. Sep 1;285(1):169-84. doi: 10.1016/j.ydbio.2005.06.014.

- Davis G.K., Jaramillo C.A., Patel N.H. (2001) Pax group III genes and the evolution of insect pair-rule patterning. *Development*. Sep;128(18):3445-58. doi: 10.1242/dev.128.18.3445.

- dos Santos A., Botelho M.T., Vannuci-Silva M., Artal M.C., Vacchi F. I., Magalhães G. R., Gomes V., Henry T. B., Umbuzeiro G. A. (2022) The amphipod *Parhyale hawaiensis* as a promising model in ecotoxicology, Chemosphere, Vol. 307, Part 2, https://doi.org/10.1016/j.chemosphere.2022.135959.

- Dunn N. A., Unni D. R., Diesh C., Munoz-Torres M., Harris N. L., Yao E., Rasche H., Holmes I. H., Elsik C. G., Lewis S. E. (2019) Apollo: Democratizing genome annotation. *PLoS Comput Biol*. 15(2):e1006790. doi: 10.1371/journal.pcbi.1006790.

- Faucheux M., Roignant J.-Y., Netter S., Charollais J., Antoniewski C. & Théodore L. (2003) batman Interacts with Polycomb and trithorax Group Genes and Encodes a BTB/POZ Protein That Is Included in a Complex Containing GAGA Factor, *Molecular and Cellular Biology*, 23:4, 1181-1195, DOI: 10.1128/MCB.23.4.1181-1195.2003

- Gerberding M., Browne W.E., Patel N.H. (2002) Cell lineage analysis of the amphipod crustacean Parhyale hawaiensis reveals an early restriction of cell fates. *Development*. Dec;129(24):5789-801. doi: 10.1242/dev.00155.

- Gong H., Yang Y., Zhang S., Li M., Zhang X. (2021) Application of Hi-C and other omics data analysis in human cancer and cell differentiation research, Computational and Structural Biotechnology Journal, Volume 19, Pages 2070-2083. https://doi.org/10.1016/j.csbj.2021.04.016

- Gontarz P., Fu S., Xing X. et al. (2020) Comparison of differential accessibility analysis strategies for ATAC-seq data. *Sci Rep* 10, 10150 https://doi.org/10.1038/s41598-020-66998-4

- Goryshin I.Y., Miller J.A., Kil Y.V., Lanzov V.A., Reznikoff W.S. (1998) Tn5/IS50 target recognition. *Proc Natl Acad Sci USA*. Sep 1;95(18):10716-21. doi: 10.1073/pnas.95.18.10716

- Gramates L.S., Agapite J., Attrill H., Calvi B.R., Crosby M., dos Santos G. Goodman J.L., Goutte-Gattat D., Jenkins V., Kaufman T., Larkin A., Matthews B., Millburn G., Strelets V.B., and the FlyBase Consortium (2022) FlyBase: a guided tour of highlighted features. *Genetics*, Volume 220, Issue 4, April 2022, iyac035

- Gusmao E. G., Dieterich C., Zenke M., Costa I. G. (2014) Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*. 30(22):3143-51. doi: 10.1093/bioinformatics/btu519.

- Huang W., Loganantharaj R., Schroeder B., Fargo D., Li L. (2013) PAVIS: a tool for Peak Annotation and Visualization, *Bioinformatics*, 29, 23, 3097–3099, https://doi.org/10.1093/bioinformatics/btt520

- Illumina Inc (2017) An introduction to Next-Generation Sequencing Technology https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

- Jaksik R., Drobna-Śledzińska M., Dawidowska M. (2021) RNA-seq library preparation for comprehensive transcriptome analysis in cancer cells: The impact of insert size, *Genomics*, 113, 6, p. 4149-4162 https://doi.org/10.1016/j.ygeno.2021.10.018

- Kao D., Lai A.G., Stamataki E., Rosic S., Konstantinides N., Jarvis E., Di Donfrancesco A., Pouchkina-Stancheva N., Sémon M., Grillo M., Bruce H., Kumar S., Siwanowicz I., Le A., Lemire A., Eisen M.B., Extavour C., Browne W.E., Wolff C., Averof M., Patel N.H., Sarkies P., Pavlopoulos A., Aboobaker A. (2016) The genome of the crustacean *Parhyale hawaiensis*, a model for

animal development, regeneration, immunity and lignocellulose digestion. *Elife*. Nov 16;5:e20062. doi: 10.7554/eLife.20062.

- Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. (2002) The human genome browser at UCSC. *Genome Res.* Jun;12(6):996-1006. http://genome.ucsc.edu

- Kent W.J., Zweig A.S., Barber G., Hinrichs A.S., Karolchik D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. Sep 1;26(17):2204-7. doi: 10.1093/bioinformatics/btq351.

- Kim D., Paggi J. M., Park C. et al. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915.

- Kingston R. E., Tamkun J. W. (2014) Transcriptional regulation by trithorax-group proteins. *Cold Spring Harb Perspect Biol.* 6(10):a019349. doi: 10.1101/cshperspect.a019349

- Klemm S. L., Shipony Z. & Greenleaf W. J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 20, 207–220. https://doi.org/10.1038/s41576-018-0089-8

- Koressaar T. and Remm M. (2017) Enhancements and modifications of primer design program Primer3 *Bioinformatics* 23(10):1289-91.

- Krueger F., James F., Ewels P., Afyounian E. & Schuster-Boeckler B. (2021) FelixKrueger/TrimGalore: v0.6.4 - DOI via Zenodo (0.6.4). Zenodo. DOI 10.5281/zenodo.5127898

- Langmead B., Salzberg S. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9:357-359.

- Lavigne M. D., Vatsellas G., Polyzos A., Mantouvalou E., Sianidis G., Maraziotis I., Agelopoulos M., Thanos D. (2015) Composite macroH2A/NRF-1 Nucleosomes Suppress Noise and Generate Robustness in Gene Expression. *Cell Rep*. 19;11(7):1090-101. doi: 10.1016/j.celrep.2015.04.022.

- Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

- Li W., Notani D. and Rosenfeld M. G. (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev*. Genet., 17, 207–223.

- Li X., Noll M. (1994) Evolution of distinct developmental functions of three *Drosophila* genes by acquisition of different *cis*-regulatory regions. *Nature* 367, 83–87 https://doi.org/10.1038/367083a0

- Li Z., Schulz M.H., Look T. et al. (2019) Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* 20, 45. https://doi.org/10.1186/s13059-019-1642-2

- Liao Y., Smyth G.K. and Shi W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30.

- Love M.I., Huber W., Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550. 10.1186/s13059-014-0550-8.

- Mathelier A., Zhao X., Zhang A. W., Parcy F., Worsley-Hunt R., Arenillas D. J., Buchman S., Chen C-y, Chou A., Ienasescu H., et al. (2013) Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 42(D1):142–7.

- Paris M., Wolff C., Patel N.H., Averof M. (2022) The crustacean model Parhyale hawaiensis. *Curr Top Dev Biol.* 147:199-230. doi: 10.1016/bs.ctdb.2022.02.001. Epub 2022 Mar 14.

- Patil I. (2021) Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, 6(61), 3167. doi:10.21105/joss.03167.

- Pervez M.T., Hasnain M.J.U., Abbas S.H., Moustafa M.F., Aslam N., Shah S.S.M. (2022) A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *Biomed Res Int.* Sep 29;2022:3457806. doi: 10.1155/2022/3457806.

- Pfurio (2014) bed2gtf.py https://github.com/pfurio/bed2gtf

- Popp C., Dean W., Feng S., Cokus S. J., Andrews S., Pellegrini M., Jacobsen S. E., Reik W. (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*. 463(7284):1101-5. doi: 10.1038/nature08829.

- Prpic N. M., Telford M.J. (2008) Expression of homothorax and extradenticle mRNA in the legs of the crustacean Parhyale hawaiensis: evidence for a reversal of gene expression regulation in the pancrustacean lineage. *Dev Genes Evol.* Jun;218(6):333-9. doi: 10.1007/s00427-008-0221-4.

- Priyam A., Woodcroft B.J., Rai V., Moghul I., Munagala A., Ter F., Chowdhary H., Pieniak I., Maynard L.J., Gibbins M.A., Moon H., Davis-Richardson A., Uludag M., Watson-Haigh N.S., Challis R., Nakamura H., Favreau E., Gómez E.A., Pluskal T., Leonard G., Rumpf W., Wurm Y. (2019) Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol Biol Evol.* 36(12):2922-2924. doi: 10.1093/molbev/msz185.

- Qin D. (2019) Next-generation sequencing and its clinical application. *Cancer Biol Med.* Feb;16(1):4-10. doi: 10.20892/j.issn.2095-3941.2018.0055.

- Quijano J. C., Wisotzkey R. G., Tran N. L., Huang Y., Stinchfield M. J., Haerry T. E., Shimmi O., Newfeld S. J. (2016) lolal Is an Evolutionarily New Epigenetic Regulator of dpp Transcription during Dorsal-Ventral Axis Formation. *Mol Biol Evol*. 33(10):2621-32. doi: 10.1093/molbev/msw132.

- Ramírez F., Ryan D.P., Grüning B., Bhardwaj V., Kilpert F., Richter A.S., Heyne S., Dündar F., and Manke T. (2016) deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. *Nucleic Acids Research.* doi:10.1093/nar/gkw257.

- Rieckhof G.E., Casares F., Ryoo H.D., Abu-Shaar M., Mann R.S. (1997) Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. *Cell*. Oct 17;91(2):171-83. doi: 10.1016/s0092-8674(00)80400-6.

- Robinson M. D., McCarthy D. J., Smyth G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 26(1):139-40. doi: 10.1093/bioinformatics/btp616.

- Sanger F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. Dec;74(12):5463-7. doi: 10.1073/pnas.74.12.5463

- Stark R., Grzelak M. & Hadfield J. (2019) RNA sequencing: the teenage years. *Nat Rev Genet* 20, 631–656. https://doi.org/10.1038/s41576-019-0150-2

- Sun D., Bredeson J., Bruce H., Patel N. (2022) Identification and classification of cis-regulatory elements in the amphipod crustacean *Parhyale hawaiensis*. *Development (Cambridge)*, 149, 11. doi:10.1242/dev.200793

- The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research*, Volume 51, Issue D1, 6 p. D523–D531. https://doi.org/10.1093/nar/gkac1052

- Thompson B., Davidson E.A., Liu W., Nebert D.W., Bruford E.A., Zhao H., Dermitzakis E.T., Thompson D.C., Vasiliou V. (2021) Overview of PAX gene family: analysis of human tissue-specific variant expression and involvement in

human disease. *Hum Genet.* Mar;140(3):381-400. doi: 10.1007/s00439-020-02212-9.

- Tsompana M., Buck M.J. (2014) Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* 7, 33. https://doi.org/10.1186/1756-8935-7-33

- Varet H., Brillet-Guéguen L., Coppee J.-Y. and Dillies M.-A. (2016) SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data, *PLoS One*. doi: http://dx.doi.org/10.1371/journal.pone.0157022

- Wang J., Kong L., Gao G., Luo J. (2013) A brief introduction to web-based genome browsers. *Brief Bioinform.* Mar;14(2):131-43. doi: 10.1093/bib/bbs029. Epub 2012 Jul 3. PMID: 22764121.

- Wang Z., Gerstein M. & Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57–63. https://doi.org/10.1038/nrg2484

- Yan F., Powell D.R., Curtis D.J. et al. (2020) From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 21, 22. https://doi.org/10.1186/s13059-020-1929-3

- Xue L., Noll M. (1996) The functional conservation of proteins in evolutionary alleles and the dominant role of enhancers in evolution. *EMBO J.* Jul 15;15(14):3722-31.

- Yi L., Liu L., Melsted P., Pachter L. (2018) A direct comparison of genome alignment and transcriptome pseudoalignment. *bioRxiv*, 444620.

- Zhang H., Lu T., Liu S., Yang J., Sun G., Cheng T., Xu J., Chen F., Yen K. (2021) Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification. *NAR Genom Bioinform.* Oct 27;3(4):lqab094. doi: 10.1093/nargab/lqab094.

- Zhang Y., Liu T., Meyer C.A. et al. (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 https://doi.org/10.1186/gb-2008-9-9-r137

- Zhou Y., Zhou B., Pache L. et al. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10, 1523. https://doi.org/10.1038/s41467-019-09234-6

- Zielezinski A., Vinga S., Almeida J. et al. (2017) Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol* 18, 186. https://doi.org/10.1186/s13059-017-1319-7

Sources of figures:

Figure 1A:
David C., Basturea G. (2016) RNA-seq Using Next Generation Sequencing A comprehensive review of RNA-seq methodologies, *MATER METHODS*, 3 DOI: 10.13070/mm.en.3.203

Figure 1B:
www.genewiz.com

Figure 1C:
Buenrostro J., Giresi P., Zaba L. et al. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218.
https://doi.org/10.1038/nmeth.2688

Figure 1D:
Yan, F., Powell, D.R., Curtis, D.J. et al. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 21, 22 (2020).
https://doi.org/10.1186/s13059-020-1929-3

Figure 2A:
Ramos A. (2017) Exploring Sensory Function and Evolution in the Crustacean Visual System. Neurobiology. Université de Lyon.

Figure 2C:
Thompson B., Davidson E.A., Liu W., Nebert D.W., Bruford E.A., Zhao H., Dermitzakis E.T., Thompson D.C., Vasiliou V. (2021) Overview of PAX gene family: analysis of human tissue-specific variant expression and involvement in human disease. *Hum Genet.* 140(3):381-400. doi: 10.1007/s00439-020-02212-9

Figure 3A-C, 11:
Andrews S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data Available online at:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Figure 3D:
https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lectures/alignment_quantification.pdf

Figure 4, 12, 13:
García-Alcalde F., Okonechnikov K., Carbonell J., Cruz L. M., Götz S., Tarazona S., Dopazo J., Meyer T. F., Conesa A. (2012) Qualimap: evaluating next-generation sequencing alignment data, *Bioinformatics*, 28, 20, 2678–2679

Figure 5:
Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. (2002) The human genome browser at UCSC. Genome Res. Jun;12(6):996-1006. http://genome.ucsc.edu

Figure 6A-D, 14, 15:
Love M.I., Huber W., Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550. 10.1186/s13059-014-0550-8.

Figure 6E:
Huang W., Loganantharaj R., Schroeder B., Fargo D., Li L. (2013) PAVIS: a tool for Peak Annotation and Visualization, *Bioinformatics*, 29, 23, 3097–3099, https://doi.org/10.1093/bioinformatics/btt520

Figure 7A-C:
Varet H., Brillet-Guéguen L., Coppee J.-Y. and Dillies M.-A. (2016) SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data, *PLoS One*. doi: http://dx.doi.org/10.1371/journal.pone.0157022

Figure 7D, 16B:
Chen H., Boutros P. C. (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12, 35 https://doi.org/10.1186/1471-2105-12-35

Figure 8, 17:
Patil I. (2021) Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, 6(61), 3167. doi:10.21105/joss.03167.

Figure 9, 18:
Gu Z., Eils R., Schlesner M. (2016). "Complex heatmaps reveal patterns and correlations in multidimensional genomic data." *Bioinformatics*. doi:10.1093/bioinformatics/btw313.

Figure 16A:
Oliveros, J. C. (2007-2015) Venny. An Interactive Tool for Comparing Lists with Venn's Diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html

Figure 19:
Zhou Y., Zhou B., Pache L. et al. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10, 1523. https://doi.org/10.1038/s41467-019-09234-6

Figure 22:
Browne W. E., Price A. L., Gerberding M., Patel, N. H. (2005) Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiensis*. *Genesis* 42, 124-149. doi:10.1002/gene.20145

Figure 23:
Tsompana M., Buck M. J. (2014) Chromatin accessibility: a window into the genome. Epigenetics & Chromatin 7, 33. https://doi.org/10.1186/1756-8935-7-33