Spatiotemporal control of gene expression in *Caenorhabditis elegans*

Jacques Serizay

Department of Genetics

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Churchill College

March 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text or in the Acknowledgment and Foreword sections.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification.

This thesis does not exceed the prescribed word limit for the Biology Degree Committee.

Jacques Serizay
March 2020

Abstract

Cell-type specific regulation of transcription drives the production of the myriad of different cells generated during development. Profiling genome-wide gene expression landscapes in different tissues has improved our understanding of the physiological and pathological processes taking place during development. Yet, the mechanisms underlying cell-type specific transcription are not well understood. Promoters and enhancers are the key loci that orchestrate spatiotemporal patterns of gene expression. Their activities can range from ubiquitous to highly cell-type specific, and their composition and arrangement define the regulatory grammar directing gene transcription across development. More comprehensive *in vivo* studies of these regulatory grammars would improve our understanding of how different patterns of gene expression are obtained across tissues.

Caenorhabditis elegans is an important model organism for studying developmental processes. At the beginning of my PhD, I helped characterize the dynamics of gene expression and chromatin activity across development and aging. Following this, I aimed to identify and characterize the regulatory elements involved in tissue-specific control of transcription in *C. elegans.* I jointly profiled chromatin accessibility and gene expression landscapes across the five main tissues of the adult nematode. To achieve this, I developed a method to sort fluorescently labelled nuclei from individual *C. elegans* tissues. Analyzing the datasets I generated, I first showed that around 80% of the regulatory elements in *C. elegans* are specifically active in subsets of tissues. I then revealed fundamental differences in the genetic structure and regulatory architecture of genes expressed ubiquitously or in individual tissues, and I defined two distinctive regulatory grammars associated with specific sets of genes. I also uncovered striking and unsuspected differences in nucleosome arrangement and sequence features of ubiquitous and germline-specific promoters compared to somatic promoters. Finally, I optimized a single nucleus method to analyze chromatin accessibility and gene expression during embryogenesis and did a pilot study of early embryo development.

My work provides a comprehensive resource of chromatin accessibility and transcription patterns in the different tissues of *C. elegans.* It sheds light on fundamental differences between the mechanisms of transcription regulation of germline-active genes or somatic tissue-specific genes. The outcome of this work will greatly enable and push forward *C. elegans* transcription regulation research. The first datasets jointly profiling chromatin accessibility and nuclear transcription across the majority of tissues in a multicellular organism will also be of benefit for the broader community studying gene regulation in eukaryotes.

Acknowledgements

At some point during the summer 2019, I was discussing about the structure that an Acknowledgements section should have with friends and family. I would like to start by thanking them for giving me the opportunity to consider the importance and the meaningfulness of this part of my thesis. In a way, they opened my eyes on what these Acknowledgments and my thesis in general could represent, for others but also for myself.

I would like to express all my gratitude to my supervisor, Professor Julie Ahringer. I have always found her door open, and even though I would usually leave her office with a list of questions longer than that of results I would have brought to her, through our endless conversations, I believe she has contributed to make me a better scientist.

This work would certainly have been a lot duller without the people who surrounded me. The Italian team largely contributed to a joyful lab. Francesco, thank you for always be willing to give your input on my (sometimes silly) questions. Andrea, your help in ranking Parisian chouquette manufacturers has been invaluable. Arianna, you saved my fourth year by being around. Chiara, your loud enthusiasm and your genuineness have been refreshing throughout my entire PhD. Thank you for being like that. Navin, you were not in the lab, but you have dragged me out of it for beers many times, still not enough though. Our discussions outside of the Gurdon have been incredibly helpful to me, scientifically as well as personally.

I had the chance to interact with great supervisors along the way, and I am particularly grateful to Maite Huarte and Kathrin Plath. Maite has welcomed my in one of the most friendly lab environments I have ever known. Kathrin has taught me that being open to any challenge, dedication and hard work are always paying off.

Amy: I have learnt much more than RNA extraction while working with you, and you know that. Since I met you, you have always been here and I hope you will always be.

Life in Cambridge would not have been the same for me without the friends I met along the way. Some stuck around until the end. Joce, thanks for bringing some music in my life and for showing me that a PhD is not all about the science. Your multiple talents are truly inspirational. Tom, thanks for being regularly around at random hours of the night, I felt less lonely! I'm sure we'll eventually get somewhere with those crazy, exciting ideas emerging from our tired brains late at night – and sorry I have not baked a single cake for you. Captain Karen, the way you live your life is so impressive. I wish I were living mine with half the motivation and dedication you have. Stay the way you are! Other people decided to leave the house too early. Victor, I know you are still in Cambridge, but the house is not the same without you. Hannah, working from home is not as pleasant when you are not around. And you definitely deserved that Best Baker eggie award, I miss your cakes. And the garden misses you as well. Michael, you accompanied me through all the hurdles of the PhD literally from day 1. It was great to have you along my side throughout the journey. Alex, I didn't think I would every meet someone as kind and generous as you are. Stay like this, you are an example to all of us.

J'ai la chance immense d'avoir eu ma famille présente autour de moi, depuis bien avant, tout au long de cette aventure et au-delà. Grand-père, tes questions ont fait de moi quelqu'un de curieux; je n'ai toujours pas de réponses pour la plupart d'entre elles, mais j'y travaille. Grand-mère, mes collocs te trouvent incroyable à chaque fois que je parle de toi, et ils ont bien raison. Ta gentillesse et ta bienveillance sont source d'inspiration pour nous tous. Philippe, merci de m'avoir supporté par la pensée quand la distance m'empêchait de venir vous voir aussi souvent que je l'aurais souhaité.

Pierre et Paul, nos différences m'ont aidé à me définir et à mieux savoir qui je suis, et c'est largement grâce à vous si j'en suis là aujourd'hui. Merci d'avoir toujours été présents, malgré nos tendances migratoires. Avoir deux frères est une chance incroyable, et avoir deux frères comme vous l'est encore plus. Vous m'avez énormément apporté tout au long de ma vie.

Maud, tu m'as accompagné tous les jours depuis plus de quatre ans. Tu m'as appris plus que je n'aurais pu l'imaginer. Tu m'as façonné en quelqu'un de meilleur, chaque jour un peu plus, et je sais que tu continueras à le faire pendant longtemps.

Enfin, je n'aurais pas pu accomplir cette étape de ma vie sans la présence de mes deux parents. Ils m'ont tout donné, chacun à leur manière, et je leur en suis infiniment reconaissant. Si j'ai l'occasion d'écrire ce texte aujourd'hui, c'est non seulement grâce à l'éducation qu'ils m'ont offert, mais aussi grâce à la liberté qu'ils ont laissé à leurs enfants d'explorer leurs propres voies. Je suis immensément fier d'être l'un de leurs fils.

All my life, I have found comfort in science. It helps give meaning to the many things we cannot control. It brings a degree of order to the chaos that surrounds us. But whilst we may be able to explain the science behind an aureole, or the falling snow, it is not possible to account for its beauty.

— James Glaisher

Contents

List of Figures

List of Tables

Nomenclature

Roman Symbols

CAGE	Cap analysis of gene expression
CTD	Carboxy-terminal domain
FACS	Fluorescent-activated cell sorting
mRNA	messenger RNA
NDR	Nucleosome-depleted region
PIC	Pre-initiation complex
PSD	Power spectrum density
PTM	Post-translational modification
RE	Regulatory element
RNAPII	RNA polymerase II
TAD	Topological associating domain
TBP	TATA-binding protein
TFBS	Transcription factor binding site
TIC	Transcription initiation cluster
TRF	TBP-related factor
TSS	Transcription start site

Foreword

During my PhD, I have been lucky enough to get help from many people. I mentioned them in relevant sections of the results, but I would also like to specifically acknowledge them here. Michael Chesney originally designed a GFP reporter and generated two strains I used during my PhD. Chiara Cerrato helped me by injecting other GFP marker constructs I created. Yan Dong contributed to my PhD project by preparing Illumina RNA-seq libraries from the samples I generated. Alex Appert participated in the design and the optimization of alternative chromatin mark profiling methods, and Rhys McDonough generated the CUT&RUN datasets mentioned in this thesis together with me. Yan Dong, Michael Schoof and Jürgen Jänes were all investigating the chromatin dynamics during development and aging when I joined this project. Jürgen Jänes designed and executed pipelines to generate outputs required for some of my analyses.

Some parts of the introduction and the results have been published or are under review for submission in scientific journals. Notably, the section about 3D organization has been adapted from Serizay and Ahringer (2018) (see 1.2.2.2). The first chapter of the results focusing on coordinated regulatory elements during *C. elegans* development and aging, has been adapted from my work in Jänes *et al.* (2018). Some parts of the second, third and fourth chapters of the results, presenting the different regulatory architectures in adult *C. elegans*, are adapted from a manuscript now available on biorXiv preprint server (Serizay *et al.*, 2020).

Chapter 1

Introduction

To place my study in the general context of genome organization and transcription regulation, the first section of this introduction provides an overview of the current understanding of the nucleosomal architecture. The second section further presents the notion of regulatory elements as well as the different mechanisms regulating chromatin transcriptional activity and introduces the aspects of gene regulation in development. The third section presents *Caenorhabditis elegans* as a model system to investigate gene regulation in development. The fourth and last section of this introduction establishes the main aims of my PhD.

1.1 Nucleosomes are the fundamental unit of chromatin

The molecules of DNA contained in eukaryotic cells can reach up to several meters when stretched out. Yet, the nuclei where they are stored are one to fifty million times smaller than this. To reach this state of compaction, DNA is packaged with a complex set of proteins into a substance called chromatin. In Chapter 1.1, I present the role of nucleosomes as the fundamental structural unit of chromatin.

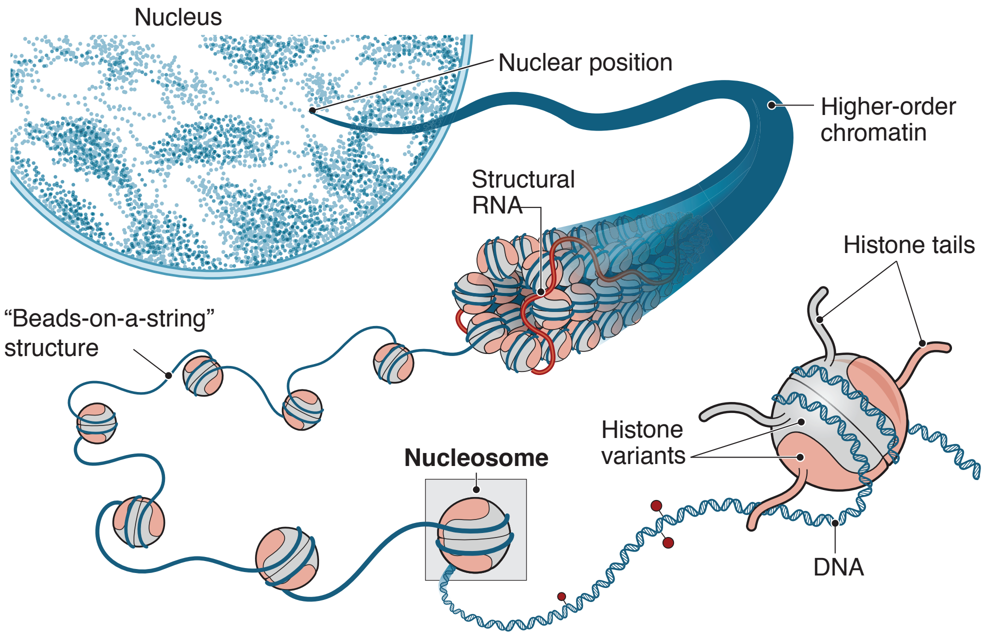


FIGURE 1.1 – Hierarchical organization of the molecule of DNA in an eukaryotic nucleus (adapted from Yadav *et al.*, 2018).

1.1.1 Nucleosomes are the basic structural unit of the genome

1.1.1.1 DNA is wrapped around globular protein complexes

The nucleosome is the basic structural unit of the chromatin. It facilitates the organization of DNA double-helix into chromatin, a highly compact nucleoprotein entity (Figure 1.1 and Cutter and Hayes, 2015).

The core of a nucleosome is made of a protein complex comprising two copies of four histone proteins: H2A, H2B, H3 and H4 (Figure 1.2A and Cutter and Hayes, 2015). A147-bp-long fragment of DNA can wrap around this relatively small octamer. Each nucleosome resembles a globule with a diameter of approximately 10 nm and is separated from the next nucleosome by a 10- to 80-bp-long DNA linker (Figure 1.1 and Figure 1.2A-B). With DNA wrapped around histones, the chromatin fiber is frequently compared to a "beads-on-a-string" structure (Figure 1.1).

In comparison, the dimensions of a 150 bp-long DNA linear segment would be roughly 2 nm x 50 nm. Thus, the nucleosome is an efficient way to package DNA in nuclei. In physiological conditions, when the additional histone H1 is present and binds to DNA linkers, specific spatial organization of the nucleosomes can form higher-order chromatin fibers, allowing for greater compaction of the DNA in

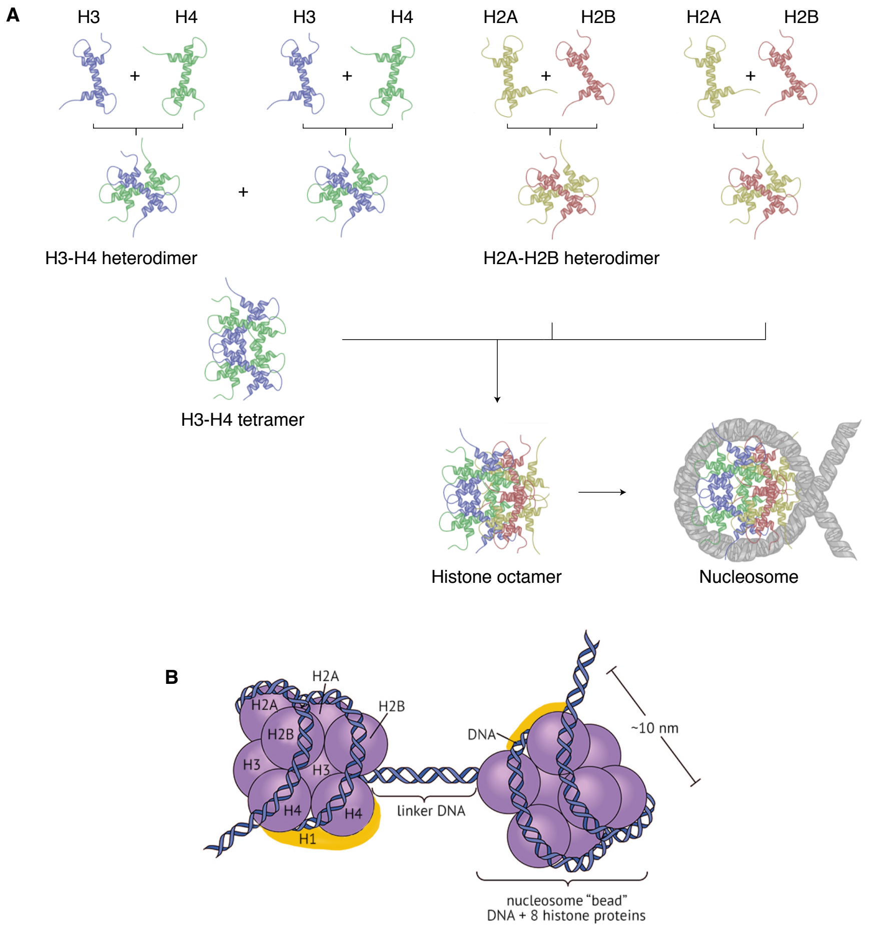


Figure 1.2 – Molecular constituents of a nucleosome. **A-** A nucleosome is constituted of a protein core (a combination of eight histone proteins) with DNA wrapped around it. **B-** Drawing of a dinucleosome (two adjacent nucleosomes separated by a DNA linker) (adapted from Cutter and Hayes, 2015).

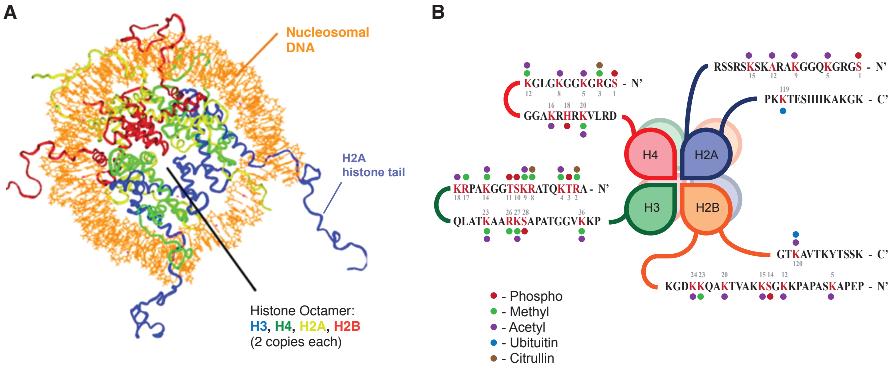


Figure 1.3 – Nucleosome molecular structure and post-translation modifications. **A-** Detailed molecular structure of a nucleosome. Note the histone tails sticking out of the nucleosome core globule (Protein Data Bank 1KX5). **B-** Histone post-translational modifications found in eukaryotes (reviewed in Zhao and Garcia, 2015).

nuclei (Figure 1.1).

1.1.1.2 Nucleosomes confer a high plasticity to chromatin

Chromatin is the resulting complex of DNA and histones interacting with additional proteins and RNA. It is the main constituent of the nucleus in eukaryote cells. The structure of the chromatin is highly dependent on the nucleosome organization and the extent to which DNA is compacted can vary accordingly (see below). This has a predominant role in the nuclear biology. For instance, the overall structure adopted by chromatin during interphase can in some cases be relatively loose to allow gene expression. However, during cell division, conformation of the chromatin undergoes drastic remodeling and adopt a much denser organization into chromosomes, ensuring protection and appropriate segregation of the heritable genetic material. This illustrates one of the most important aspects of the chromatin: its plasticity is a crucial feature coming into play in many developmental processes, from cell differentiation to stress response. In the next sections, I present the different features of nucleosomes important for chromatin regulation.

1.1.2 Histone post-translational modifications

1.1.2.1 Histone can be post-translationally modified

Histones are the key protein components of nucleosomes. They are characterized by their prominent N-terminal tails exposed on the surface of the assembled

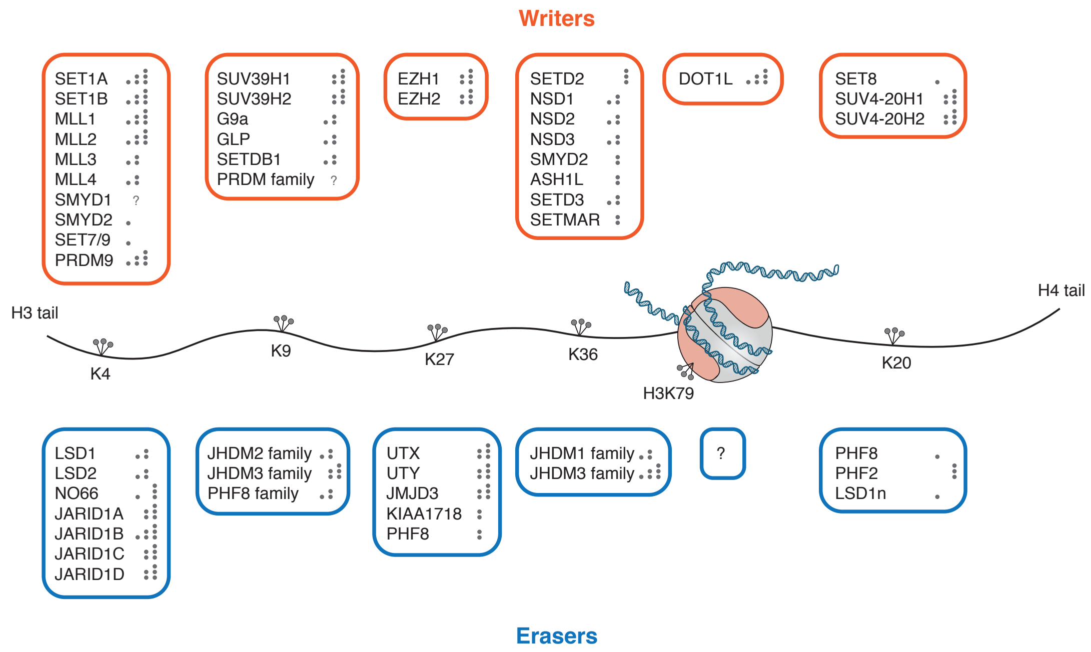


Figure 1.4 – Known human histone lysine methylation writers (methyltransferases) and readers (demethylases). The activity of the writers/erasers is depicted next to their name. Note that histone writers and readers can be part of larger protein complexes which combine both activities, such as the Polycomb Repressive Complex 2 (adapted from Hyun *et al.*, 2017).

nucleosome (Figure 1.3A). These tails feature evolutionary conserved amino-acids subject to dozens of covalent post-translational modifications (PTM) (Figure 1.3B and Ho *et al.*, 2014). For instance, Lysine 4 of the Histone 3 (His3) protein can be methylated one, two or three times; such histone modifications are referred to as H3K4me1, H3K4me2 or H3K4me3. A variety of histone PTMs exist (*e.g.* acetylation, methylation, ubiquitinylation or phosphorylation, among others) and affect N-terminal tails of all histones as well as the C-terminal tail of H2A and H2B (Kouzarides, 2007).

The proteins responsible for histone PTMs form a large family of histone modifying complexes known as “writers” which is still growing as of today. In many cases, “erasers” (*i.e.* factors with antagonist activity form “writers”) also exist, allowing for the covalent histone PTMs to be reversible (Figure 1.4). For example, histone acetyl transferases and histone deacetylases have opposite enzymatic activity on lysine residues of histones (Kouzarides, 2007).

Most histone modifying complexes have orthologues across metazoans. The Polycomb Repressive Complex 2 (PRC2) is a well-studied chromatin remodeler constituted of at least three main subunits (EZH2, EED and SUZ12) in human and responsible for H3K27 trimethylation. All of these subunits are conserved in *Drosophila melanogaster* and two of them are found in *Caenorhabditis elegans*: MES-2 (ortholog of EZH2) and MES-6 (ortholog of EED).

1.1.2.2 Histone modifications and gene regulation

Many histone PTMs modify the electrostatic constrains of the nucleosome (Tessarz and Kouzarides, 2014). For instance, the unacetylated Lysine 56 from His3 (H3K56) is positively charged and enhances the interaction between the DNA backbone (negatively charged) and the histone. Acetylation of this lysine removes its positive charge, decreasing the affinity between the histone and the DNA backbone. This has been described as "nucleosome breathing" and results in a relaxed wrapping of DNA around nucleosomes with H3K56ac modifications. Other histone tail modifications, such as the addition of methyl groups on lysine residues, can create steric hindrance either between the nucleosome surface and the DNA double-strand or between subunits of the nucleosome (Bowman and Poirier, 2015). Finally, PTMs at the histone-histone interface (such as H4K91ac) can disrupt nucleosome intrinsic stability by altering the interactions between each histone protein.

Overall, histone modifications largely impact the stability of nucleosomes as well as their interactions with DNA, and thus play a central role in regulating chromatin organization and gene expression (Bannister and Kouzarides, 2011).

On top of modulating nucleosome stability and interactions with DNA, some histone modifications can also participate to the recruitment of transcription factors. H3K4me3 is a canonical histone modification found at the first nucleosome downstream of transcription start sites. It has been shown to be read by the PHD domain of TAF3, a subunit of TFIID, a basal transcription factor required for the activation of large sets of genes in metazoans (Vermeulen *et al.*, 2007).

In other cases, histone modifications can lead to the recruitment of inactivating proteins. H3K9me2/3 modifications are bound by the chromodomain of Heterochromatin Protein 1 (HP1) (Fischle *et al.*, 2005) which, when recruited to a given locus, will lead to inactivation of transcription (Danzer and Wallrath, 2004). Importantly, histone modifications can act cooperatively to enhance binding of a given factor. For instance, PHF8 binding to H3K4me3 is enhanced by the acetylation of H3K9 and H3K14 (Bannister and Kouzarides, 2011).

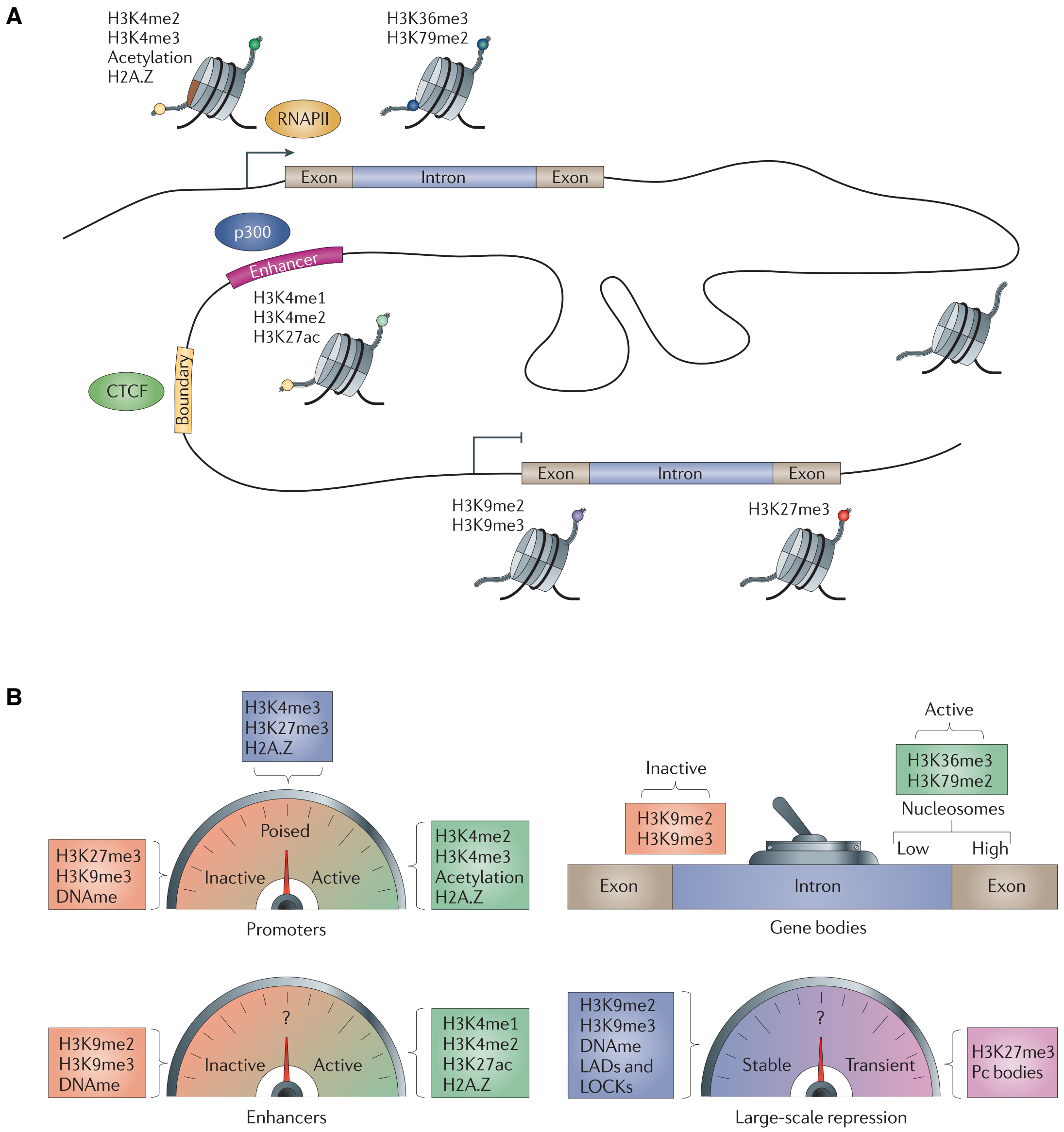


FIGURE 1.5 – **A-** Different combinations of histone modifications demarcate functional elements in mammalian genomes. **B-** Association of histone modifications with either an active or an inactive chromatin conformation for different functional elements (adapted from Zhou *et al.*, 2011).

1.1.2.3 Histone code and chromatin states

Different histone PTMs are associated with certain chromatin loci (*e.g.* promoter, enhancer, transcription, repeats, inactive chromatin, ...) (Ernst and Kellis, 2010; Kundaje *et al.*, 2015 and Figure 1.5A). H3K4me2 and H3K4me3 are usually found in cis of distal gene regulatory loci. H3K36me2 and H3K36me3 are found enriched across a transcribed gene, in combination with H3K79me1/2/3 at its 5' end. H3K9me1/2/3 and H3k27me2/3 are generally found in inactive chromatin, either in large domains or over narrow loci. Importantly, the reversible nature of histone PTMs reflects how a given genomic locus can switch between an active and an

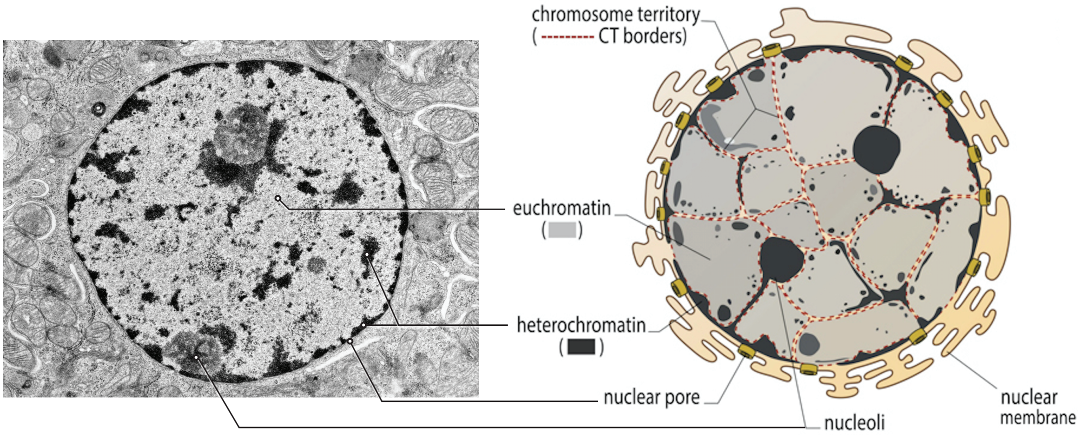


Figure 1.6 – Organization of euchromatin and heterochromatin in a mammalian nucleus. Left inset: electron microscopy photograph of a nucleus (adapted from medcell.med.yale.edu). Right inset: Schematic of the organization of chromatin within a nucleus, with chromosome territories and the different types of chromatin depcited (adapted from https://www.mechanobio.info/).

inactive conformation (Figure 1.5B). This notion of a histone code defining specific chromatin environment, also known as “chromatin states”, has been instrumental in the understanding of the role of the chromatin and its dynamic changes for gene regulation, cell physiology and development.

The notion of histone PTMs cross-talk is crucial to understand how combinations of histone modifications are integrated together. Histone PTM cross-talks occur at multiple levels. First, competitive antagonism between histone modifications occurs, particularly at the lysines of histone tails which are subject to a range of possible modifications (Kouzarides, 2007). Secondly, histone modifications can depend on one another. For example, methylation of H3K4 and H3K79 can only happen when H2BK123 is ubiquitylated in yeast (Lee *et al.*, 2007). Alternatively, binding of HP1 to H3K9me3 is disrupted by the phosphorylation of H3S10 during mitosis (Fischle *et al.*, 2005). Finally, the order of implementation of histone modifications can affect the impact of these modifications on gene transcription (Lee *et al.*, 2010).

1.1.3 Nucleosomes and chromatin accessibility

1.1.3.1 Accessibility in euchromatin and heterochromatin relies on nucleosome structure and organization

Chromatin is segregated in the nucleus into euchromatin and heterochromatin (Figure 1.6 and Chen and Dent, 2014; Evans *et al.*, 2016). Euchromatin is characterized by a loose nucleosome arrangement, active histone modifications (*e.g.* H3K36me3 and H3K79me2) and the association with transcription factors (Figure 1.7). Euchromatin is in a decondensed state which fills up most of the nucleoplasm (Figure 1.6). In contrast, heterochromatin is characterized by H3K9me3 and/or H3K27me3 modifications, the association with heterochromatin proteins such as HP1 and a high level of histone compaction (Figure 1.7). Heterochromatin is present in the nucleus (i) interacting with the nuclear lamina at the nuclear envelope, (ii) condensed into granules within the nucleus or (iii) around the nucleoli (Figure 1.6). Heterochromatin can be separated into two states. "Constitutive heterochromatin" is found at specific regions of the genome which are permanently condensed into an inactive state across all physiological and most pathological contexts, and is usually marked by H3K9me2 and H3K9me3. On the contrary, "facultative" heterochromatin is a non-permanent inactive state of the chromatin which can be altered depending on the cell context, and is usually marked by H3K27me3 (Figure 1.7).

More generally, depending on their structure and organization, nucleosomes confer different levels of accessibility to the chromatin. Inactive chromatin-related histone modifications generate compact arrays of nucleosomes; this creates an environment where only heterochromatin proteins can interact with DNA. In contrast, active chromatin-related histone modifications relax nucleosomes, allowing for transcription activators to bind to DNA.

1.1.3.2 Profiling of chromatin accessibility landscapes

Nucleosomes define the first level of chromatin accessibility. Thus, profiling nucleosome presence (or absence) across a given locus is a way to profile chromatin accessibility. Historically, this has been done by using nucleases. By adding nucleases to nuclei containing DNA in its native conformation, cuts are preferentially generated where the DNA is accessible and unprotected, *i.e.* where the DNA is not in contact with nucleosomes. Relying on this idea, deoxyribonuclease I (DNase I) and Micrococcal nuclease (MNase) have been used in analogous ways to profile chromatin accessibility (Zentner and Henikoff, 2012). DNase preferentially cuts in "hyper-sensitive" accessible regions of the DNA; upon sequencing, regions characterized by a high accessibility are revealed by DNase-seq signal. MNase, on the contrary, cuts DNA between nucleosomes; upon paired-end sequencing,

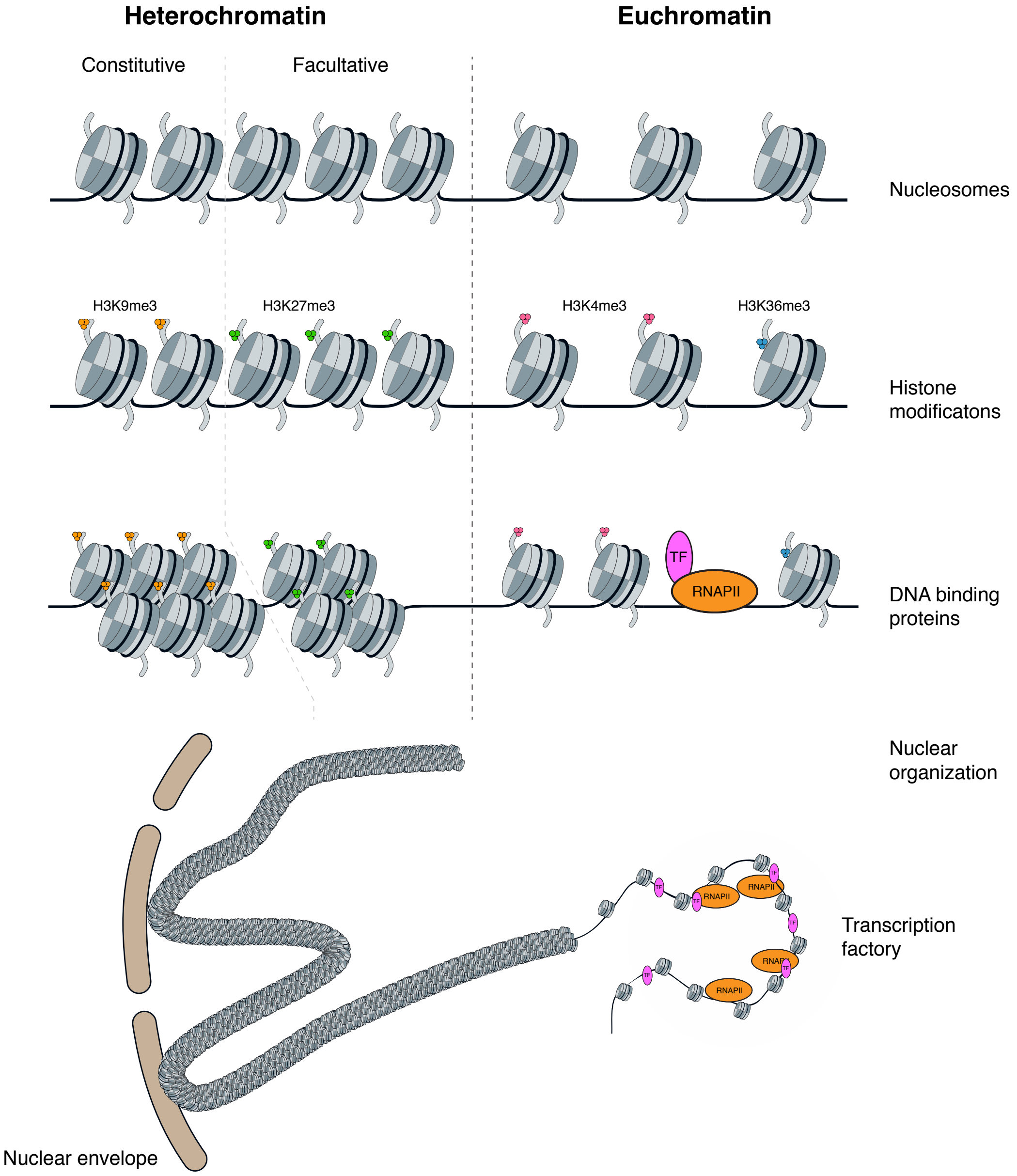


FIGURE 1.7 – Features of euchromatin and heterochromatin. Heterochromatin is associated with genomic features such as H2K9me3/K27me3 histone modifications, dense nucleosomes and attachment to nuclear envelope, while euchromatin is associated with loose nucleosome structure, higher chromatin accessibility and enriched interactions with transcription factories (adapted from Zhou *et al.*, 2011).

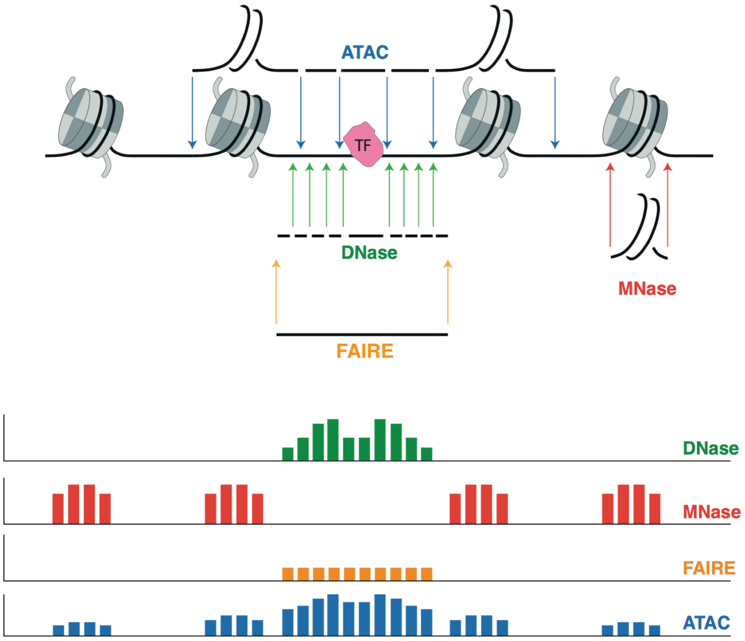


Figure 1.8 – Different genomic assays can be used to profile chromatin accessibility. ATAC-seq relies on the Tn5 transposase to capture accessible chromatin; DNase-seq relies on DNAse to capture DNase I hypersensitive sites; MNase-seq relies on MNase to capture nucleosomal DNA; FAIRE-seq relies on mechanical shearing to capture nucleosome-depleted under-crosslinked DNA.

regions occupied by a nucleosome will be revealed and accessible regions can be inferred by the absence of MNase-seq signal (Figure 1.8). Each method has its pros and cons: DNase-seq allows one to model the footprint of a protein on the DNA double-strand to study its mechanism of binding but does not give any information of the neighboring nucleosome organization. In contrast, MNase-seq reveals the arrangement (occupancy and positioning) of nucleosomes at accessible regions and throughout the genome but will not shed light on how other proteins may be interacting with DNA.

Of note, other profiling methods also provide information on nucleosome arrangement. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) allows one to profile the binding pattern of a protein (which could be a histone subunit of nucleosomes, for instance) (Park, 2009). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) is another method relying on cross-linking of nucleosomal DNA and enrichment of non-crosslinked internucleosomal DNA (Giresi *et al.*, 2007). However, these methods do not have the same sensitivity as nuclease-based assays, which are preferred when focusing on profiling chromatin accessibility.

More recently, ATAC-seq (Assay for Transposase-Accessible Chromatin sequencing) was designed to profile chromatin accessibility (Buenrostro *et al.*, 2013). It relies on a transposase (Tn5) to insert short DNA oligomers (transposomes) in accessible regions; after isolation of the "tagmented" DNA and library preparation, paired-end sequencing can reveal accessible loci where the Tn5 enzyme was able to integrate its transposome (Figure 1.8). Importantly, ATAC-seq not only reveals hyper-sensitive sites, similarly to DNase-seq, but it also gives additional information on the nucleosome organization flanking the accessible site. Computational tools such as nucleoATAC (Schep *et al.*, 2015) have now been developed to use ATAC-seq datasets to infer nucleosome occupancy and positioning, alleviating the hurdle of the MNase-seq poor reproducibility.

ATAC-seq is a very efficient method to map open chromatin loci and study nucleosome organization. Its early experimental procedure has been further optimized and in adequate conditions, as few as 500 nuclei can yield a reasonably good chromatin accessibility profile (Corces *et al.*, 2017). However, single-cell techniques are now emerging and would theoretically allow one to profile the chromatin accessibility landscape of thousands of nuclei individually. This represents a formidable opportunity to start focusing on new questions. Notably, the field of developmental biology will certainly benefit from these techniques as single-cell approaches represent the way to study the differentiation events occurring early on during embryonic development. Such topics are already being tackled (*e.g.* by Cusanovich *et al.*, 2018a) to try and shed light on the early molecular events leading to the formation of a multi-cellular organism.

1.1.4 Interplay between nucleosomes and transcription factors

Transcription factors (TFs) are proteins interacting at regulatory loci of the genome to control gene expression (Spitz and Furlong, 2012). Many of them directly interact with DNA by recognizing specific binding motifs thanks to a DNA-binding domain, though some can indirectly interact with DNA when engaged in multi-protein complexes. A canonical example is the vitamin D receptor (VDR) transcription factor. In unstimulated cells, monomeric VDR is unable to form stable protein-DNA interactions at its binding motif. When its ligand $_{1,25}(OH)_2D_3$ is present in cells, VDR recognizes it and changes conformation, triggering its heterodimerization

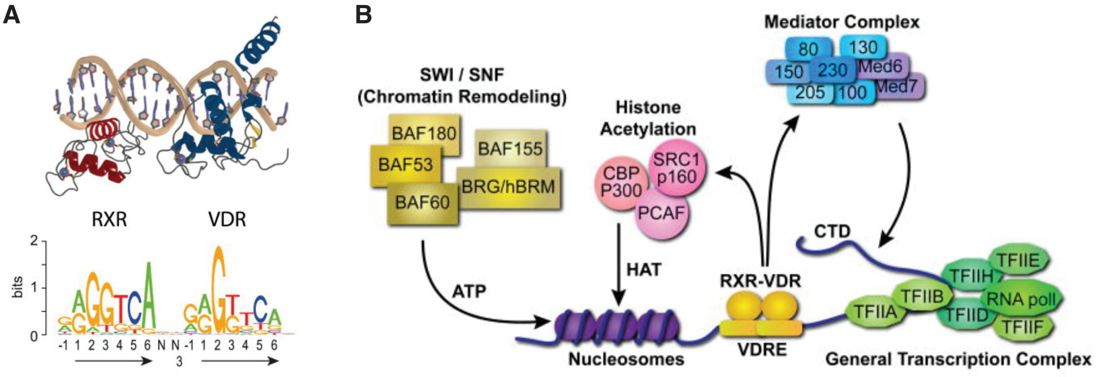


FIGURE 1.9 – Vitamin D: a canonical example of transcription factor-mediated gene regulation. **A-** Structure of the heterodimer VDR-RXR bound to DNA (top) and representation of its DNA binding site (bottom) (adapted from Carlberg and Campbell, 2013). **B-** VDR-RXR-mediated recruitment of a battery of proteins leading to modification of the local chromatin environment: chromatin remodeling complexes (yellow), histone acetyl-transferases (pink), general transcription factors (in green, see 1.2.1.1) and mediator (in blue, see 1.2.1.2) (adapted from Pike and Meyer, 2010).

with the retinoid X receptor (RXR). The dimer can then physically interact with a specific DNA motif found in vitamin D response elements (VDREs) (Figure 1.9-A and Carlberg and Campbell, 2013).

TF binding sites (TFBS) may be "buried" within nucleosomal DNA rather than in between nucleosomes (Neph *et al.*, 2012). In this context, most transcription factors can not recognize their binding site. Nucleosome remodelers are conserved protein complexes that are usually required to first displace or remove nucleosomes, an ATP-dependent active process (Clapier and Cairns, 2009). As a consequence, the local organization of nucleosomes is altered and initially hidden TFBS can be unmasked, allowing transcription factors to bind (Figure 1.10 and Jiang and Pugh, 2009; Spitz and Furlong, 2012).

Reversely, the binding of transcription factors to their DNA binding site (*e.g.* the VDR-RXR heterodimer to VDREs) also leads to important remodeling of the local chromatin environment (Figure 1.9B and Pike and Meyer, 2010). Transcription factors can interact with chromatin remodelers or transcription machinery to modify local transcriptional activity, and in some cases largely remodel nucleosome organization.

Importantly, "pioneer" transcription factors (*e.g.* PHA-4 in *C. elegans*) are a specific class of transcription factors that can bind nucleosomal DNA (Hsu *et al.*, 2015; Zaret and Carroll, 2011). Their binding can initiate cooperative interactions

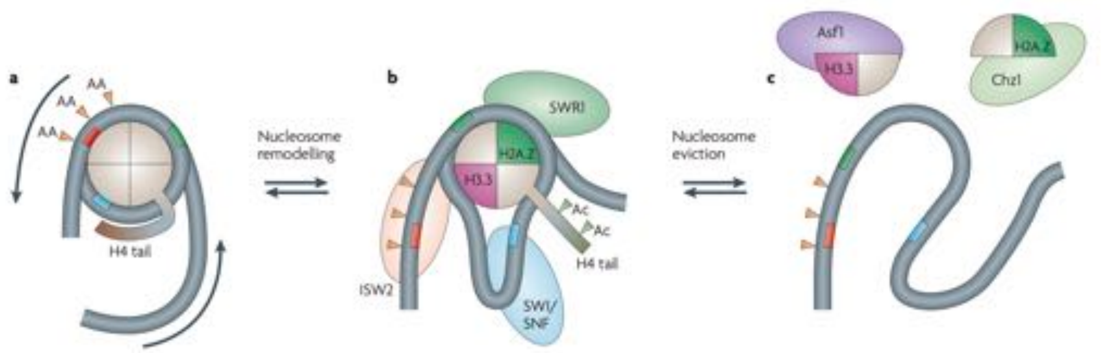


FIGURE 1.10 – Mechanisms altering DNA accessibility by nucleosome remodeling. **A-** A stable nucleosome. **B-** A remodeled nucleosome. **C-** An evicted nucleosome. Three transcription factor binding sites are shown in red, green and blue, respectively. The red and blue sites only become accessible after remodeling (**B**), either by nucleosome sliding or by chromatin remodeling complexes that 'extract' DNA from the nucleosome surface (*e.g.* ISW2, SWR1 and SWI/SNF). The green site is always accessible in the various states as it is exposed at the outer side of the nucleosome. Nucleosome eviction (**C**) might be necessary to assemble large protein complexes and to transcribe the underlying DNA (adapted from Jiang and Pugh, 2009).

with other regulatory proteins, eventually leading to loosening and opening of the local chromatin. Upon pioneer factor binding, chromatin can become competent for activation or even become transcriptionally active. Thus, pioneer factors are the first to come into play during cell differentiation, a process inherently requiring large genome-wide chromatin remodeling (Iwafuchi-Doi and Zaret, 2014).

1.2 Principles of gene regulation in metazoans

Nucleosomes are fundamental units of chromatin and are required for packaging DNA into nuclei. However, these "beads on a string" also represent a hurdle obstructing the complex machinery in charge of gene transcription. In Chapter 1.2, I present the structure of the different classes of regulatory elements (REs) and introduce the importance of gene regulation during development.

1.2.1 Regulatory elements modulate transcription

1.2.1.1 Promoters regulate the initiation of productive transcription

Fundamental promoter organization Tens of thousands of promoters exist in metazoan genomes and are located at the 5' end (or "upstream") of coding sequences. When favorable conditions are met, a complex set of proteins are recruited

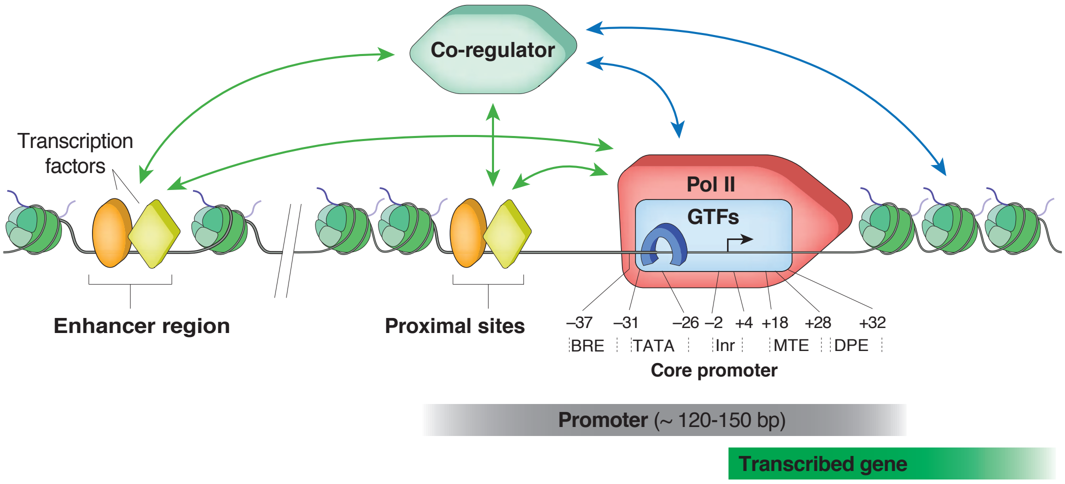


Figure 1.11 – Schematic of the different protein complexes interacting at the extended promoter (core promoter and proximal TF binding motifs) and at distal sites. Red box: RNA polymerase II; blue box: general transcription factors (*e.g.* TFIID); yellow diamond and orange oval: specific transcription factors; green hexagon: co-regulators (*e.g.* p300/CBP) (adapted from Fuda *et al.*, 2009).

to a promoter locus (Figure 1.11) to initiate transcription of the downstream region. RNA polymerase II (RNAPII) is the enzymatic complex generally responsible for the synthesis of messenger RNA (mRNA) from protein-coding genes. mRNAs synthesized during the transcription undergo co- and post-transcriptional modifications (such as capping, splicing and poly-A-tailing, depending on its nature) and are exported out of the nucleus to be translated.

The very first base of a promoter to be transcribed is called a "Transcription Start Site" (TSS). In a given promoter, the TSS is not necessarily located at a reproducible position for every initiation event (Figure 1.12). Instead, TSSs are organized in clusters (*i.e.* Transcription Initiation Clusters, or TICs, also referred to as Transcription Clusters or TCs, Carninci *et al.*, 2006). The base with the highest transcription initiation signal is usually used to annotate the dominant TSS (Figure 1.12). TSS profiling is the most specific approach to annotate promoters and has led to the identification of different groups of promoters, based on their transcription initiation characteristics (reviewed in Lenhard *et al.*, 2012 and detailed in 1.2.1.1).

Still, precisely annotating TSSs may be challenging in specific contexts. In *C. elegans,* trans-splicing affects ~ 70% of the transcripts (Allen *et al.*, 2011); the 5' end of these transcripts (the "outron") can be spliced out and replaced by a splice leader sequence. Importantly, the trans-spliced outron can be several kb

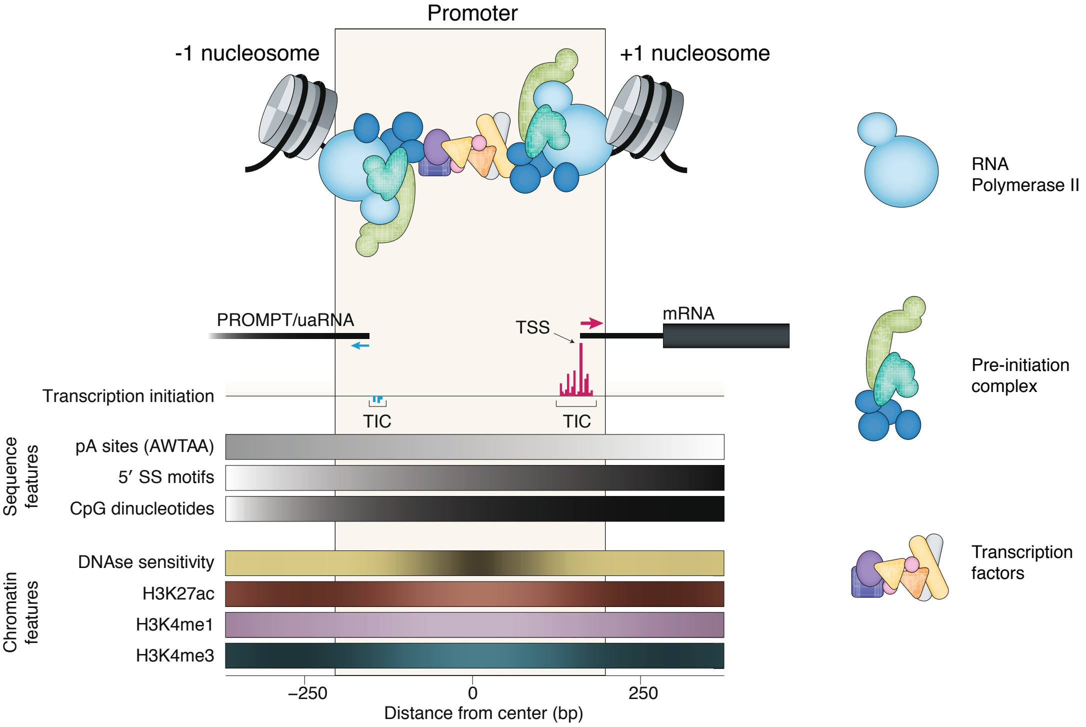


Figure 1.12 – A schematic of the promoter and its flanking nucleosomes, together with transcription factors (TFs), the pre-initiation complex (PIC) and the RNA polymerase II (RNAPII). Below the schematic, the first row shows the synthesized RNAs initiating at the TICs (both the mRNA and the upstream antisense RNAs are represented) and the second row shows typical signals of transcription initiation obtained by cap analysis of gene expression (CAGE). The sequence features typically found in promoters (polyA (pA) sites, 5´ splice sites (5´ SSs) and CpG dinucleotides) are visualized as greyscale bars showing their average pattern density, where black indicates higher density. The representative DNase sensitivity is shown as well as the typical patterns of H3K27ac, H3K4me1 and H3K4me3 histone modifications (adapted from Andersson and Sandelin, 2019).

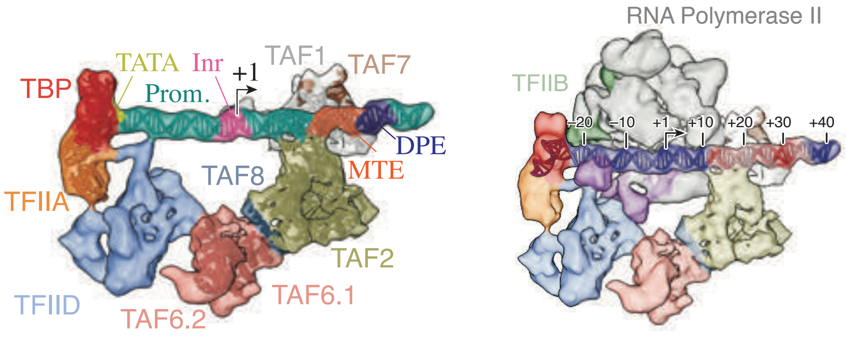


Figure 1.13 – Structure of the Pre-Initiation Complex (PIC) without RNA Pol II and TFIIB (left inset) and with RNAPII and TFIIB (right inset). TBP: TATA-Binding Protein; TAF: TBP-Associated Factors; Inr.: Initiator (adapted from Louder *et al.*, 2016).

long, effectively preventing the precise mapping of promoters in this organism. Alternative methods have been used to capture the position of the initial 5' end have been successfully applied to map TSSs and promoters in the nematode (Chen *et al.*, 2013; Jänes *et al.*, 2018; Kruesi *et al.*, 2013; Saito *et al.*, 2013).

Promoters harbor specific sequences and chromatin features which can be used to infer the position of promoters in genomic studies. For instance, splicing sites and CpG dinucleotides as well as H3K4me3 and H3K27ac histone modifications are enriched around or immediately downstream of promoters (Figure 1.12).

Formation of Pre-Initiation Complex at promoters for productive transcription The Pre-Initiation Complex (PIC) is assembled at core promoters (Figure 1.11 and Figure 1.12). It is required to activate RNAPII and its position determines where transcription initiates. The PIC is a large protein complex that orchestrates transcription initiation (Figure 1.13). The main components of this complex are the general transcription factors (GTF): TFIIA, -B, -D, -E, -F and -H and RNAPII. Each member of the TFII family is composed of many subunits, resulting in a massive complex of more than fifty proteins bound in the promoter region (Kornberg, 2007). On top of this minimal PIC assembly, co-activators are generally required for transcription of messenger RNAs (Figure 1.11).

Initiation of transcription is a multi-step process starting with the assembly of the PIC at an open promoter (Figure 1.13 and Figure 1.14). First, TFIID interacts at an open promoter. This interaction is stabilized by TFIIA and TFIIB brings DNA into a configuration which can enter the active site of the RNA Polymerase

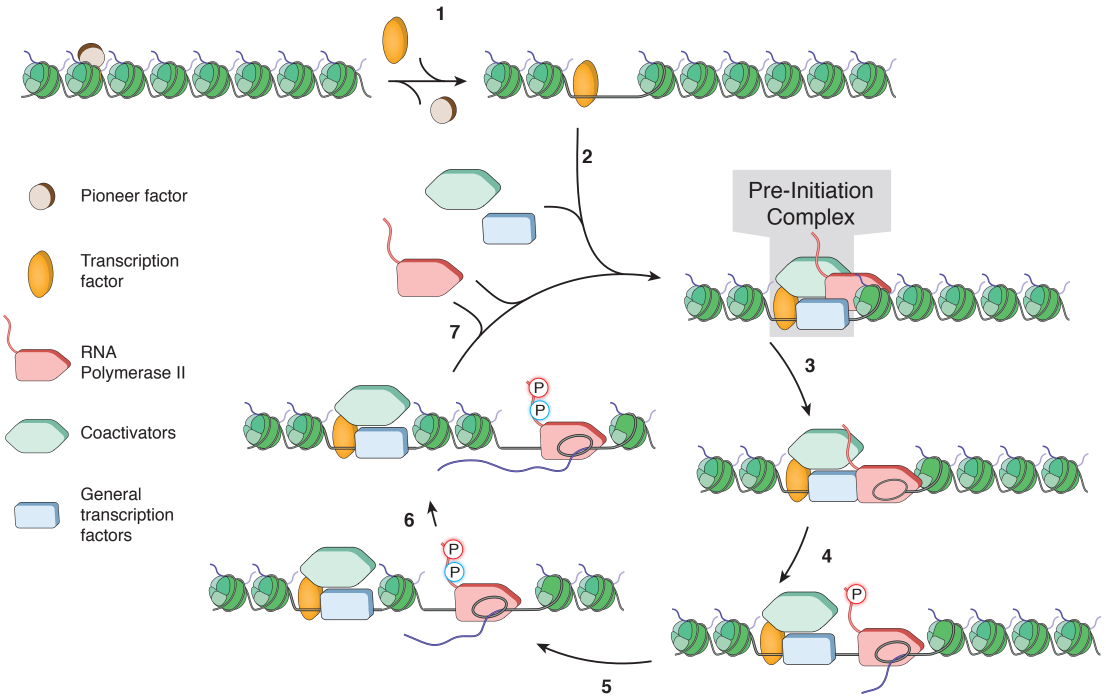


FIGURE 1.14 – Molecular steps leading to productive transcription of an mRNA. 1) Upon opening of the chromatin by a pioneer factor, a TF binds to the promoter. 2) This effectively initiates the assembly of the pre-initiation complex at the open promoter. 3) DNA is unwound to form an open complex, 4) RNAPII engages transcription and rapidly pauses. 5) A second phosphorylation of RNAPII C-terminal domain leads to its escape. 6) Eventually, the RNAPII encounters a termination signal and 7) it is recycled in a new PIC (adapted from Fuda *et al.*, 2009). The modalities of RNA transcription initiation vary for different types of promoters and across organisms.

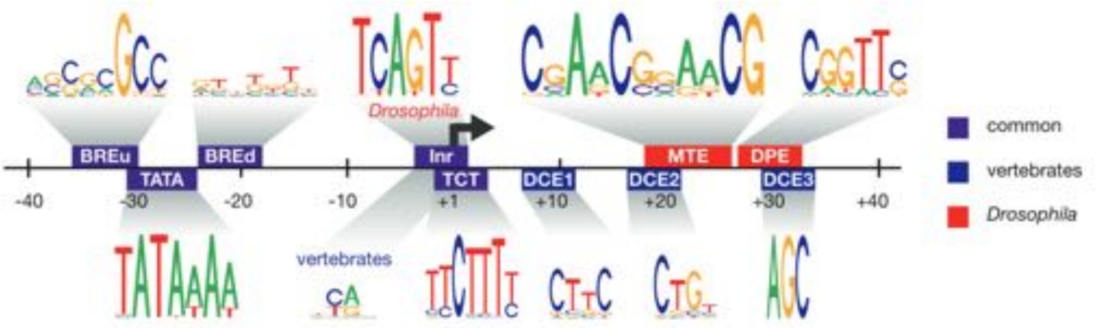


Figure 1.15 – Sequences enriched in core promoters and their conserved location (adapted from Haberle and Lenhard, 2016). Note that this is an aggregated consensus promoter and that no promoter would present all these sequences at once.

II. RNAPII and TFIIF are recruited, followed by TFIIE ad TFIIH. The initial recruitment of TFIID to the promoter is variable depending on its organization and can require either a TATA-Binding Protein (TBP) subunit or TBP-related factors (TRFs). Once the PIC is assembled, transcription is initiated at the TSS after unwinding the DNA, forming an open complex. RNAPII then breaks contacts with promoter-bound factors, transcribes 20–50 bases downstream of the TSS and pauses. At this stage, RPB1, the main subunit of RNAPII, has its carboxy-terminal domain (CTD) phosphorylated. A second phosphorylation of the CTD leads to RNAPII escape from pausing to enter in a productive elongation state, ending with transcription termination. At this point, RNAPII can be reused in a new transcription initiation event (Fuda *et al.*, 2009).

Conserved sequences in promoters are important for regulation Core promoters contain canonical sequence features found across all eukaryotes or in specific clades (Figure 1.15 and Lenhard *et al.*, 2012). DNA motifs are bound by different factors and contribute to transcription initiation. However, canonical sequences are not found all in every core promoter. For instance, the TATA-box is the motif recognized by TBP at TATA promoters (Buratowski *et al.*, 1989) but in its absence, the Sp1 protein can act as a molecular bridge between the DNA and TFIID to enable proper PIC assembly (Kadonaga *et al.*, 1986; Tan and Khachigian, 2009). Moreover, specific promoter sequences have been identified in individual species (*e.g.* MTE and DPE in *Drosophila*, Figure 1.15).

Extended promoters (~ 120-150 bp around the TSS) can also harbor additional transcription factors binding sites (TFBS). TFBS are sometimes unmasked upon chromatin opening by pioneer factors and can then be bound by specific transcription factors. Once bound, specific transcription factors can modulate the activity of the promoter. Thus, the presence or absence of TFBS is a good indicator of where, when and how sets of promoters are activated. For instance, ELT-2 transcription factor is the master regulator of gut differentiation in *C. elegans* and more than 80% of the genes active in the gut feature its binding motif in their promoter (McGhee *et al.*, 2009). Thus, identifying conserved motifs enriched in sets of promoter sequences aids understanding the mechanisms of regulation of these promoters (Kulakovskiy and Makeev, 2013).

Nucleosome organization at promoters Accessibility of DNA at promoters is crucial for protein factors to bind and regulate transcription (see 1.1.4), and nucleosomal organization at promoters is largely conserved from yeast to metazoans (Mavrich *et al.*, 2008b; Schones *et al.*, 2008; Valouev *et al.*, 2008; Yuan *et al.*, 2005, Figure 1.12). Active promoters are generally depleted of nucleosomes, generating a so-called Nucleosome-Depleted Region (NDR). Flanking upstream and downstream nucleosomes are identified as the -1 and +1 nucleosomes. These nucleosomes usually have a high occupancy and a tight positioning, *i.e.* they are generally localized at the same loci upstream and downstream of the NDR. In yeast, the nucleosomes downstream of the +1 nucleosomes are also well arranged, creating an array of nucleosomes (Figure 1.16 and Mavrich *et al.*, 2008a). Such arrays are also detected in mammals at ubiquitous or strongly expressed genes, albeit to a lesser extent than in yeast (Lenhard *et al.*, 2012).

In the absence of any other factor, the position of a nucleosome on DNA is influenced by the underlying DNA sequence (Albert *et al.*, 2007; Dreos *et al.*, 2016; Field *et al.*, 2008; Forrest *et al.*, 2014; Haberle *et al.*, 2014; Ioshikhes *et al.*, 1996, 2011; Pich *et al.*, 2018; Satchwell *et al.*, 1986; Segal *et al.*, 2006; Struhl and Segal, 2013; Wang and Widom, 2005). To which extent this is relevant for *in vivo* nucleosome positioning along the genome and particularly in the regions flanking NDRs has been under a lot of investigation and still remains debated (reviewed in Struhl and Segal, 2013; Travers *et al.*, 2010). An emerging model is that specific underlying DNA sequence features (such as 10-bp periodic dinucleotides or stretches of $(A/T)_{n,\ n>6}$, also referred to as poly(dA:dT) tracks) confer specific physical properties to the DNA double-helix, such as bendability or stiffness, which can favor

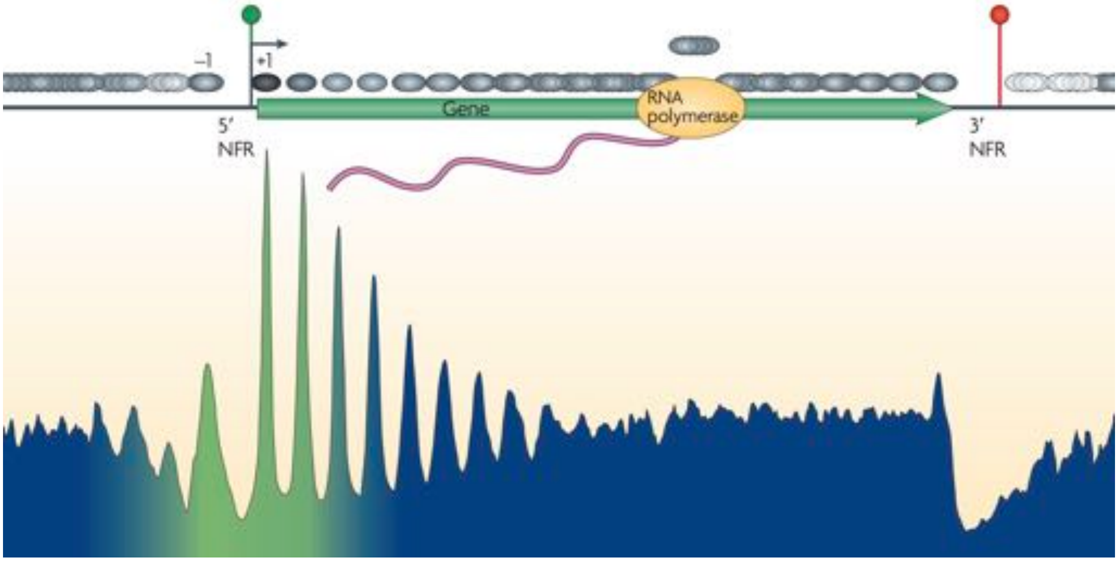


FIGURE 1.16 – Nucleosomal landscape of yeast genes. The peaks and valleys represent similar positioning relative to the transcription start site (TSS). Nucleosomes are shown as grey ovals and the green–blue shading represents the transitions observed in nucleosome composition and phasing (green represents high H2A.Z levels, acetylation, H3K4 methylation and phasing, whereas blue represents low levels of these modifications). Note the strong phasing of the +1 nucleosomes (adapted from Jiang and Pugh, 2009).

or disfavor nucleosome positioning (Figure 1.17 and Struhl and Segal, 2013; Travers *et al.*, 2010). Such sequence features at and around promoters likely contribute to the generation of nucleosome-depleted regions flanked by -1 and +1 nucleosomes (Struhl and Segal, 2013).

Defining classes of promoters Promoter classes have been defined according to several different features (reviewed in Haberle and Lenhard, 2016; Lenhard *et al.*, 2012). A predominant classification relies on the structure of the Transcription Initiation Clusters (TICs). TICs can be described as either sharp or broad, based on the transcription start profile obtained by CAGE-seq (Carninci *et al.*, 2006). Promoters with sharp TICs ("type I promoters") are enriched for core promoter elements such as the TATA box while those with broad TICs ("type II promoters") are enriched for other elements such as the Initiator. In vertebrates, this classification overlaps to some extent with CpG islands : usually only one short CpG island is found over the TSSs of type I promoters while larger CpG islands cover type II promoters. A third class, "type III promoters", encompasses promoters covered by very large CpG islands often extending well into the gene body. Sharp and broad promoters also have distinct patterns of nucleosome positioning and histone

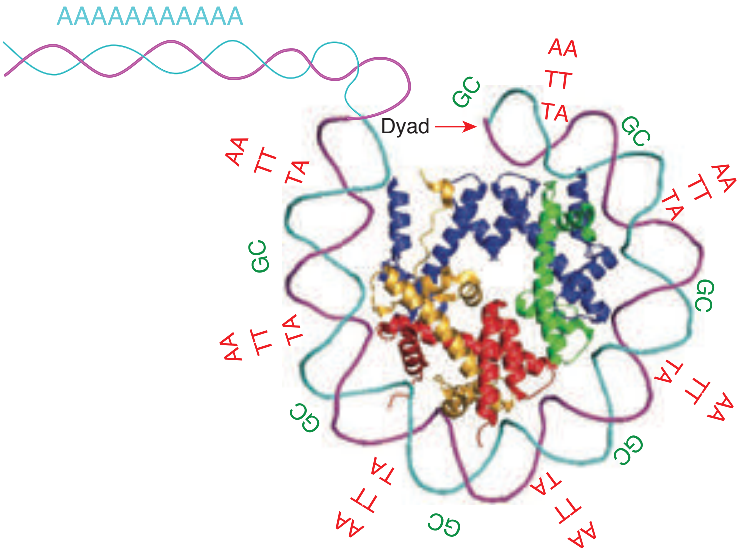


FIGURE 1.17 – Biases in nucleosomal sequences. Long stretches of A/T disfavor nucleosome positioning and are enriched in nucleosome-depleted regions, while 10-bp periodic W (A/T) or S (G/C) dinucleotide signals are enriched in nucleosomal sequences (adapted from Struhl and Segal, 2013).

modification.

A functional annotation of these promoter classes has been proposed based on GO enrichment analyses (Carninci *et al.*, 2006; Haberle and Lenhard, 2016; Lenhard *et al.*, 2012). Mostly based on GO term enrichment analysis, it has been suggested that type I promoters are associated with genes with tissue-specific expression in adult tissues whereas type II promoters are associated with genes expressed across all tissues, and type III promoters control developmentally regulated genes. This proposed functional classification was initially useful to refine the textbook notion of promoters, and contributed to clarify the relationship between the structure and the function of a promoter. However, it did not take into account the fact that genes can harbor promoters of different types (*e.g.* a ubiquitously active promoter and a tissue-specific promoter). Furthermore, only a minority of the promoters have a "sharp" TIC or a TATA-box (~ 20 to 25%, Chen *et al.*, 2013; Sandelin *et al.*, 2007) and could not explain the overall abundance of tissue-specific gene expression genome-wide. Thus, directly determining characteristics shared by functional sets of promoters with similar activity (*e.g.* tissue-specific promoters) could help to refine our understanding of the relationship between the structure and the function of a promoter.

1.2.1.2 Enhancers remotely modulate transcription

Enhancers resemble promoters but do not lead to productive transcription In addition to promoters, enhancers represent the other major class of regulatory elements. Enhancers are specific loci of the genome which can modulate the transcriptional output of associated promoter(s) (Figure 1.18A and Shlyueva *et al.*, 2014).These loci can be located in close proximity (~ 200 bp to 1 kb) or up to several kilobases upstream or downstream from the promoter they regulate (Levine, 2010; Shlyueva *et al.*, 2014) and in some extreme cases, enhancers can be located hundreds of kilobases to few megabases away from their target promoters (Joshi *et al.*, 2015). Many enhancers are located within the transcribed region of the gene they regulate, particularly in the first intron (Park *et al.*, 2014). Enhancers are typically enriched for transcription factor binding sites which, when unmasked upon nucleosome remodeling, recruit specific transcription factors. These transcription factors can modulate the transcriptional activity of an associated promoter. Thus, distant enhancers can drastically increase the transcriptional output of promoters, sometimes more than 100-fold (Figure 1.18A).

Importantly, the differences in the structural features of promoters and enhancers responsible for their specific functions still remain unclear. Enhancers share a lot in common with broad bi-directional promoters (Andersson *et al.*, 2015; Core *et al.*, 2014): both classes of regulatory elements feature NDRs, they can recruit functional PICs and they usually lead to bi-directional initiation of transcription. The major difference currently observed between the two classes actually lays *outside* of the regulatory regions (Figure 1.19). In some organisms, enhancers are enriched for polyadenylation sites immediately around their NDR while splicing sites are located in proximity downstream of a promoter NDR. This leads to differences in RNA stability: enhancer RNAs (eRNA) and upstream antisense RNAs (uaRNA) are rapidly degraded by exosomes while messenger RNAs are recognized and handled by spliceosomes (Andersson and Sandelin, 2019).

The mechanisms by which a transcription factor bound to a distal enhancer modulates the transcriptional output of a promoter has been a long-standing question. It is now clear that DNA adopts a hierarchical multi-level spatial organization on top of the fundamental chromatin fiber (see 1.2.2.2 and Serizay and Ahringer, 2018). Regulatory loops are a major type of higher-order organization of the

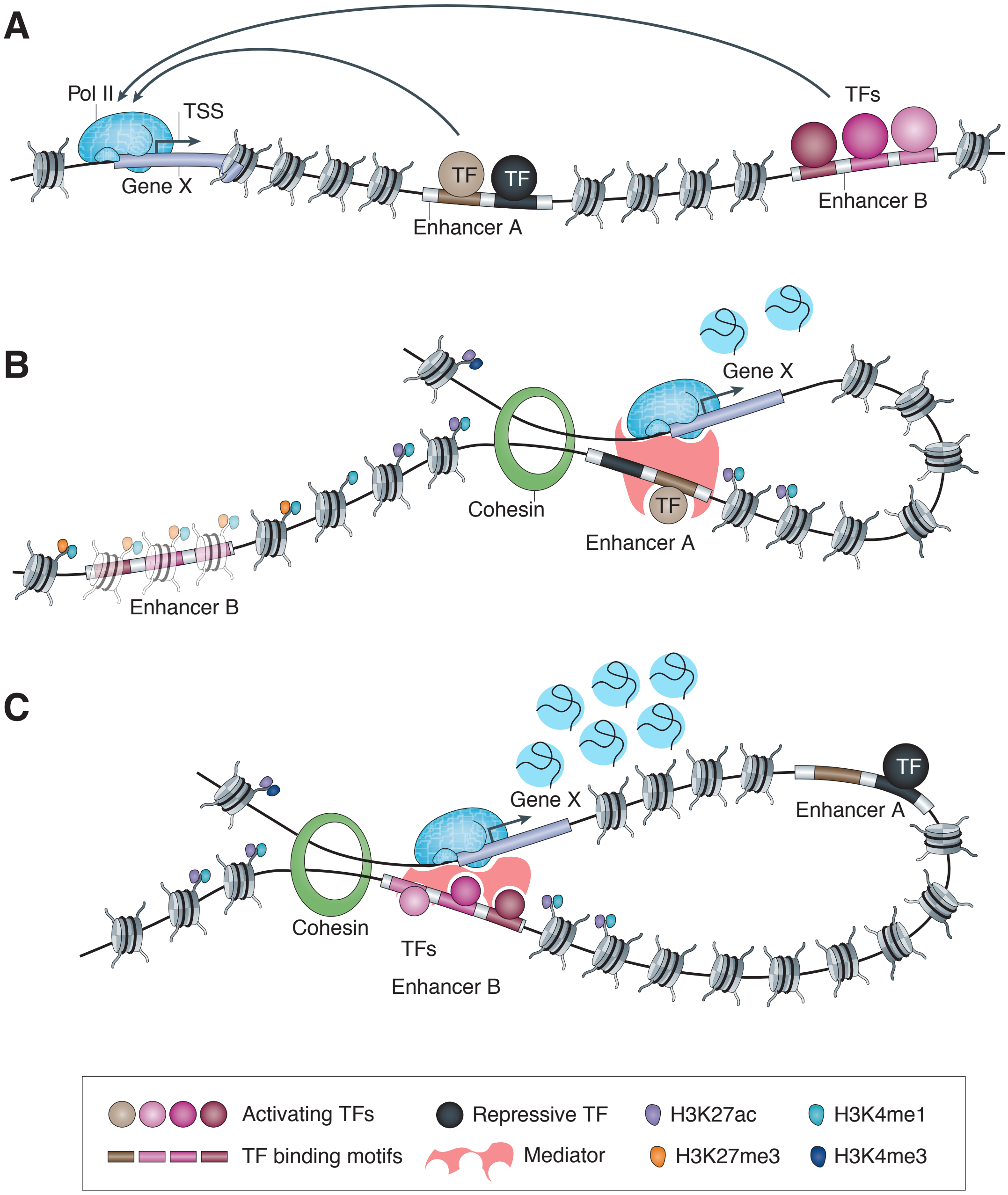


FIGURE 1.18 – Enhancers can modulate promoter transcription activity. **A-** Enhancers recruiting different transcription factors can modulate the transcriptional activity of an associated promoter. **B-** and **C-** Given the cellular context, promoter-enhancer physical interactions – usually mediated by the Mediator protein complex (in red) – differently modulate the transcriptional activity of an associated promoter (adapted from Shlyueva *et al.*, 2014).

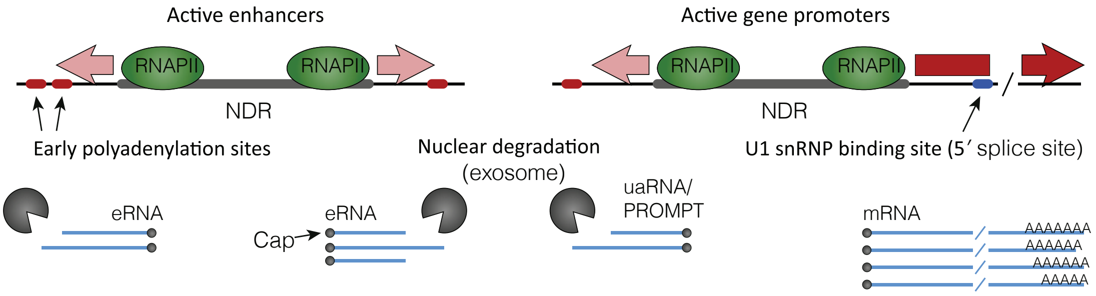


FIGURE 1.19 – Differences between promoters and enhancers. The stability of the transcripts originating from both ends of an NDR determine the main activity of a regulatory element. At promoters, an elongating transcript rapidly encounters 5' splice sites which promote its splicing, leading to an increase in stability (adapted from Andersson *et al.*, 2015).

chromatin fiber (Robson *et al.*, 2019; Schoenfelder and Fraser, 2019). Physical chromatin looping could directly link two regulatory elements together and conceptually explains how enhancers can act from afar by bringing transcription factors in spatial proximity to a promoter (Spitz, 2016). In many organisms, cohesin and Mediator are the major protein complexes involved in enhancer-promoter (E-P) loops: cohesin has a ring-like structure and actively extrudes chromatin into a loop while Mediator is thought to act as a bridge, bringing enhancers and promoters close together (Figure 1.18B-C and Fudenberg *et al.*, 2016; Shlyueva *et al.*, 2014).

1.2.1.3 Insulators limit the communication between enhancers and promoters

Insulators are another class of regulatory elements that function as physical barrier. Contrary to enhancers and promoters, insulators indirectly influence transcription. One example is CTCF, which binds DNA in a sequence-specific manner. When interacting with its binding site, CTCF *de facto* generates an insulator by acting as a bulky factor physically limiting interactions (Figure 1.20). In collaboration with the cohesin factor, it is thought to participate to chromatin loop extrusion, eventually forming a self-contained isolated neighborhood with limited interactions with regulatory elements outside of this domain (Figure 1.20 and Dowen *et al.*, 2014; Hnisz *et al.*, 2016).

CTCF is evolutionarily conserved across most bilaterians but is absent in other metazoans, plants and fungi (Heger *et al.*, 2012, 2009). However, within bilaterians, Platyhelminthes and some nematodes including *C. elegans* appear to have lost

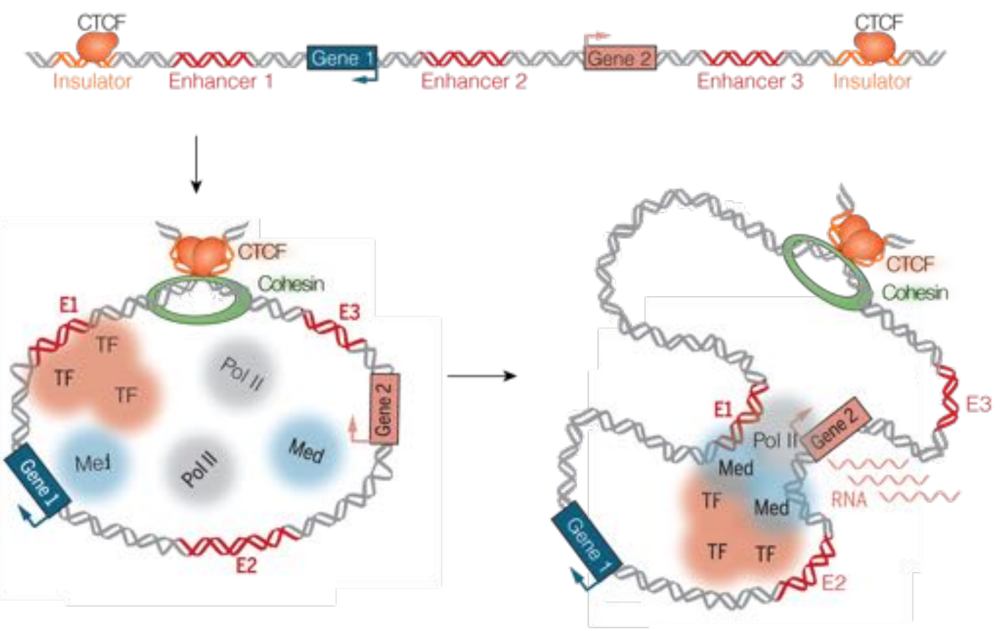


Figure 1.20 – Role of insulators in chromatin higher-order structure. Insulator proteins such as CTCF bind to insulator loci and form homodimers effectively blocking loops from being further extruded, thus creating an isolated neighborhood. Within this environment, E-P loops can be formed without interfering with the rest of the genome (adapted from Furlong and Levine, 2018).

CTCF, yet retain an organized chromatin architecture (Crane *et al.*, 2015; Huang *et al.*, 2018). This raises the question of which chromatin loci and which factors are important to modulate chromatin organization in these species.

1.2.2 Regulation of gene expression: beyond regulatory elements

1.2.2.1 Nascent mRNA are co-transcriptionally processed

Initiation of transcription is a highly regulated biological process. However, elongating and mature transcripts still face additional regulation (Figure 1.21 and Li and Manley, 2006; Maniatis and Reed, 2002; Saunders *et al.*, 2006; Shatkin and Manley, 2000). Soon after transcription initiation, a cap is added to the 5' end of the nascent mRNA. Additional protein complexes also bind to the elongating mRNA, acting as chaperone to ensure its adequate packaging. Transcripts are spliced while being synthesized and a polyadenylated 3' tail is added to them when their transcription is achieved. Improperly processed transcripts are recognized and degraded by surveillance complexes in the nucleus.

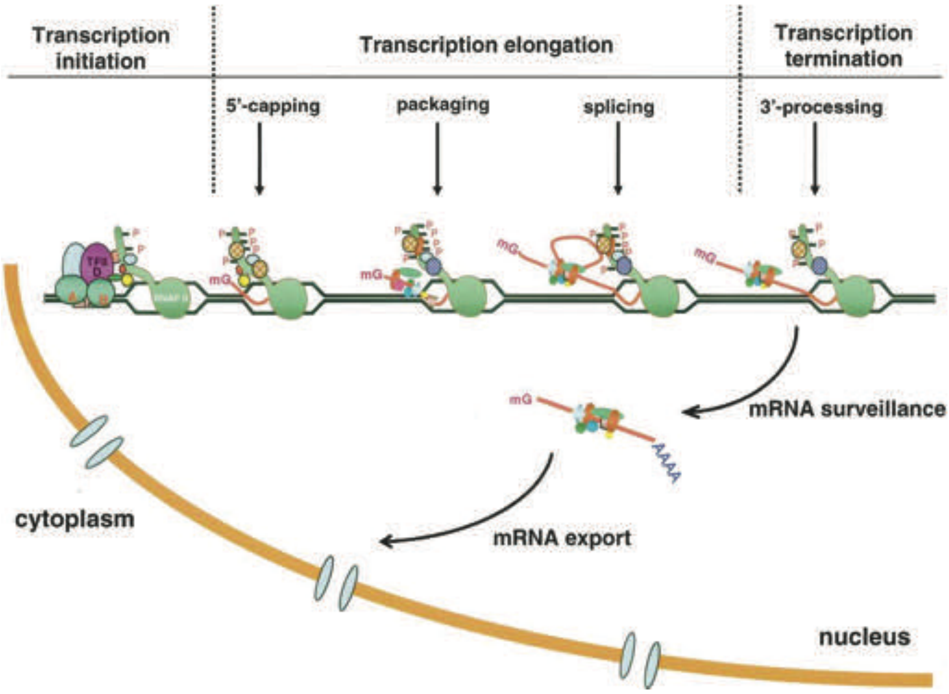


Figure 1.21 – Co- and post-transcriptional control of RNA. After initiation of transcription, the nascent RNA is capped, packaged and spliced while elongating. Transcription terminates when RNAPII encounters termination signals and the RNA is polyadenylated. These modifications protect the mRNA from being degraded by surveillance complexes so that it can be safely exported outside of the nucleus (adapted from Li and Manley, 2006).

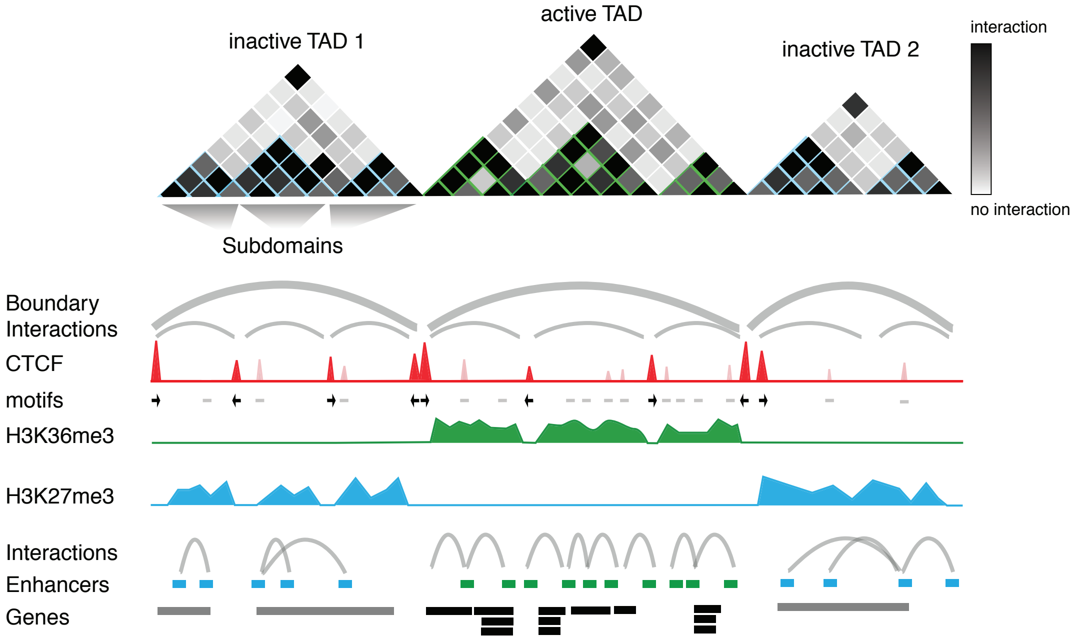


Figure 1.22 – Large-scale chromatin organization in Topological Associating Domains (TADs). Enhancers and promoters located within a TAD preferentially interact with each other (adapted from Ali *et al.*, 2016).

1.2.2.2 Chromatin is spatially organized into higher-order domains

Gene expression is also regulated by the hierarchical multi-level spatial organization of chromatin. Beyond the aforementioned enhancer-promoter interactions, the chromatin adopts a higher-order 3D architecture. Topological Associating Domains (TADs) are a larger-scale fundamental feature of the genome spatial architecture, essentially acting as megabases-large isolated neighborhoods (see 1.2.1.3). Fine regulatory interactions such as E-P loops are contained within each TAD and inter-TAD physical interactions are inhibited (Figure 1.22 and Ali *et al.*, 2016). Since their initial identification in metazoans (Dixon *et al.*, 2012; Hou *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012), the TADs have been under a lot of scrutiny and their main mechanisms of formation and structural characteristics are now better defined.

I published a more exhaustive review of the spatial organization of the chromatin at different scales as well as the differences observed across species (Serizay and Ahringer, 2018), which can be found in the Appendix Chapter A.

1.2.3 Combinations of regulatory elements control gene expression in development

1.2.3.1 Genome-wide atlases of regulatory elements in development

Regulatory elements play a central role in modulating chromatin transcriptional output, and the last decade has witnessed increasing efforts to comprehensively annotate them. Active regulatory elements are characterized by a nucleosome-depleted region, and genome-wide DNA accessibility profiling has been largely used as a proxy to identify putative regulatory elements. Tens of thousands of accessible genomic loci have been mapped across metazoans using genome-wide chromatin accessibility profiling techniques (mainly DNase-seq and/or ATAC-seq, see 1.1.3.2), thus generating atlases of cis-regulatory elements in many metazoans (Andersson *et al.*, 2014; Bulut-Karslioglu *et al.*, 2018; Jänes *et al.*, 2018; Liu *et al.*, 2019a; Nègre *et al.*, 2011; Quillien *et al.*, 2017; Thurman *et al.*, 2012).

Regulatory element accessibility has been extensively studied during development in worm (Daugherty *et al.*, 2017; Jänes *et al.*, 2018), fly (Bozek *et al.*, 2019; Cusanovich *et al.*, 2018a; Haines and Eisen, 2018; Thomas *et al.*, 2011), fish (Pálfy *et al.*, 2019; Uesaka *et al.*, 2019), mouse (Uesaka *et al.*, 2019; Wu *et al.*, 2016) and human (Gao *et al.*, 2018; Liu *et al.*, 2019b). Moreover, some studies have performed tissue-specific accessibility profiling, relying either on single-cell approaches (Cusanovich *et al.*, 2018a) or bulk tissue analysis (Liu *et al.*, 2019a; Werber *et al.*, 2014). These studies showed that a majority of regulatory elements undergo dynamic changes of accessibility during development.

In *C. elegans*, the Ahringer lab has pioneered the genome-wide functional annotation of regulatory elements in development. Sequencing of short nuclear RNA with a 5'-cap was used to profile transcription initiation clusters (TICs, equivalent to CAGE tag clusters) ("short-capped RNA-seq", Chen *et al.*, 2013). Simultaneously, longer capped nuclear transcripts were sequenced to profile transcription elongation events (*i.e.* productive elongation) ("long-capped RNA-seq", Chen *et al.*, 2013). Together, these data provided insights in the transcription landscape across the entire genome and throughout development. Later, ATAC-seq experiments were performed and led to the annotation of 42,247 accessible chromatin loci (Jänes *et al.*, 2018). The integration of the transcription initiation landscape (short

Table 1.1 – Regulatory classes of accessible elements annotated in Jänes *et al.*, 2018.

Type of elements	#	%
Uni-directional promoters	11,620	27.51
Bi-directional promoters	1,976	4.68
Putative enhancers	19,231	42.52
Non-coding RNA	824	1.95
Pseudogene promoters	291	0.69
Unknown promoters	1,791	4.24
Other elements	6,512	15.41

capped RNA-seq), the productive elongation landscape (long capped RNA-seq) and the chromatin accessibility landscape (ATAC-seq) allowed to functionally annotate these chromatin loci into sets of promoters, enhancers and other classes (Table 1.1). Accessible loci overlapping one or more TICs and with productive transcription were annotated as promoters. This resulted in 13,596 promoters being annotated to 11,196 genes of the 20,222 protein-coding genes annotated in the *C. elegans* genome. 19,231 other accessible loci overlapping TICs but without productive elongation were annotated as putative enhancers. Finally, the remaining accessible sites were annotated based on the genetic feature they overlap with (*e.g.* non-coding RNA). The exact strategy used to functionally annotate accessible chromatin loci is further detailed in Jänes *et al.* (2018) (Appendix Chapter B). In this study, nucleosome occupancy and histone modifications landscapes have also been generated, confirming and improving the current understanding of regulatory element molecular organization (Jänes *et al.*, 2018).

1.2.3.2 Regulation of gene expression during development

Promoter activity is dynamic during development During development, a single cell gives rise to progenitors, which further differentiate in various lineages. Ultimately, tissues and organs are formed and fulfill specific biological functions. This requires large-scale remodeling of chromatin, and many genes become expressed in restricted cell types at specific developmental stages. Regulatory elements control these distinct spatiotemporal patterns of gene expression. The genome-wide "atlases" of regulatory elements have been instrumental to investigate the dynamics of regulation of gene expression during development.

Transcription factors also play a central role in gene regulation at each develop-

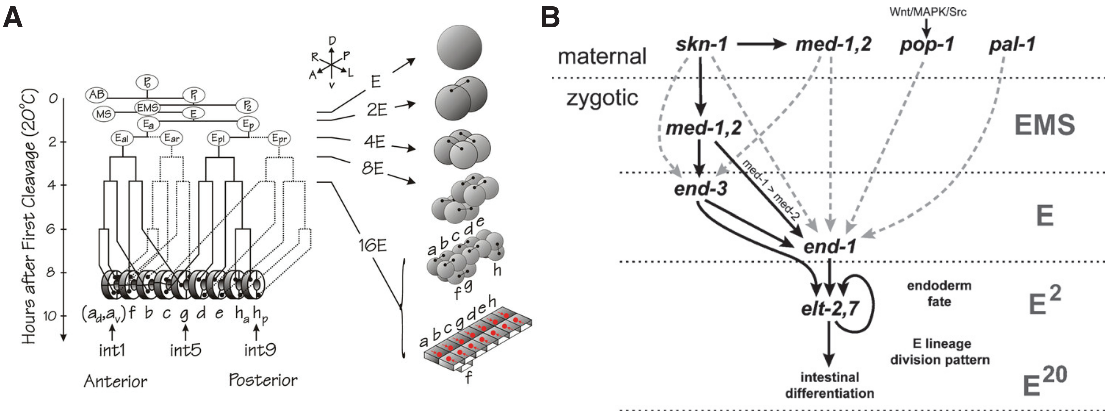


Figure 1.23 – Differentiation of the endoderm in *C. elegans*. **A-** The lineage of the twenty clonal cells forming the worm intestine originates from the E cell (adapted from McGhee, 2007). **B-** The differentiation of the E cell precursor into twenty daughter cells depends on the specific transcription factors *med-1*, *med-2*, *end-1*, *end-3*, *elt-2* and *elt-7* regulating each other through feed-forward loops (adapted from Maduro *et al.*, 2007).

mental stage, from cell specification to organogenesis and later on in post-embryonic development. They coordinate the expression of functional sets of ubiquitous or tissue-specific genes involved in similar processes. A canonical example of transcription factors precisely regulating the expression of tissue-specific genes is the skn-1—med-1,2—end1,3—elt-2,7 cascade in *C. elegans*. This cascade regulates the transcription of intestinal genes early on during embryogenesis development (Figure 1.23). Thus, characterizing the spatiotemporal activity of transcription factors would bring insights in their contribution to tissue-specific gene regulation during development.

Finally, chromatin architecture largely varies during development and contributes to gene regulation. For example, in fly, both local and global chromatin remodeling events occur during embryogenesis and contribute to the activation of different sets of genes (Figure 1.24).

Thus, characterizing the activity of regulatory elements and transcription factors as well as the chromatin architecture during development is crucial to better understand the mechanisms of tissue-specific gene regulation.

Different intrinsic promoter sequences are associated with positioning of TSSs during development The mechanism by which transcription is initiated at promoters can also vary during development. Haberle et al. profiled transcription initiation landscape throughout Zebrafish embryonic development (Haberle *et al.*,

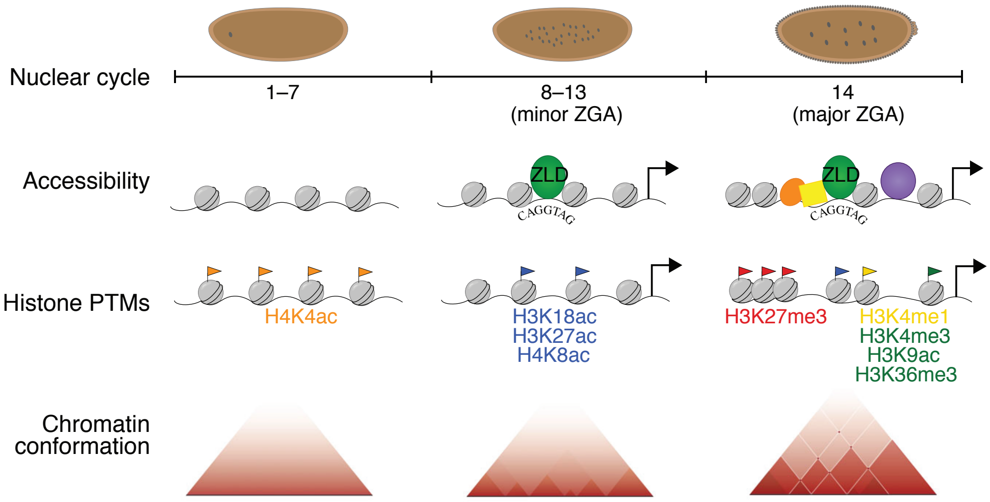


Figure 1.24 – Remodeling of chromatin architecture during embryogenesis in fly. Between nuclear cycle 8 and 13, Zelda pioneer factor is synthesized and binds to nucleosomal DNA, leading to rearrangement of nucleosome organization at a battery of promoters. This contributes to inducing the major ZGA at NC 14. At this stage, newly synthesized transcription factors initiate gene transcription and genome zygotic 3D architecture is acquired (adapted from Hamm and Harrison, 2018).

2014). This study revealed that within individual promoters, the nature of the DNA sequences associated with the position of transcription initiation change throughout development. In oocytes, the position of the dominant TSS of thousands of constitutively expressed genes appears aligned ~ 30 bp downstream of a degenerated TATA-box ("W-box"). Once the genome is activated in zygotic cells, TSSs can shift up to few tens of bp and appear aligned ~ 50 bp upstream of +1 nucleosomes, whose positioning is facilitated by the underlying DNA sequence (Figure 1.25 and Haberle *et al.*, 2014). This groundbreaking study clearly showed that the position of TSSs could be directly aligned (or not) with nucleosome-positioning sequences.

A limitation of this study is the comparison of the transcription initiation grammar used in one cell type (the oocytes) with that used in cells from embryos harboring differentiating tissues. This mixed-tissues state prevented the authors of the study from identifying tissue-specific transcription initiation grammars. Going further, it would be particularly interesting to investigate the differences in transcription initiation grammars at promoters active in different sets of fully differentiated tissues.

Regulatory grammars contribute to gene regulation Gene regulatory grammars are defined by the composition, arrangement and activities of regulatory

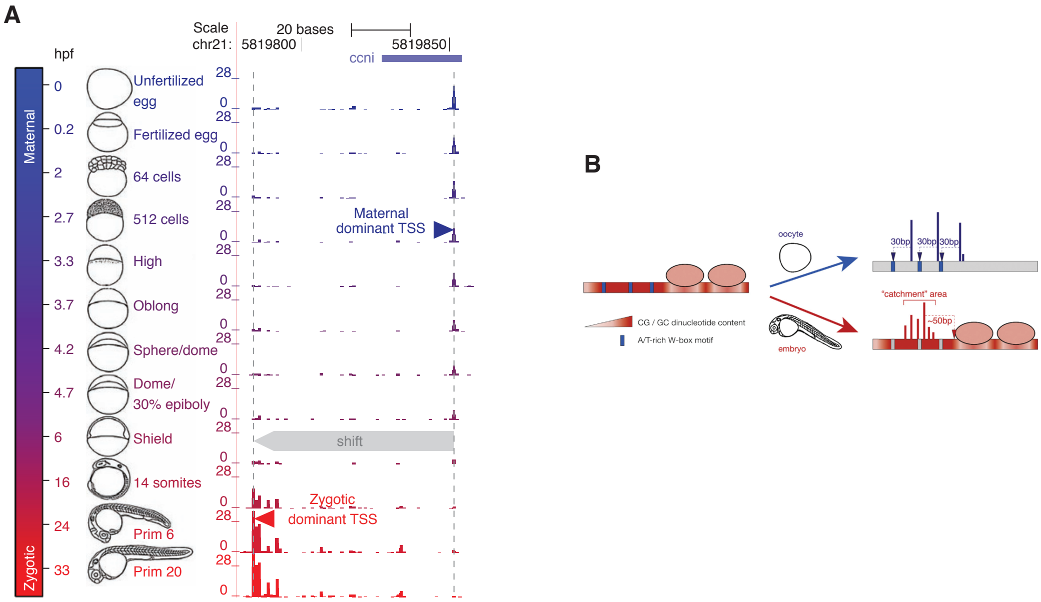


FIGURE 1.25 – Switch between two overlapping transcription initiation grammars during ZGA in zebrafish. **A-** Transcription initiation landscape at the cyclin 1 locus at different time points throughout the embryonic development. Note the shift of TSS clusters before and after ZGA. **B-** Two different underlying transcription initiation grammars (adapted from Haberle *et al.*, 2014 and Haberle and Lenhard, 2016).

elements that control transcriptional patterns of gene expression (Levine, 2010; Spitz and M Furlong, 2012; Yáñez-Cuna *et al.*, 2013). Different regulatory grammars have been described, ranging from single promoters to complex structures of multiple alternative promoters, and regulatory elements can operate redundantly, hierarchically, additively or synergistically (Bahr *et al.*, 2018; Davuluri *et al.*, 2008; Guerrero *et al.*, 2010; Herr, 1993; Osterwalder *et al.*, 2018; Whyte *et al.*, 2013). The striking patterns of expression observed for structural genes in fly embryos are canonical examples of cooperation between regulatory elements. For instance, several enhancers are present around the *eve* locus and each one drives a specific pattern of spatial expression for *eve*. The combination of all the enhancers leads to the overall banding pattern observed for *eve* expression (Figure 1.26A).

Importantly, the function of regulatory elements can change throughout development. For instance, the same set of enhancers can activate two different promoters in different cellular contexts. Perhaps the most famous example of such developmentally regulated process is the transition from fetal γ-globin to adult β-globin (Figure 1.26B and Sankaran and Orkin, 2013). A cluster of enhancers known as the Locus Control Region (LCR) is located ~ 50 kb upstream of two globin

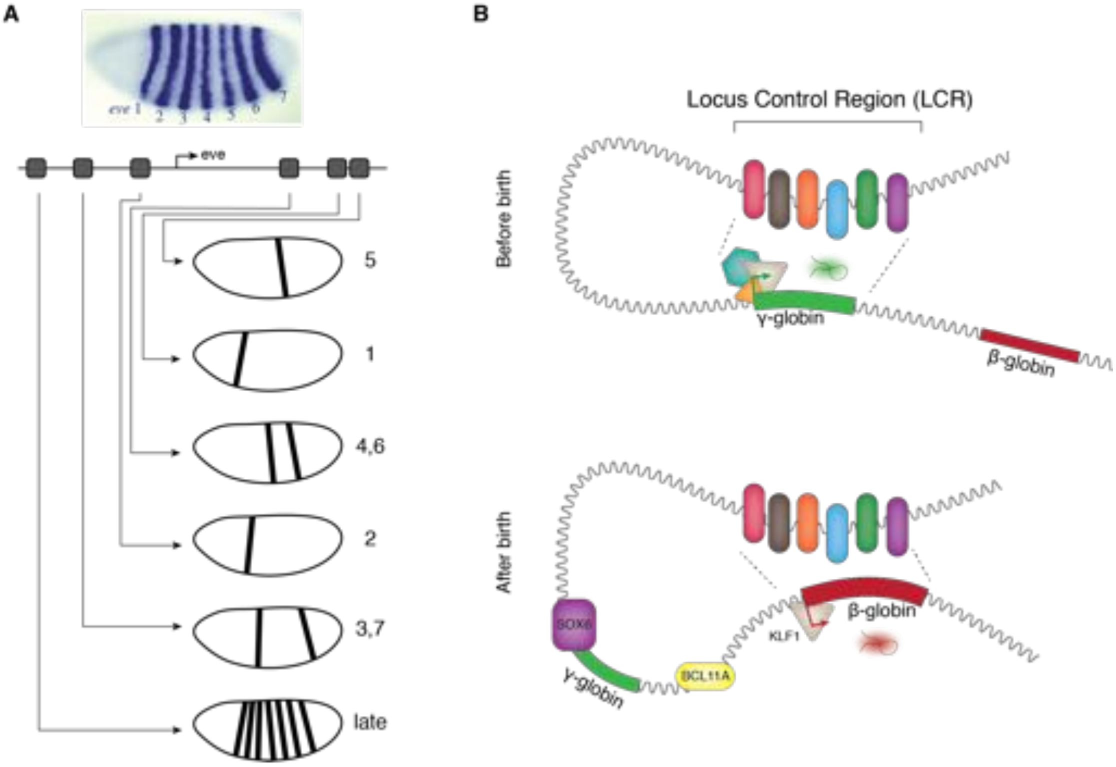


FIGURE 1.26 – Regulatory grammars in development. **A-** Pattern of expression of the gene *eve* (top, photo from Andrioli *et al.*, 2002) and contribution of six different associated enhancers to the overall pattern of expression (bottom). **B-** The human Locus Control Region (LCR) regulating the transcription of a γ-globin before birth switches to regulating the transcription of a neighboring β-globin after birth.

genes, γ-globin and β-globin. During embryonic development and up to the birth, the γ-globin is transcribed. However, the LCR–γ-globin interactions occurring during embryogenesis are disrupted after birth by new transcription factors such as SOX6, FOP and BCL11A, and KLF1, which rewire the LCR interactions with the β-globin promoter. Furthermore, alternative promoters of the same gene can also be activated in different contexts, a mechanism known as "promoter switching". This can play a major role in development as the resulting mRNAs isoforms could harbor different coding sequences (Pecci *et al.*, 2001; Pozner *et al.*, 2007). Finally, perturbation of the linear and/or spatial arrangement of regulatory elements can lead to pathologies (Chatterjee and Ahituv, 2017; Lupiáñez *et al.*, 2015; Parker *et al.*, 2013; Schaub *et al.*, 2012).

Overall, these examples illustrate the importance of characterizing the combinatorial activity of regulatory elements, which defines regulatory grammars, to understand the mechanisms of gene regulation during development.

1.3 Using *C. elegans* to study gene regulation in development

In Chapter 1.2, I presented the molecular mechanisms of gene regulation in metazoans and the central role played by regulatory elements. Based on these mechanisms, a variety of gene expression patterns are obtained in different cellular and organismal contexts. In Chapter 1.3, I introduce *Caenorhabditis elegans* as a model system to study gene regulation during development.

1.3.1 *C. elegans* is a powerful system to study cell-type specific control of gene expression

1.3.1.1 *C. elegans* has pioneered the field of modern genetics and genomics

Characterized by a rapid 3-day life cycle and a large brood size (~ 200 progeny), *Caenorhabditis elegans* stands as an attractive system to study development. Moreover, the ability to precisely edit its genome, the fact that it is (mostly) a self-fertilizing hermaphrodite and that males can be used for genetic crosses make *C.*

elegans an ideal model organism to investigate the principles governing *in vivo* gene regulation in metazoans (Riddle *et al.*, 1997). The worm has emerged as a model system for genetics relatively recently compared to *Drosophila*, but the fact that it was the first metazoan to have its genome fully sequenced is a testament to how quickly it gained popularity among the community of geneticists toward the end of the 1990 decade (C. elegans Sequencing Consortium, 1998).

The 100 megabase *C. elegans* genome has many similarities with more complex metazoans, notably featuring gene structure with introns and exons, alternative splicing, regulatory sequences as well as intergenic regions with repeated sequences, albeit generally smaller than in other metazoans (Spieth *et al.*, 2018). *C. elegans* research has pioneered many breakthrough genetic approaches including large-scale genetic screens, the use of GFP as a biological marker and the use of RNA interference. It also paved the way for biological discoveries including mechanisms of dosage compensation, genes involved in apoptosis, mechanisms of acquisition of embryo polarity and mechanisms of trans-generational inheritance (Table 1.2 and Corsi *et al.*, 2015).

1.3.1.2 *C. elegans* tissues are formed by an invariant cell lineage

Perhaps the most striking feature of *Caenorhabditis elegans* development is its fixed cell lineage, which leads to consistency of cell number and cell position from individual to individual during development. Its cell lineage has been entirely determined from the single-cell embryo to the mature adult by microscopic observation of cell divisions and cell migrations in the transparent animal (Figure 1.27 and Sulston and Horvitz, 1977; White *et al.*, 1986). It has proven to be an excellent model system with a high connectivity to human biology, and worm-based studies have unlocked a wealth of knowledge on multiple facets of biology.

Benefiting from its constant cell lineage and the collection of techniques available, many cell differentiation regulatory pathways have been identified in the worm (Corsi *et al.*, 2015). For instance, the gut develops clonally from the unique E blastomere, born at the 8-cell embryo stage. During embryogenesis, this blastomere gives rise to twenty cells whose identity and function are maintained throughout the life of the worm. The cascade of key transcription factors involved in this differentiation process has been extensively studied (Figure 1.23 on page 33 and

TABLE 1.2 – Selected discoveries in *C. elegans* research (adapted from Corsi *et al.*, 2015).

Year	Discovery	References
1974	Identification of mutations that affect animal behavior	Brenner 1974; Dusenberry et al. 1975; Hart 2006
1977	First cloning and sequencing of a myosin gene	Macleod et al. 1977
1977	Genetic pathways for sex determination and dosage compensation described	Hodgkin and Brenner 1977; Meyer 2005; Zarkower: 18050479
1981	Identification of mutations affecting touch sensitivity	Sulston et al. 1975; Chalfie and Sulston 1981
1981	First germline stem cell niche identified	Kimble and White 1981; Kimble and Crittenden 2005
1983	First complete metazoan cell lineage	Sulston and Horvitz 1977; Kimble and Hirsh 1979; Sulston et al. 1983
1983	Discovery of apoptosis (cell death) genes	Hedgecock et al. 1983; Ellis and Horvitz 1986; Yuan and Horvitz 1992; Yuan et al. 1993; Conradt and Xue 2005
1984	Identification of heterochronic genes	Ambros and Horvitz 1984; Slack and Ruvkun 1997
1986	First complete wiring diagram of a nervous system	White et al. 1986; Jarrell et al. 2012; White 2013
1987	Discovery of the first axon guidance genes	Hedgecock et al. 1987, 1990; Culotti 1994
1988	Asymmetric distribution of cellular components in embryos by par proteins	Kemphues et al. 1988; Gönczy and Rose 2005
1993	Demonstration of a role for insulin pathway genes in regulating lifespan	Friedman and Johnson 1988; Kenyon et al. 1993; Kimura et al. 1997; Collins et al. 2007
1993	First microRNA (lin-4) and its mRNA target (lin-14) described	Lee et al. 1993; Wightman et al. 1993; Vella and Slack 2005
1993	Identification of nonsense-mediated decay genes	Pulak and Anderson 1993; Hodgkin 2005
1994	Introduction of GFP as a biological marker	Chalfie et al. 1994; Boulin et al. 2006
1998	First metazoan genome sequenced	C. elegans Sequencing Consortium 1998; Schwarz 2005
1998	Discovery of RNA interference (RNAi)	Fire et al. 1998
2000	Development of genome-wide RNAi screening	Fraser et al. 2000; Kamath et al. 2001
2000	Transgenerational inheritance and its mediation by piRNA	Grishok et al. 2000; Ashe et al. 2012
2005	First use of channelrhodopsin optogenetics in an intact animal	Nagel et al. 2005 16360690

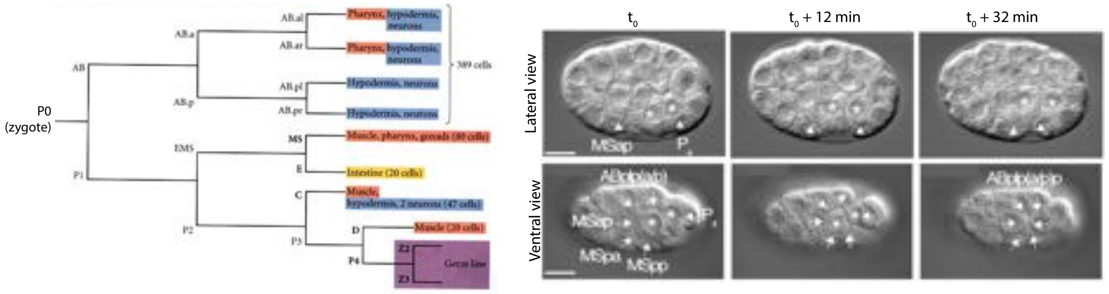


FIGURE 1.27 – Cell fate specification during early *C. elegans* embryonic development. Left: abbreviated cell lineage tree (up to the 8-cell stage) and nature and number of derived cells at the time of hatching (adapted from Gilbert, 2000). Right: Example of annotation of cells by confocal imaging during gastrulation. Asterisks indicate the differentiating Ea and Ep cells, and neighboring cells are labeled with arrows. Note that Ea and Ep ingress towards the center of the embryo during gastrulation and are eventually surrounded by MSap and P4 (adapted from Lee and Goldstein, 2003).

Maduro *et al.*, 2007; McGhee, 2007). Tissue-specific gene regulation of biological processes has also been tackled in post-embryonic development. For instance, the different steps of proliferation and maturation of the germline in hermaphrodite worms have been extensively studied.

Overall, the mechanisms of gene regulation involved in tissue-specific processes are still poorly appreciated. Improving our understanding of the tissue-specific chromatin organization regulating networks of genes during development would help to shed light on many key biological processes, from the control of cell differentiation to the physiological functions of individual organs.

1.3.1.3 Challenges arising from studying tissue-specific genetics in *C. elegans*

C. elegans is a small organism (~ 1 mm when it has reached adulthood) protected by a resistant cuticle. The worm community clearly benefited from these characteristics which for example allow one to safely preserve thousands of strains in a single container indefinitely (Stiernagle, 2006). However, for the same reasons, isolating genetic material from individual tissues is more complicated in the worm than in other larger model organisms (*e.g.* fly, fish, mice). The easiest approach is to extract cells from embryos. This yields relatively healthy living cells and isolated blastomeres can even be cultured and differentiated *in vitro* (Sangaletti and Bianchi, 2013). However, getting access to whole cells of the worm after hatching at larval stages is challenging and requires lengthy and inefficient cuticle dissociation methods (Zhang *et al.*, 2011). This represented a major hurdle hindering the use of traditional methods widely used in other organisms such as fluorescent-activated cell sorting (FACS) or antigen-based cell purification to study tissue-specific post-embryonic development. Thus, while large-scale screens and candidate-based perturbation assays can be performed, using genomic assays to study tissue-specific gene regulation in differentiated tissues remains challenging and requires the design and optimization of a method to isolate material from individual tissues.

The embryonic development is very short in worm (14 hours at 20-22 °C). If this is an advantage when seeking to obtain a lot of individuals, it also implies that it is challenging to capture intermediate cell stages which undergo fundamental chromatin remodeling during specification and differentiation (*e.g.* there is only E cell per embryo and it only exists for few minutes before dividing into daughter cells). Traditional steady-state genome-wide approaches do not appear adapted to capture events transiently occurring during a very brief period of time.

1.3.2 Preliminary studies: focusing on gene expression

Many studies have tried to characterize gene expression in specific tissues of the worm, using a variety of approaches (Table 1.3). Originally, tissue-specific cells have been obtained from embryonic cell dissociation followed by *in vitro* differentiation and FACS-based isolation (Fox *et al.*, 2005, 2007; Zhang *et al.*, 2002) but this could not be used for RNA quantification in adult tissues *in vivo*. Soon after, mRNA tagging methods emerged, using for instance poly(A)-binding protein co-immunoprecipitation, and have been extensively used to study steady-state levels of mature transcripts in bulk tissues (Blazie *et al.*, 2015; Ma *et al.*, 2016; Pauli *et al.*, 2006; Roy *et al.*, 2002) or even single cell types (Spencer *et al.*, 2011; Stetina *et al.*, 2007; Takayama *et al.*, 2010). Around the same time, realizing the limitations of tiling microarrays used between 2002 and 2009, the worm community focused on large-scale microscopy-based reporter assays (Dupuy *et al.*, 2007; Hunt-Newbury *et al.*, 2007; McKay *et al.*, 2003; Murray *et al.*, 2008, 2012). With the improvements in flow cytometry and the emergence of high-throughput RNA-sequencing after 2010, an increase of FACS-based studies generated new high-quality datasets covering whole tissues (Haenni *et al.*, 2012; Kaletsky *et al.*, 2018; Meissner *et al.*, 2009; Warner *et al.*, 2019) or individual cell types (Kroetz and Zarkower, 2015; Spencer

Table 1.3 – Important studies quantifying gene expression or protein levels in *C. elegans* tissues or cell types.

Method of tissue isolation	**Biological material**	**Quantif. method**	**Reference**
Embryonic cell culture	Touch receptors from cultured embryo cells	Microarray	Zhang 2002
	Neuron from primary cultures of embryonic cells	Microarray	Fox 2005
	Embryonic muscle cells and cultured derivates	Microarray	Fox 2007
RNA tagging	Muscle cells in L1	Microarray	Roy 2002
	Intestine cells in L4	Microarray	Pauli 2006
	Pan-neurons and A-class cholinergic neurons in larvae	Microarray	Stetina 2007
	ASEL or ASER neurons	Microarray	Takayama 2009
	11 different subtissues in larvae	Microarray	Spencer 2011
	Intestine, pharynx and body muscle (mixed stages)	RNA-seq	Blazie 2015
	Muscle in larvae, dauer larvae and aging worms	RNA-seq	Ma 2016
Dissections	Intestines from adults	Microarray	McGhee 2007
	Dissected gonads	RNA-seq	Ortiz 2014
	Dissected gonads	3'-end-seq	West 2016
	Dissected and cryo-sectioned gonads	RNA-seq	Diag 2018
Proteomics	Seam and hyp7 epidermal cells, intestine, or neurons	Mass spec.	Waaijers 2016
	Intestine, epidermis, body wall muscle, or pharyngeal muscle	Mass spec.	Reinke 2017
Nuclei purification	Muscle cells in adults	Microarray	Steiner 2012
Reporter assays (microscopy)	802 reporters	Microscopy	McKay 2003
	~900 promoters in larvae	Microscopy	Dupuy 2007
	1886 promoters in larvae	Microscopy	Hunt-Newbury 2008
	127 genes during embryogenesis	Microscopy	Murray 2008
	93 genes in 363 specific cells from L1 stage	Microscopy	Liu 2009
FACS-based	Embryonic neurons	Microarray	Stetina 2007
	Muscle embryonic cells	SAGE	Meissner 2009
	13 different subtissues in embryo	Microarray	Spencer 2011
	Intestine nuclei in larvae	3'-end-seq	Haenni 2012
	NSM serotonergic neurons in adults	RNA-seq	Spencer 2014
	PGCs in L1	RNA-seq	Kroetz 2015
	Muscle, neuron, intestine, and epidermis cells in adult	RNA-seq	Kaletsky 2018
	Muscle, intestine, neurons, pharynx and hypodermis during emb. dev.	RNA-seq	Warner 2019
Single-cell	Very early embryos (2-cell stage)	CEL-seq	Hashimshony 2012
	L2 larvae	sciRNA-seq	Cao 2017
	Developing embryos	sciRNA-seq	Paker 2019
	Hand-picked embryos (1- to 16-cell stages)	SMART-	Tintori 2016

et al., 2014). Micro-dissections have also been useful to generate tissue-specific RNA-seq datasets in intestine (McGhee *et al.*, 2007) and in germline (Diag *et al.*, 2018; Ortiz *et al.*, 2014; West *et al.*, 2018). Finally, with the rise of single-cell techniques, incredibly precise quantification of cell-specific gene expression is now possible and single-cell experiments have been performed in embryos (Hashimshony *et al.*, 2012; Packer *et al.*, 2019; Tintori *et al.*, 2016) and larvae (Cao *et al.*, 2017).

Importantly, most of these studies have been performed in embryos or early larvae stages, and not in a systematic or comprehensive way. The only integrative tissue-specific resources covering the main tissues when all the tissues are fully formed come from Kaletsky *et al.* (2018), who performed RNA-seq after sorting nuclei from individual tissues (neurons, muscle, hypodermis and intestine) in adult worms. However, the authors did not isolate germline nuclei and consequently, one of the main tissues constituting the body of adult worms was not characterized in this study. In this study, the authors identified tissue-specific isoforms and generated a tissue-specific gene expression prediction tool. However, they did not focus on defining tissue-specific classes of genes and their structural characteristics.

Even more importantly, only a handful chromatin accessibility studies have been published in *C. elegans* (Daugherty *et al.*, 2017; Ibrahim *et al.*, 2018; Jänes *et al.*, 2018; Kolundzic *et al.*, 2018; Lee *et al.*, 2017) and the only tissue-specific dataset reported as of today is from the primordial germline and in this study, the authors focused on a mutant phenotype (Lee *et al.*, 2017). Lack of information on chromatin activity has largely hindered studies of the function of regulatory elements in gene regulation.

1.3.3 Outstanding questions

As described hereabove, patterns of gene expression throughout development have already been extensively studied. However, there is still a need for a comprehensive study of tissue-specific gene expression patterns in fully differentiated animals. Furthermore, the mechanisms of gene regulation used to obtain these patterns of tissue-specific expression remain poorly investigated. Seminal studies point to different grammars of transcription initiation between different developmental stages (Haberle *et al.*, 2014), but how the transcription of tissue-specific genes is controlled is still unclear. Notably, the differences between mechanisms of regulation of ubiquitous genes or tissue-specific genes have not been directly tackled. Finally, the role of transcription factors and chromatin organization in individual tissues has rarely been comprehensively investigated. Answering these outstanding questions would shed light on fundamental mechanisms involved in the regulation of gene expression during development.

1.4 Aims of my PhD

During my PhD, I sought to investigate regulatory elements activity in *C. elegans* to shed light on principles of temporal and tissue-specific gene regulation. I pursued three major goals described below.

Characterizing genome-wide dynamics of chromatin activity through life

At the beginning of my PhD in 2016, a project was ongoing in the lab to identify and annotate regulatory elements active at different stages of *C. elegans* development and aging. I integrated the ATAC-seq and RNA-seq datasets generated for this project along with published ChIP-seq data, to identify the general patterns of temporal chromatin activity and better characterize the regulatory element activity through development and aging of a metazoan (Figure 1.28, Aim I, Chapter 3).

Defining the regulatory architecture of tissue-specific and ubiquitous genes

I devoted the major part of my PhD to investigating the tissue-specific gene regulation in the worm (Figure 1.28, Aim II, Chapter 4, Chapter 5 and Chapter 6). I developed a sorting method to purify bulk tissues and I profiled chromatin accessibility and transcriptional activity across the five main tissues of the adult worm. With these datasets, I aimed to identify the regulatory architectures of tissue-specific and ubiquitous genes and the molecular features of their associated regulatory elements.

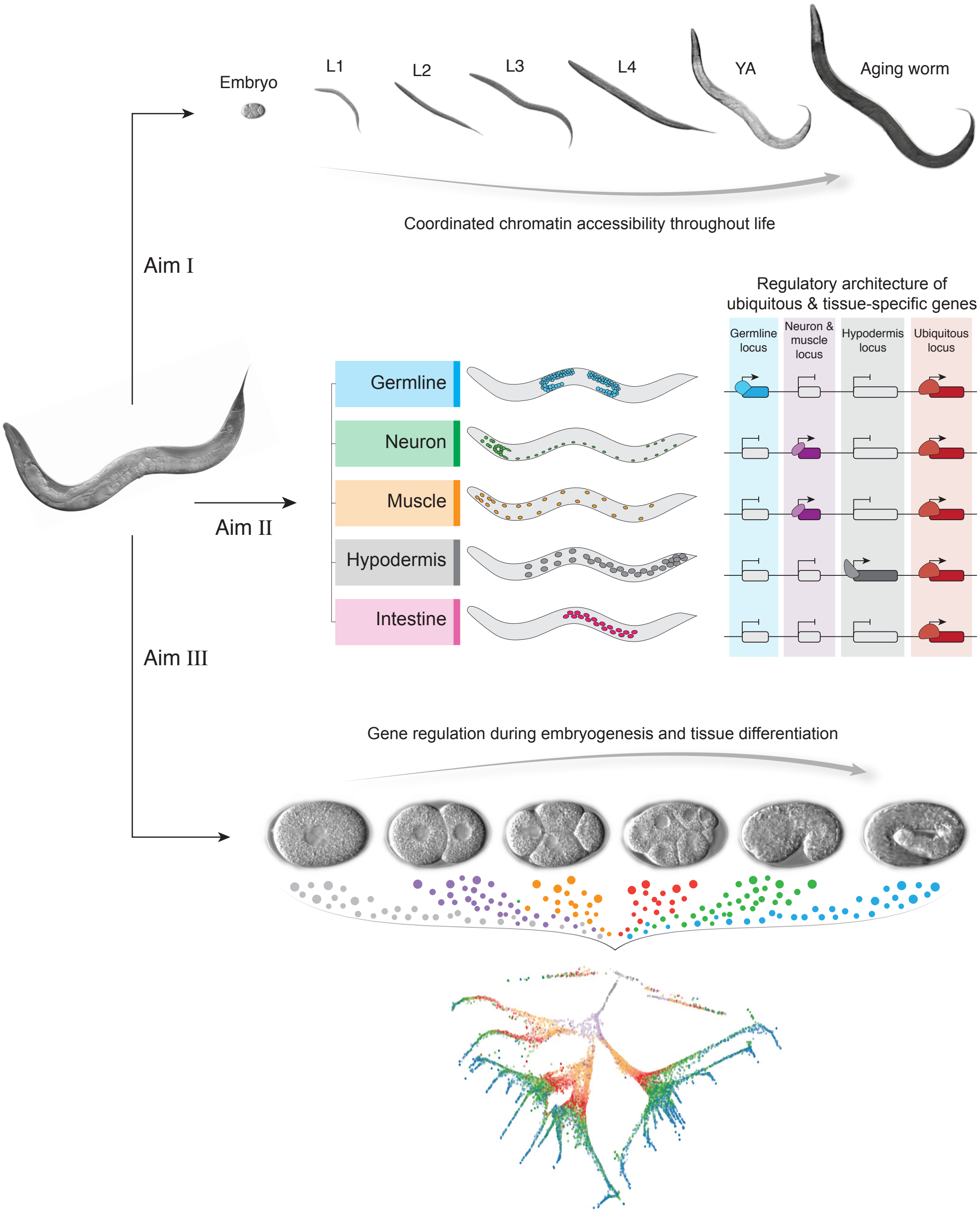

Embryo
L1
L2
L3
L4
YA
Aging worm
Coordinated chromatin accessibility throughout life
Aim I
Regulatory architecture of ubiquitous & tissue-specific genes
Germline locus
Neuron & muscle locus
Hypodermis locus
Ubiquitous locus
Germline
Neuron
Muscle
Hypodermis
Intestine
Aim II
Aim III
Gene regulation during embryogenesis and tissue differentiation

Figure 1.28 – Aims of my PhD. Each aim is further described in Section 1.4 (The schematic of Aim III is inspired from Farrell *et al.*, 2018).

Investigating the mechanisms of gene regulation during lineage specification and organogenesis

Finally, I sought to investigate how tissue-specific patterns of gene expression are acquired and regulated, specifically during lineage specification and organogenesis (Figure 1.28, Aim III, Chapter 7). For this, I focused on developing single-cell-based approaches to study regulatory element activity in individual nuclei during early embryogenesis and tissue differentiation.

Chapter 2

Methods

2.1 Experimental methods

Unless specified in the Foreword on page 1 or in relevant chapters, I performed all the experiments described in this work.

Generation of transgenic strains

C. elegans strains were maintained using standard procedures at 25 °C and fed OP50 *E. coli*. Targeting of the GFP to the nuclear envelope was achieved in two different ways: 1) by fusing a StrepTag (WSHPQFEK) to the N-terminal end of GFP (from pPD95.02, Fire lab Vector Kit) and UNC-83 (aa 1-290) to its C-terminal end, or 2) fusing the full-length NPP-9 coding sequence to the C-terminal end of GFP. The first approach was used to target GFP to the nuclear envelope in germline, muscle, hypodermis and intestine cells. The second approach was used to target GFP to the nuclear envelope in neurons. The promoter used to express the reporter in individual tissues are the *mex-5* promoter (for Germline expression, chrIV:13,353,242-13,353,729), the *egl-21* promoter (for Neuron expression, chrIV:10,481,768-10,481,932), the *myo-3* promoter (from Muscle expression, chrV:12,234,302-12,236,686), the *dpy-7* promoter (for Hypodermis expression, chrX:7,537,794-7,538,688) and the *npa-1* promoter (for Intestinal expression, chrV:7,075,526-7,075,947) (coordinates are in ce11). Three-way Gateway cloning was used to clone each tissue-specific promoter (in slot 1) upstream of the reporter coding sequence (in slot 2). *tbb-2*–3' UTR was used in slot 3 (Merritt *et al.*, 2008). The destination vector was pCFJ150 (Frøkjaer-Jensen *et al.*, 2008).

Reporter constructs were integrated in a single copy at the ttTi5605 Mos1 site located on chrII (Frøkjaer-Jensen *et al.*, 2008).

Worm collection

C. elegans strains were grown in liquid culture to the adult stage using standard S-basal medium with HB101 bacteria, animals bleached to obtain embryos, and the embryos hatched without food in M9 buffer for 24 hr at 25 °C to obtain synchronized starved L1 larvae. L1 larvae were grown in a further liquid culture at 25 °C for approximately 42h and young adult worms were then collected, washed in M9, floated on sucrose, washed again in M9, then frozen into "popcorn" by dripping embryo or worm slurry into liquid nitrogen.

Nuclear isolation

Young adult hermaphrodites were obtained by growing synchronized starved L1 larvae at 25 °C in standard S-basal medium with HB101 bacteria for 40-42h. After sucrose flotation and washing in M9 buffer, worms were frozen into "popcorn" by dripping concentrated slurry into liquid nitrogen. Nuclei were isolated as previously detailed (Jänes *et al.*, 2018), with minor modifications. ~ 20,000 to 200,000 frozen young adult worms were broken by smashing using a Biopulverizer then the frozen powder was thawed in 8 ml Egg buffer (25 mM HEPES pH 7.3, 118 mM NaCl, 48 mM KCl, 2 mM CaCl2, 2 mM MgCl2). Broken worms were pelleted by spinning at 800 g for 3 min then resuspended in 8 ml of Buffer A (0.3 M sucrose, 10 mM Tris pH 7.5, 10 mM MgCl2, 1 mM DTT, 0.5 mM spermidine 0.15 mM spermine, protease inhibitors (Roche complete, EDTA free) and 0.025 % IGEPAL CA-630). The sample was dounced (two strokes) in a 14-ml stainless steel tissue grinder (VWR) then spun at 100 g for 6 min to pellet remaining worm fragments. The supernatant was kept (nuclei batch 1) and the pellet resuspended in a further 7 ml of Buffer A and dounced for 30 strokes. This was spun at 100 g for 6 min to pellet debris and the supernatant was kept (nuclei batch 2). The first fraction was enriched for germline nuclei while the second fraction was enriched for somatic nuclei.

Nuclear sorting

Following isolation, nuclei were immunostained by adding phycoerythrin-coupled anti-GFP antibody (Biolegend # 338003) at 1:200 in 7 ml of buffer A, and 280 units of murine RNAse inhibitor (M0314S) were added to protect RNA from being degraded. Nuclei were kept slowly rotating at 4 °C in the dark for 1 to 16 hours. Debris was removed by spinning at 100 g for 6 min at 4 °C then nuclei were pelleted (2000 g for 20 min at 4 °C), washed in 6 ml of buffer A, and resuspended in buffer A containing 80 U/ml murine RNAse inhibitor at a concentration of ~ 10-15 million nuclei / ml. Finally, nuclei were filtered on 30 µm mesh (CellTrics 04-0042-2316) and stained with 0.025 µg/ml DAPI. Nuclei quality was assessed immediately before sorting by microscopy.

Nuclear sorting was performed at 4 °C using a Sony SH800Z sorter fitted with a 100 µm sorting chip and auto-calibrated. Nuclei were gated using the DAPI signal and PE-positive nuclei were gated using PE-H / BSC-A signal. DAPI gating depended on which nuclei were being sorted (e.g. intestine nuclei are 32N). A recording speed > 15,000 nuclei per second ensured a sorting efficiency higher than 80 %. Nuclei were sorted into 15 ml Falcon tubes containing 500 µl of buffer A with 800U/ml murine RNAse inhibitor. Nuclei were sorted in batches of one million and then processed for downstream applications. The purity and integrity of each batch of nuclei was assessed by recording an aliquot of sorted nuclei in a second pass in the sorter and by microscopy. All sorted samples used in this study had a purity higher than 95%.

ATAC-seq

One million nuclei were pelleted (2000 g for 20 min at 4 °C) and resuspended in 1X Tn5 Buffer (10 mM Tris pH 8, 5 mM MgCl2, 10% DMF) at a final concentration of ~ 500,000 nuclei / ml. 2.5 µl of Tn5 (Illumina FC-121-1030) were added to 47.5 µl (~ 25,000 nuclei) of the suspension. The mix was incubated for 30 min at 37 °C while mixing at 400 rpm. Tagmented DNA was purified using a MinElute column (Qiagen) and converted into a library using the Nextera kit protocol. Typically, libraries were amplified using 12-16 PCR cycles. Libraries were then cleaned up using 0.6 volumes of AMPure XP beads to remove large fragments of DNA (> 700 bp) and DNA recovered from the supernatant by adding 1.2 volumes of beads.

DNA was eluted in 50 µl water and the library was further size-selected using 0.9 volumes of beads to bind the library, leaving adaptor dimers in the supernatant.

ATAC-seq libraries were generated from two biological replicates for each tissue, and were sequenced in both single-end and paired-end modes. Single ATAC-seq libraries were made for L1 and L3 muscle (SE-sequenced) and L3 germline (PE-sequenced). PGC-specific ATAC-seq data at the L1 stage was obtained from (Lee *et al.*, 2017).

RNA-seq

Nuclear RNA was extracted from batches of one million sorted nuclei by pelleting nuclei (2000 g for 20 min at 4 °C), washing them in 1 ml of buffer A, then adding 500 µl of Trizol to pelleted nuclei. 100 µl of Chloroform were added and samples were shaken vigorously for 15 s. Samples were then spun at 12,000 g for 15 min at 4 °C and the aqueous phase was transferred into one volume of ice-cold Isopropanol. 0.5 µl of GlycoBlue (AM9515) was added and RNA was left to precipitate overnight at -20 °C. RNA was pelleted (12,000 g for 30 min at 4 °C), washed in 1 ml of ice-cold 80% Ethanol and resuspended in 12 µl of RNAse-free water. A minimum of 20 ng of total nuclear RNA was used to make long nuclear RNA-seq libraries. Long nuclear RNA (>200 nt) was isolated using Zymo Clean and Concentrate columns (#R1013) and stranded libraries were prepared with the NEB Next Ultra Directional RNA Library Prep Kit (#E7420S) after removal of rRNA using the Ribo-Zero rRNA removal kit (MRZH11124). Long nuclear RNA-seq libraries were made from two biological replicates for each tissue and were PE-sequenced.

Histone modification profiling

cChIP-seq was performed using chromatin from sorted neuron nuclei. Two million nuclei were sorted and resuspended in 200 µl FA buffer (50 mM HEPES-KOH pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% Na-DOC, 150 mM NaC). Chromatin was sheared to 100-300 bp-long fragments by sonicating the samples for 25 cycles (30” on / 30” off on Bioruptor Pico sonicator at high setting). Each sample was then topped up to 1 ml with FA buffer and 57.4 µl of Sarkosyl were added. 2.5 µg of antibody (H3K4me3: ab8580; H3K27me3: 309-95259, H3K36me3: ab9050) were added along with 60 ng of the corresponding recombinant histone and samples with antibody were incubated overnight at 4 °C slowly rotating. Dynabeads® Protein A (10001D) were also blocked in FA buffer complemented with 1% BSA and 10 µl tRNA. The next day, blocked beads were washed twice in FA buffer then the equivalent of 50 µl of Dynabeads® Protein A slurry was added to each sample and the mix was kept at 4 °C rotating for two hours. Subsequently, chromatin-bound beads were washed twice with FA buffer, once with FA 500 (same recipe than FA but with 500 mM NaCl), once with FA 1000, once with TEL (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA) and twice with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Samples were eluted twice with 150 µl elution buffer for 15 minutes at 65 °C. Samples were then treated with RNAse for 30 minutes at 37 °C, then with proteinase K at 55 °C for 1-2 hours then transferred to 65 °C overnight to reverse crosslinks. DNA was cleaned up using 1.8 volume of AMPure XP beads and eluted in 40 µl of lowTE (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA). Libraries were then made using the Accel-NGS® 2S Plus DNA Library Kit (21024) following manufacturer instructions.

CUT&RUN was performed using chromatin from sorted muscle nuclei. For each histone modification profiling, around 30,000 muscle nuclei were directly sorted into 500 µl of NE1 buffer (Skene and Henikoff, 2017). The rest of the steps were performed as described in Skene and Henikoff, 2017. Once eluted, DNA was transformed into high-throughput sequencing libraries using the procedure described in Jänes *et al.*, 2018.

Single-nucleus experiments

Single-nucleus assays were performed using the 10X Genomics Chromium workflow. Nuclei were obtained as follows: a “popcorn” of frozen early embryos was dropped into 1 ml of buffer A and quickly thawed. Embryos were then stained with DAPI and ran through a Sony SH800Z sorter fitted with a 100 µm sorting chip, and only the embryos with the 30% lowest DAPI signal were recovered. From there, nuclei were extracted using a Balch homogenizer fitted with a 18 µm ball, allowing for a 2 µm clearance. After ~ 8-10 strokes, nuclei were readily released from embryos into a high-quality single-nucleus suspension. Accurate nuclei concentration was estimated using a C-Chip disposable hemocytometer (DHC-N01). Single-nucleus emulsion and RNA-seq or ATAC-seq was performed by the Genomics facility at the CRUK Institute using 10X Genomics workflows.

2.2 Computational methods

Unless specified in the Foreword on page 1 or in relevant chapters, I performed all the computational analyses described in this work.

Data processing

Sequenced reads were trimmed using `fastx_trimmer 0.0.14` and aligned to the reference genome WBcel235/ce11 obtained from Ensembl release 92 (ftp://ftp.ensembl.org/pub/release-92/) using `bwa-backtrack 0.7.17-r1188` (Li and Durbin, 2009) in single-end (ATAC-seq) or paired-end mode (ATAC-seq, long nuclear RNA-seq). Low-quality ($q < 10$), mitochondrial and modENCODE-blacklisted (Dunham *et al.*, 2012) reads were discarded.

Normalized genome-wide accessibility tracks were computed with MACS2 (Feng *et al.*, 2012) using parameters `--format BAM --bdg --SPMR --gsize ce --nolambda --nomodel --extsize 150 --shift -75 --keep-dup all` and the bedGraphToBigWig utility (Kent *et al.*, 2010). ATAC-seq was also sequenced in paired-end mode; paired-end data were used for nucleosome occupancy and fragment density analysis.

Long nuclear RNA-seq data were processed essentially as in (Chen *et al.*, 2013). Following alignment and filtering, fragments-per-million-normalized strand-specific coverage tracks were computed by transforming the bam file into a bedGraph file using the `genomeCoverageBed v2.26.0` utility (Quinlan and Hall, 2010) with the parameters `-bg -pc -scale 10e6/${NBFRAGS} -strand ${STRAND}` (where `${NBFRAGS}` is the number of mapped fragments and `${STRAND}` is + or -). Gene annotations used throughout this study are WBcel235/ce11 obtained from Ensembl release 92 (ftp://ftp.ensembl.org/pub/release-92/).

To assess the reproducibility of biological replicate datasets, I used site accessibility or gene expression values to compute pairwise Euclidean distances between each dataset and pairwise Pearson correlation scores. The 20% least accessible sites and least expressed genes were not taken into consideration to compute the correlation scores. Both ATAC-seq and RNA-seq data from biological replicates showed high concordance.

Clustering of promoter accessibility

Accessible loci with regulated accessibility during development or aging were determined as follows. All loci (n = 42,245) were tested for a difference in ATAC-seq coverage between any two developmental time points or between any two aging time points using DESeq2 (Love *et al.*, 2014). Sites with >= 2 absolute fold change and adjusted p-value < 0.01 were defined as 'regulated' and were used in clustering analyses (n = 30,032 in development and n = 6,590 in aging); regulated promoters (n = 10,199 in development and n = 1,800 in aging).

Depth-normalized ATAC-seq coverage of each promoter was calculated at each time point in development or aging. For each accessible locus, the log2-mean-centered relative accessibility was calculated throughout development or aging. Clustering was performed using k-medoids as implemented in the `pam()` method of the `cluster` R package. Different numbers of clusters were tested for clustering of regulatory elements in developmental and aging datasets and assessed by their "silhouette"; 16 was chosen for developmental data and 10 for aging data as the normalized changes in promoter ATAC-seq signals within each cluster were relatively homogeneous. I manually merged two aging clusters showing comparable accessibility and tissue-specific gene enrichment (resulting in the cluster I + H [2]). Clusters labels were determined based on enrichment for tissue-biased gene expression within each cluster (see below).

To compare accessibility and gene expression, FPM-normalized gene-level read counts were calculated using DESeq2, and then averaged across biological replicates.

Using single-cell RNA-seq data from Cao *et al.*, 2017, I defined tissue-biased gene expression as follows: Gene expression was considered enriched in a given tissue if it had a fold-change >= 3 between expression in the tissues with highest and second highest levels and an adjusted p-value < 0.01. This defined 5,315 genes with tissue-biased expression (1,432 in Gonad, 553 in Hypodermis, 799 in Intestine, 352 in Muscle, 1,218 in Neurons, 447 enriched in Glia, 514 in Pharynx). For each developmental or aging cluster of promoters, I calculated the percentage of genes with biased expression in a given tissue relative to the total number of genes in the cluster.

Integration of transcription factor binding profiles and chromatin dynamics

modENCODE and modERN transcription factor binding datasets used in this paper were obtained from http://www.encodeproject.org or http://data.modencode.org (EOR-1) (Araya *et al.*, 2014; Kudron *et al.*, 2018). ChIP-seq profiles were manually inspected and 227 high quality datasets selected, covering 176 transcription factors. To analyze enrichment of individual factors, TF peaks were assigned to a regulatory element if their summits overlapped with the 400 bp region centered at the element midpoint. I excluded binding at so-called 'HOT' (highly occupied target) regions from enrichment analyses, as these are thought to represent non-sequence-specific TF binding or ChIP artifacts (Gerstein *et al.*, 2014; Kudron *et al.*, 2018). HOT regions were defined here as accessible sites with binding of 19 or more of the analyzed 176 TFs (sites in the top 20% of binding, excluding sites with no binding). Only transcription factors with more than 200 peaks overlapping 'non-HOT' regulatory elements were kept, to ensure sufficient data for analysis. Following this stringent filtering, 89 transcription factors could be assayed for binding enrichment. Transcription factor binding enrichment in each cluster was estimated using the odds ratio and the transcription factors with an enrichment higher than 2 in at least one cluster and an associated p-value < 0.01 were kept (Fisher's exact test).

Temporal activity of transcription factors during development and aging

Temporal activity of transcription factors during development or aging is estimated as follows: for each factor, its binding enrichment across all the promoter clusters was multiplied by the average promoter accessibility in corresponding clusters, separately during development and aging. The resulting matrix product was then row-scaled.

Annotation of new regulatory elements

In a previous study, 42,245 accessible sites were identified across development and aging and annotated them into functional classes (coding promoters, non-coding promoters, unassigned promoters, putative enhancers, inactive elements) based on nuclear RNA-seq patterns (Jänes *et al.*, 2018). Jürgen Jänes ran his annotation pipeline using the previously generated data together with the newly generated adult tissue-specific ATAC-seq and RNA-seq generated in this study. This resulted in the detection and annotation of 5,269 new accessible sites, bringing the total sites to 47,514.

GO term enrichment analyses

GO term enrichment analyses were performed using the `gProfileR 0.6.7` package (Reimand *et al.*, 2007), filtering for redundant GO terms using the `hier_filtering = moderate` option. To compare GO enrichment across several groups, the `clusterProfiler 3.10.1` package (Yu *et al.*, 2012) was used, filtering for redundant terms using REVIGO (http://revigo.irb.hr/). Only GO terms with Bonferroni-adjusted p-values lower than 0.05 were kept.

Comparison of gene annotation with published gene sets

I compared the main classes of tissue-specific and ubiquitous genes obtained in this study with previously published gene sets from Cao *et al.*, 2017, Spencer *et al.*, 2011 and Kaletsky *et al.*, 2018 (Figure 5.9 on page 111). None of these studies formally defined a class of ubiquitous genes.

Annotation of DREAM targets

DREAM targets were defined as the set of genes regulated by DREAM complex identified in Latorre *et al.*, 2015 (Figure 5.16 on page 119).

Distribution in chromatin domains

Distribution of genes and accessible sites over active or regulated domains, borders and chromosome X (Figure 5.19 on page 123) was obtained using the chromatin domain annotation from Evans *et al.*, 2016.

Physical chromatin interactions and networks of inferred interactions in individual tissues of young adult worms

Chromatin interactions mapped using ARC-C in whole worms at the L3 stage were obtained from Huang *et al.*, 2018. I discarded 8,547 interactions mapped in L3 anchored at at least one accessible site classified as "Unclassified" or "Low" in young adults, and 589 other interactions anchored at two regulatory elements not active in the same tissue(s) in young adult (*e.g.* one germline and one neuron regulatory element). I then deconvoluted the remaining 24,152 interactions based on the activity of the two regulatory elements to which each interaction is anchored. For instance, an interaction anchored to a ubiquitous promoter on one side and to a promoter active in both neurons and muscles on the other side would be assigned to the two sets of neuron and muscle interactions. This resulted in five sets of inferred interactions in the five main tissues of the YA worm.

Networks of inferred interactions in individual tissues were generated and investigated using `igraph 1.2.4` package (Csardi and Nepusz, 2006). Simulated networks of inferred interactions in individual tissues were generated by first deleting specific interactions (*e.g.* interactions bridging loci located further than X bp from each other) or specific loci (*e.g.* sites accessible across all four somatic tissues) then re-computing the resulting networks. Interaction frequency, communities and modularity were determined using `igraph 1.2.4` package.

ATAC-seq fragment density plots

ATAC-seq fragment density plots, also known as V-plots (Henikoff *et al.*, 2011), were generated using the `VplotR 0.4.0` package (Serizay, 2020b).

Nucleosome occupancy tracks and putative +1 nucleosome mapping

Processed bam files from paired-end ATAC-seq duplicates of each tissue or from whole organism young adults (Jänes *et al.*, 2018) were merged. For each class of promoter (germline, neuron, muscle, hypodermis, intestine and ubiquitous promoters), the nucleoATAC python package (Schep *et al.*, 2015) was used to compute the probability of nucleosome occupancy from -1kb to + 1kb from promoter centers in each tissue (germline, neuron, muscle, hypodermis, intestine and whole organism). Promoter centers were defined by the summits of peaks in chromatin accessibility signal.

Putative +1 and -1 nucleosome positions were determined for each set of tissue-specific promoters using the corresponding tissue-specific nucleosome occupancy probability track and for ubiquitous promoters using whole organism nucleosome occupancy probability track (Jänes *et al.*, 2018). I assigned the center of the putative +1 nucleosome to the local maximum of the nucleosome occupancy probability within 200 bp downstream from the forward TSS mode (TSSs were annotated using short capped nuclear RNA-seq, see Chen *et al.*, 2013; Jänes *et al.*, 2018). Similarly, the center of the -1 nucleosome summit was assigned to the local maximum of the occupancy probability within 200 bp upstream of the reverse TSS mode. Only coding promoters with experimentally determined forward and reverse TSSs were considered.

Motif identification and enrichment analyses

Motifs enriched in different sets of promoters (-75 bp to +105 bp from promoter centers) were identified using MEME in stranded mode and a 0-order background model (`-markov_order 0`). MEME mode was set to 'Any Number of Repetitions' (`-mod anr`) and motif widths were restricted to 6 to 25 bp. The five motifs found most enriched (with an E-value threshold of 0.05) were retrieved. Unstranded motifs (found twice as complementary sequences, since MEME was run in stranded mode) were manually combined. PWMs for the Initiator (Inr) and the TATA motif were obtained from Jin *et al.*, 2006. Motif mapping to promoters was performed in R using the `Biostrings 2.50.2` package, the `GenomicRanges 1.34.0` package and the `TFBSTools 1.20.0` package with a relScore threshold set to 0.8.

Dinucleotide periodicity

To estimate dinucleotide periodicity in sets of sequences (e.g. -50 to +300 bp sequences around ubiquitous, germline or somatic-tissue-specific TSSs in Figure 6.9 on page 147, or -50 to +300 bp sequences around TSSs from different organisms in Figure 6.15 on page 154), the `getPeriodicity()` function from the `periodicDNA 0.2.0` package was used with default parameters (Serizay, 2020a). Briefly, the distribution of distances between all possible pairs of dinucleotides in the set of sequences was computed and corrected for distance decay, smoothed by a moving average window of 3 and power spectral densities were retrieved by applying a Fast Fourier Transform to the normalized distribution.

To generate 10-bp dinucleotide periodicity score tracks, the `generatePeriodicityTrack()` function from the `periodicDNA 0.2.0` package was used with default parameters (Serizay, 2020a). Briefly, a running 10-bp dinucleotide periodicity score was calculated by applying a Fast Fourier Transform (`stats 3.5.2` package) on the distribution of distances between pairs of dinucleotides (e.g. WW......WW) found in 100-bp long sequences (2-bp increments).

Phasing of nucleosomal sequences

To observe the 10-bp periodic occurrence of a dinucleotide in putative +1 nucleosomes, sequences (400 bp centered at the nucleosome dyads) were first clustered by k-means based on the dinucleotide occurrences in each sequence, then the clusters were rephased within a -/+5 bp range using the lag value estimated by the `ccf()` function from the `stats 3.5.2` package.

Sets of annotations in fly, fish, mouse and human

In worms, experimentally annotated TSSs were used (Jänes *et al.*, 2018). In fly and zebrafish (respectively dm6 and danRer10 genome versions), TSSs were assigned to the first base of the genes using `TxDb.Dmelanogaster.UCSC.dm6.ensGene 3.4.4` and `TxDb.Drerio.UCSC.danRer10.refGene 3.4.4` gene models with the `GenomicFeatures 1.34.7` package in R. In mouse and human, FANTOM5 CAGE datasets were used to retrieve the dominant TSS closest to the gene annotation (Lizio *et al.*, 2015).

Nucleosome occupancy in fly, fish, mouse and human

Nucleosome occupancy tracks were generated as described for worms using `nucleoATAC` with the following ATAC-seq datasets: SRR6171265 in fly (Haines and Eisen, 2018), SRR5398228 in zebrafish (Quillien *et al.*, 2017), SRR5470874 in mouse (Benchetrit *et al.*, 2019) and SRR891268 in human (Buenrostro *et al.*, 2013).

Coefficient of variation of gene expression

Coefficient of variation of gene expression (CV) values were retrieved from Gerstein *et al.*, 2014 for worm, fly and human or computed using gene expression datasets from Pervouchine *et al.*, 2015 for mouse and White *et al.*, 2017 for zebrafish. Genes with the 20% lowest CVs were considered broadly expressed and those with the 20% highest CVs were considered regulated.

Other visualization tools

Figures were generated in `R 3.5.2`, using either base or `ggplot2 3.1.1` plotting functions. Genome browser screenshots were obtained from `IGV 2.4.8`. Genome tracks in the bigWig format were imported in R using the `rtracklayer 1.42.2` package.

Chapter 3

Coordinated regulatory element activity during *C. elegans* development and aging

When I started my PhD, Yan Dong and Michael Schoof were profiling chromatin accessibility and gene expression by ATAC-seq and RNA-seq, throughout *C. elegans* development and aging. In parallel, Jürgen Jänes was leading the computational analyses. I joined this project to investigate the dynamics of regulatory element accessibility during *C. elegans* development and aging. In Chapter 3, I present the characteristics of sets of regulatory elements which have coordinated changes in accessibility during *C. elegans* life. Most of the results presented here have been published in Jänes *et al.* (2018) (see Appendix Chapter B).

Collaboration note: Yan Dong and Michael Schoof generated the ATAC-seq and RNA-seq libraries used in this chapter and Jürgen Jänes developed the bioinformatic pipeline to map and annotate regulatory elements from these datasets.

3.1 Chromatin accessibility dynamics during development and aging

3.1.1 Annotation of regulatory elements in *C. elegans*

Regulation of transcription during development is a key process to achieve spatiotemporal patterns of gene expression. Promoters and enhancers play a central role in this process, either by initiating or by modulating transcription, and can be under spatial and/or temporal control. Thus, an essential step for understanding the transcriptional circuits that control development and physiology is the genome-wide identification and characterization of regulatory elements (see Chapter 1.2 and Chapter 1.3). Still, no study had yet investigated regulatory element usage across the life of an animal, from the embryo to the end of life. *Caenorhabditis elegans* is an ideal model organism to map regulatory elements and study their dynamics through development and aging, as it has a simple anatomy, well-defined cell types, and short development and lifespan.

At the beginning of my PhD, I joined a project to identify and annotate regulatory elements used during *Caenorhabditis elegans* development and aging by integrative analysis of different types of high-throughput sequencing datasets (Jänes *et al.*, 2018). ATAC-seq datasets generated from embryo to aging worms were used to define 42,245 genomic loci accessible at any point during the worm life cycle. Nuclear RNA-seq data were then used to functionally annotate the accessible sites. This identified 15,678 putative promoters (13,596 of them associated with protein-coding genes) and 19,231 putative enhancers. 824 accessible sites were found to overlap small ncRNAs and 6,512 other accessible sites did not have any transcriptional activity and remained uncharacterized.

3.1.2 Sets of coordinated promoters regulate gene expression during development

I sought to investigate changes in the accessibility of regulatory elements across *C. elegans* life. I first focused on chromatin accessibility dynamics between embryonic development and adulthood. 71% of the annotated accessible sites showed a significant difference in accessibility within this period of time. To investigate how accessibility relates to gene expression, I then focused on the 13,596 regulatory elements annotated as promoters of protein-coding genes; 10,199 of these promoters (75%) showed significant changes in accessibility in development. Using a k-medoids clustering approach, I grouped them into sixteen sets of promoters characterized by similar temporal variations of accessibility (Figure 3.1). The remaining 3,397 promoters show stable accessibility. Importantly, I observed that within each cluster, promoter accessibility and nuclear RNA levels of the associated genes are

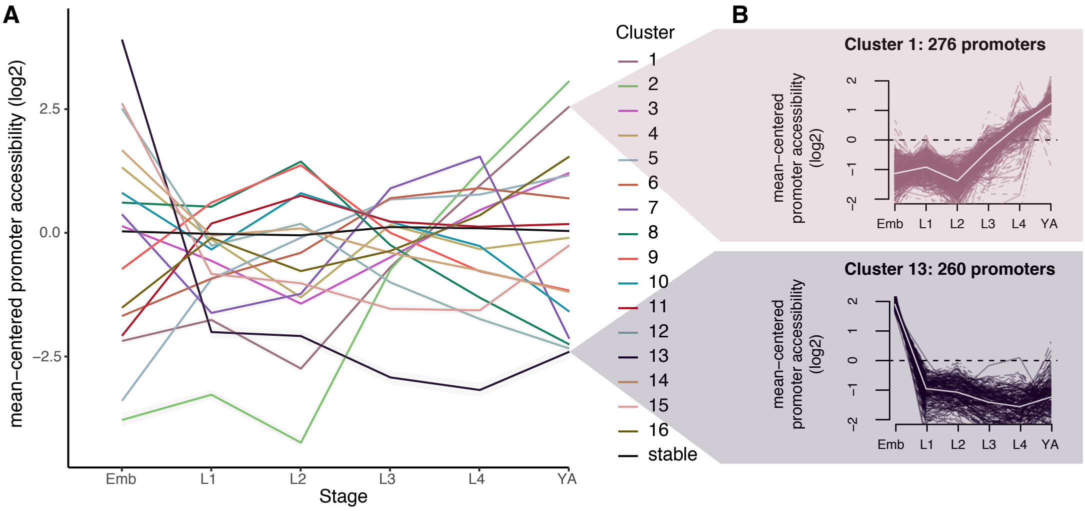


FIGURE 3.1 – *k*-medoids based clustering of promoter accessibility changes during development. **A-** Mean-centered promoter accessibility in each cluster during development. Solid lines represent the average value of chromatin accessibility and grayed ribbons represent the confidence intervals. **B-** Example of promoter accessibility changes in cluster 1 and cluster 13.

relatively well correlated (mean r = 0.47, sd = 0.11 across all clusters), indicating that indeed accessibility is a good metric of promoter activity and overall gene expression (Figure 3.2). These results suggest that most promoters are dynamically accessible during development, and that subsets of promoters with coordinated accessibility regulate temporal expression of their associated genes.

I hypothesized that some of the sets of promoters with coordinated accessibility could be associated with genes involved in tissues-specific processes. I took advantage of a single-cell dataset measuring levels of gene expression in different tissues in L2 stage to annotate each cluster based on its association with tissue-specific genes (Figure 3.3, Figure 3.4 and Cao *et al.*, 2017). I found that eight clusters of coordinated promoters are associated with genes specifically expressed in one or two tissues (four gonad-related promoter clusters (G1-G4), two intestine-related promoter clusters (I1, I2), one hypodermis-related promoter cluster (H) and one promoter cluster associated with genes expressed in neurons and muscle (N + M)). Interestingly, promoters associated with genes with the same tissue-specificity can exhibit similar trends of chromatin accessibility variations but with different amplitudes. For instance, G1 and G2 gonad-related promoter clusters are both characterized by an increase of chromatin accessibility starting in L3 (when germline starts proliferating); however, the amplitude of this increase is 1.5-fold greater in G2 than in G1 (Figure 3.3). This is consistent with higher levels of

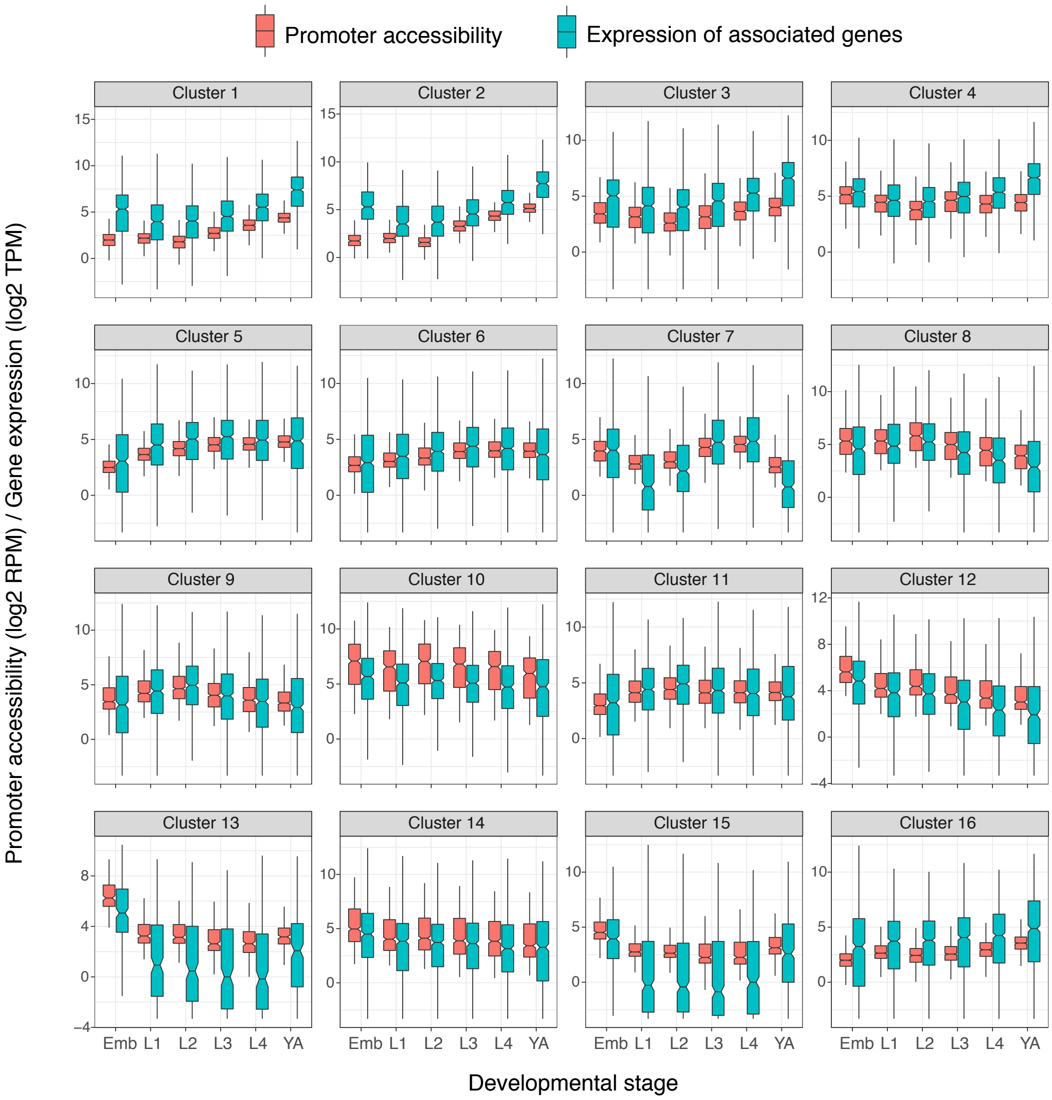


Figure 3.2 – Promoter accessibility and associated gene expression in developmental clusters of coordinated promoters. Note the correlation between changes in promoter accessibility and changes in expression of the associated genes.

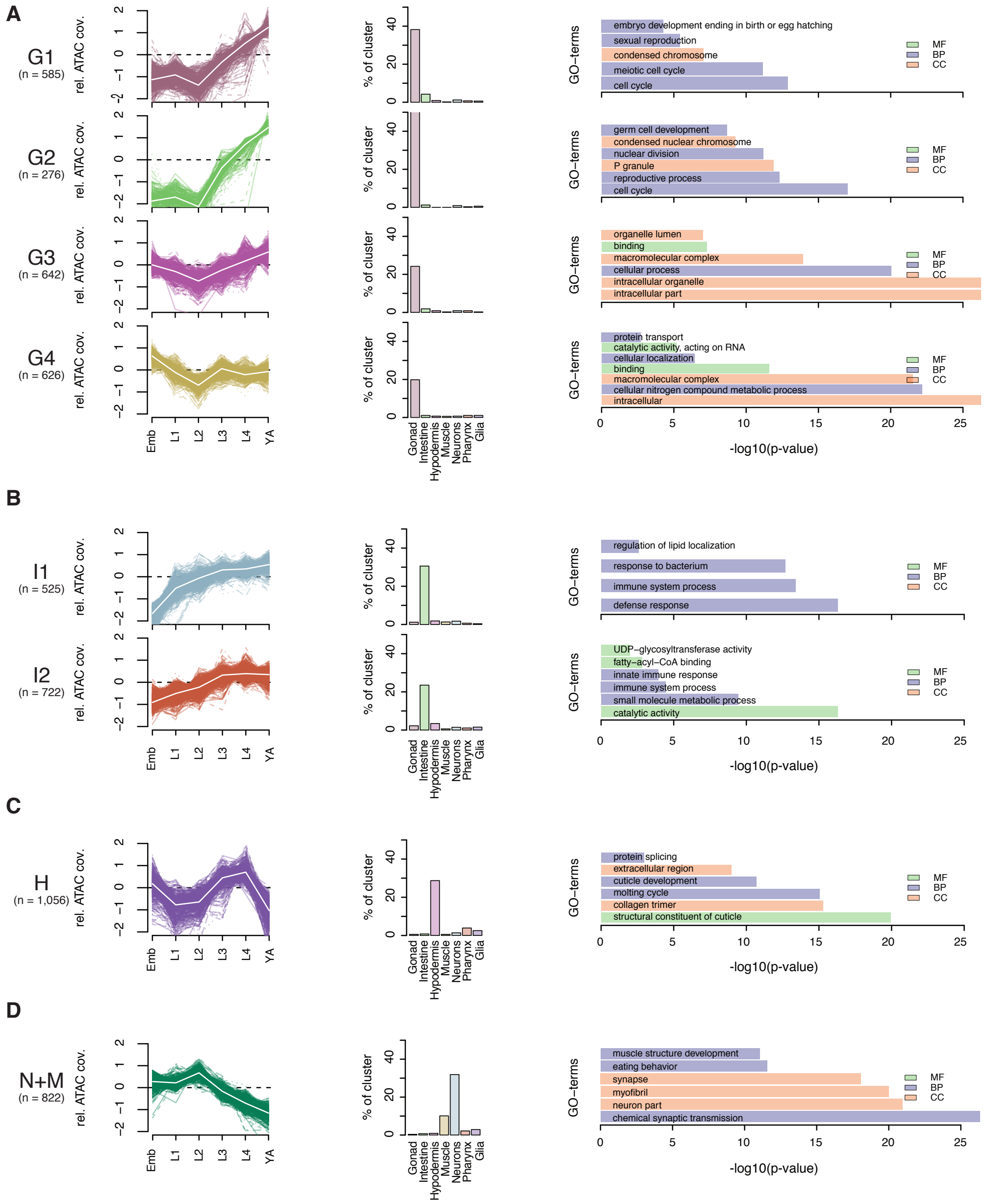


Figure 3.3 – Association between promoter clusters and biological processes (1). The left column represents promoter accessibility changes during development. The middle column represent the % of genes from each cluster specifically expressed in a given tissue in L2 (data from Cao *et al.*, 2017). The right column shows the main GO terms enriched in each cluster (colors indicate the type of GO term; blue: Biological Process; green: Molecular Function; orange: Cellular Component). Clusters are grouped by the tissue in which the associated genes are generally expressed: **A-** Gonad-related clusters, **B-** Intestine-related clusters, **C-** Hypodermis-related cluster and **D-** Neurons and muscle-related cluster. Only the eight promoter clusters with tissue-specific-related functions are shown here (the other eight clusters are shown in Figure 3.4)

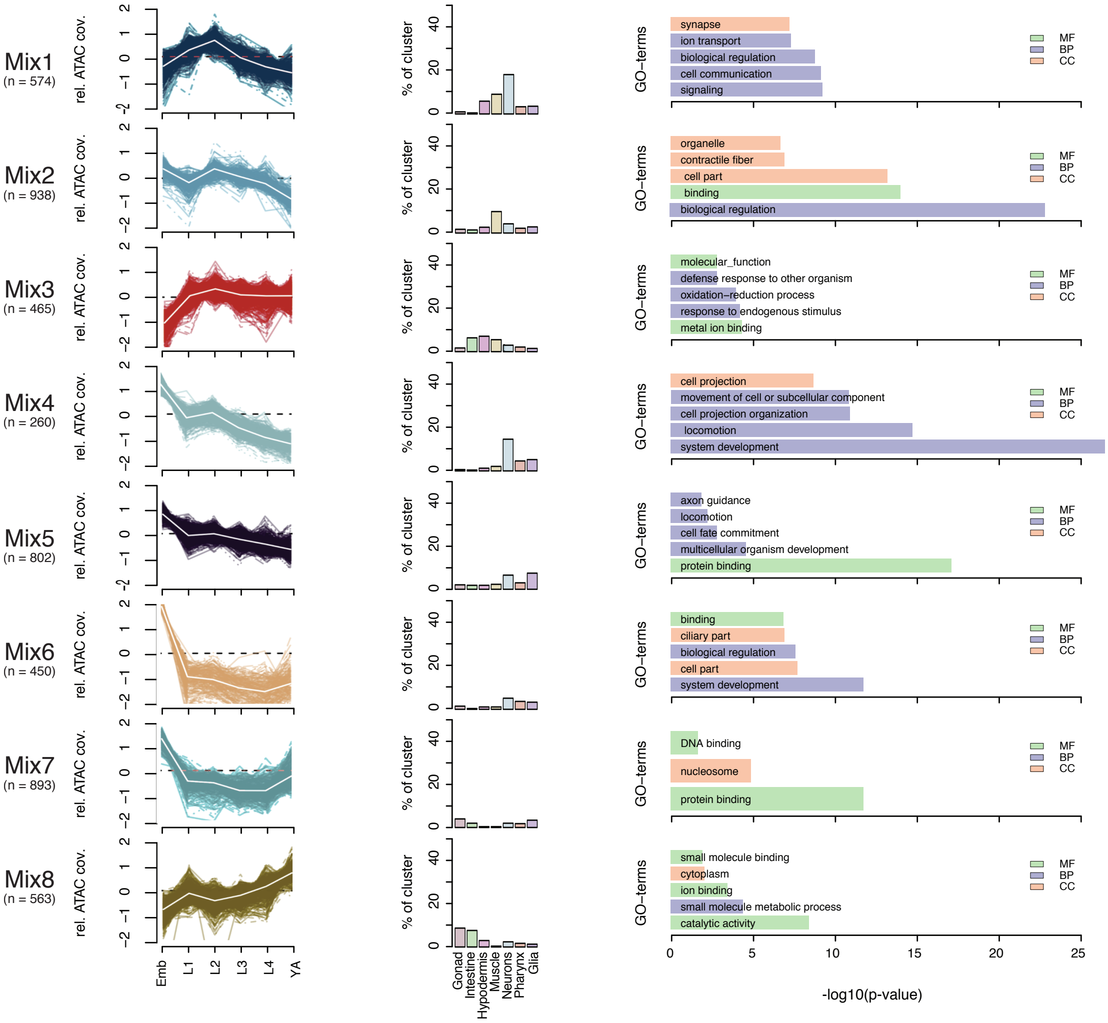


Figure 3.4 – Association between promoter clusters and biological processes (2). The eight clusters of promoters with no clear tissue-specific-related functions are represented here. Legends are the same than in Figure 3.3.

expression for the genes associated with a promoter from the G2 cluster compared to those associated with a promoter from the G1 cluster (Figure 3.2). Overall, the coordinated promoters in these eight clusters are regulating the temporal expression of tissue-specific genes. In contrast, genes associated with the remaining eight promoter clusters (Mix1–8) are generally not tissue-specific genes, but still have a variable expression during development (Figure 3.1 and Figure 3.4). Thus, the coordinated promoters in the "Mix" clusters are regulating the temporal expression of genes expressed more broadly across tissues.

I performed Gene Ontology term enrichment analyses on the genes associated with each set of promoters to investigate the functional role of coordinated promoters (Figure 3.3). As expected, I found that promoter clusters associated with tissue-specific genes are enriched for GO terms related to that tissue. For instance, the cluster H contains promoters associated with genes highly expressed in hypodermis and GO terms linked them to cuticle development. Interestingly, promoter clusters regulating similar tissue-specific groups of genes can be associated with different biological functions. For example, the I1 cluster contains promoters associated with intestinal genes involved in organismal defense against pathogens, while the I2 cluster contains promoters associated with intestinal genes generally involved in metabolism. "Mix" promoter clusters associated with broadly expressed genes are also enriched for specific biological functions. For instance, the cluster Mix4 is associated with genes expressed during embryogenesis and early larval stages and involved in cell migration (Figure 3.4).

Taken together, these results suggest that sets of promoters with coordinated activity during development temporally regulate functional sets of tissue-specific or broadly expressed genes.

3.1.3 Sets of coordinated promoters regulate gene expression during aging

Following the same approach, I also characterized chromatin accessibility changes during aging. In contrast to the development time course, only 1,800 of the 13,596 promoters (13%) show significant changes in accessibility. As for the development time course, I clustered these promoters according to their changes in accessibility and identified eight clusters of promoters with coordinated chromatin

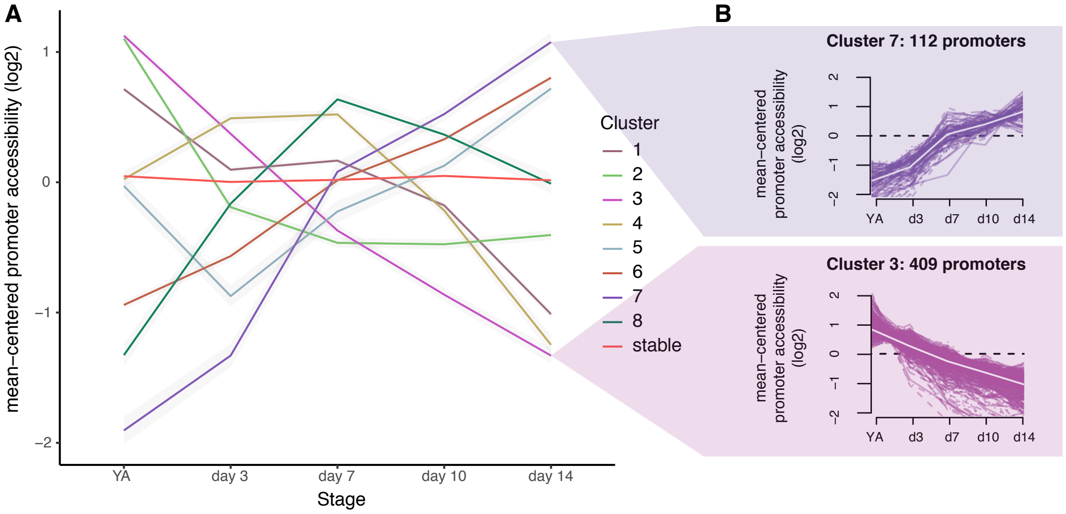


FIGURE 3.5 – *k*-medoids based clustering of promoter accessibility changes during aging. **A-** Mean-centered promoter accessibility in each cluster during aging. Solid lines represent the average value of chromatin accessibility and grayed ribbons represent the confidence intervals. **B-** Example of promoter accessibility changes in cluster 3 and cluster 7. YA: Young adults; d3: day 3 of life (*i.e.* one day older than YA); d7: day 7 of life; d10: day 10 of life; d14: day 14 of life.

accessibility during aging. I annotated these promoter clusters based on their association with tissue-specific genes (Figure 3.5 and Figure 3.6). This identified one cluster of intestine-related promoters (I), two clusters of promoters associated with genes enriched in intestine or hypodermis (I + H) and five other clusters (Mix1-5). Interestingly, most of the promoters that changed accessibility during aging underwent a decrease of accessibility (Figure 3.6). This is in line with recent reports suggesting focal heterochromatinization in aging nuclei (Sen *et al.*, 2016).

These results show that many of the promoters undergoing accessibility changes during aging are associated with intestine and hypodermis processes, supporting a central role of these tissues in aging and lifespan (Gelino *et al.*, 2016; Herndon *et al.*, 2002; McGee *et al.*, 2011).

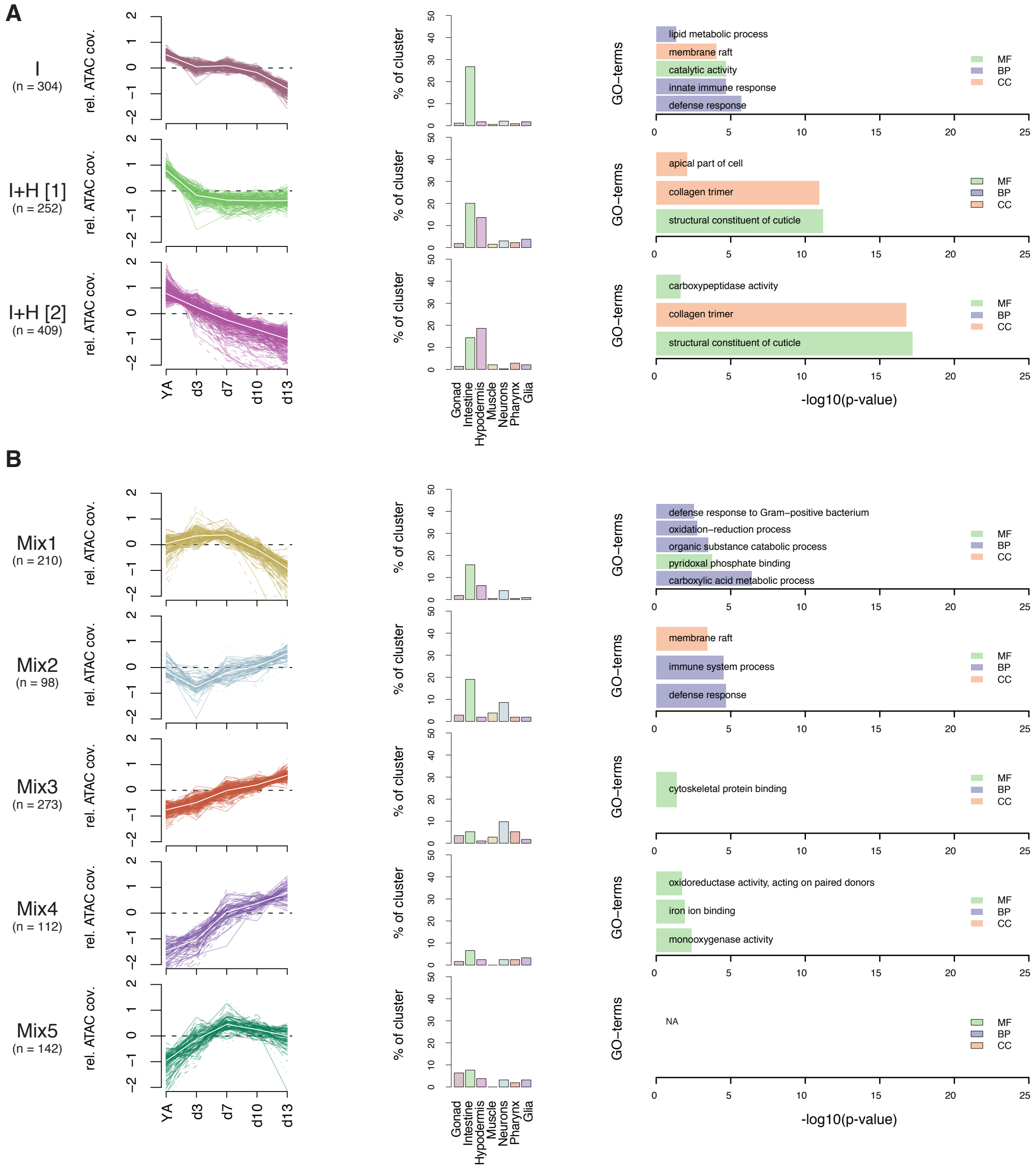


Figure 3.6 – Association between aging promoter clusters and biological processes. The left column represents promoter accessibility changes during aging. The middle column represent the % of genes from each cluster specifically expressed in a given tissue in L2 (data from Cao *et al.*, 2017). The right column shows the main GO terms enriched in each cluster (colors indicate the type of GO term; blue: Biological Process; green: Molecular Function; orange: Cellular Component). Clusters are grouped by the tissue in which the associated genes are generally expressed: **A-** Intestine or Intestine and Hypodermis-related clusters, **B-** Promoters clusters with no clear tissue-specific-related functions.

3.2 Transcription factors coordinating regulatory element accessibility

3.2.1 Transcription factors associated with developmental promoter clusters

The sets of developmentally coordinated promoters I identified are associated with genes with shared biological functions. I hypothesized that transcription factors could bind to specific sets of promoters to coordinate the expression of their associated genes across developmental stages, as seen in Figure 3.2. I obtained genome-wide transcription factor binding profiles from the modENCODE and modERN databases (Kudron *et al.*, 2018; ModENCODE, 2011) and computed the odds ratio of TF binding in each promoter cluster (Figure 3.7). In many cases, transcription factors were specifically associated with one or several individual clusters of promoters. For example, the intestinal transcription factor ELT-2 (Fukushige *et al.*, 1998) is enriched at promoters of intestinal clusters I1 and I2. Similarly, the hypodermal transcription factors BLMP-1 (Horn *et al.*, 2014), NHR-25 (Gissendanner and Sluder, 2000) and ELT-3 (Gilleard *et al.*, 1999) bind to promoters of the hypodermal cluster H and the germline XND-1 factor (Wagner *et al.*, 2010) binds to promoters of germline clusters G1 to G4 (Figure 3.7).

I aimed to characterize the activity of transcription factors throughout development. The levels of expression of a given transcription factor do not necessarily recapitulate the TF binding activity. For instance, a transcription factor can be expressed but restricted to cytoplasm, or would require the presence of a co-factor to bind to DNA. I sought to leverage (i) the TF binding profile at promoters and (ii) the accessibility dynamics of the promoters to provide an alternative estimation of the regulatory activity of a transcription factor, based on its binding patterns rather than on its expression. For a given transcription factor (*e.g.* XND-1) at a given time point (*e.g.* L1 stage), I defined its temporal activity $\alpha_{XND-1,L1}$ as the sum of the products between the TF binding enrichment ($Enr_{XND-1,clust.i}$, defined as the percentage of promoters bound by the factor within a cluster i) and the average promoter accessibility in L1 ($Acc_{,clust.i,L1}$, defined as the average promoter accessibility in cluster i), for each of the annotated clusters.

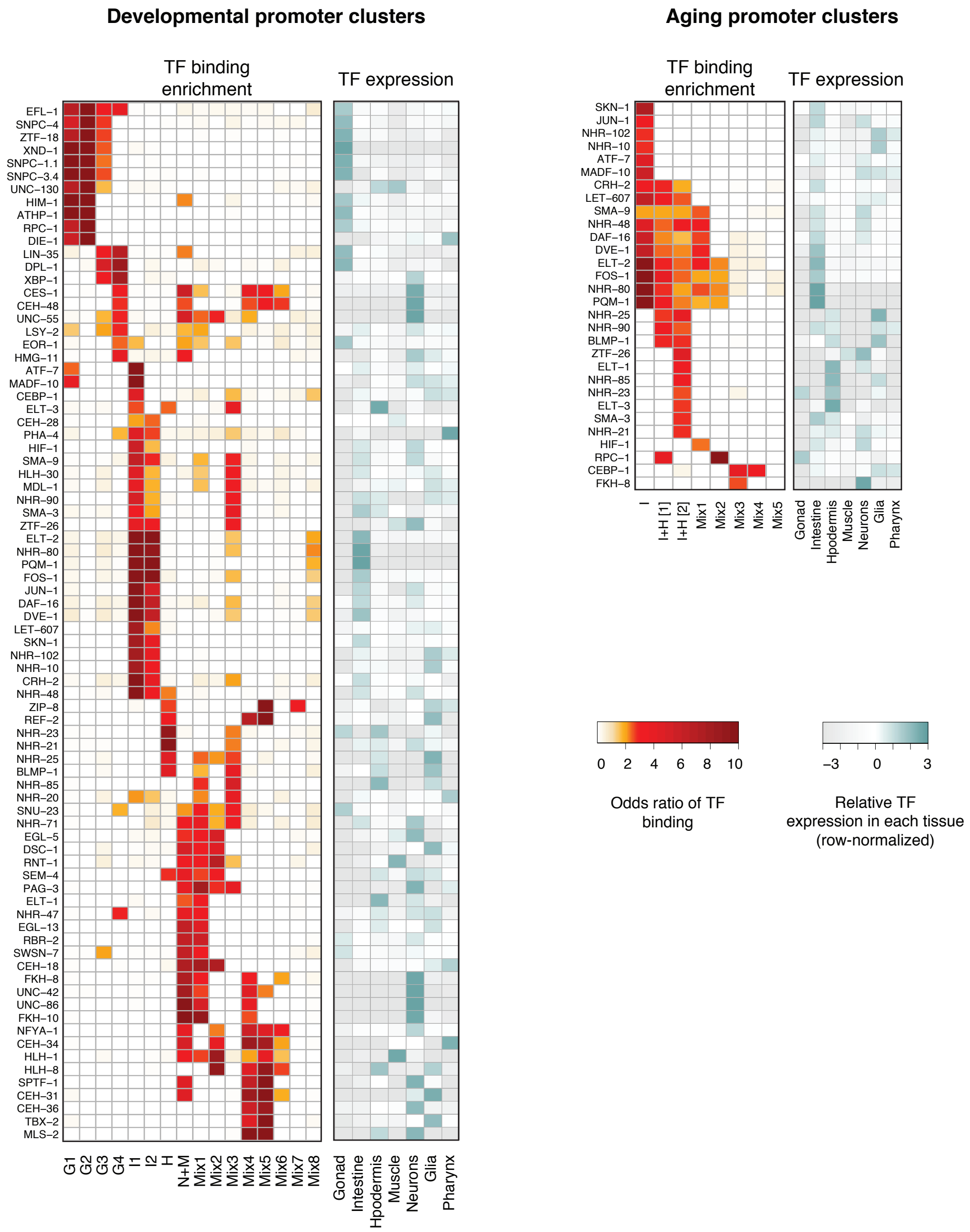


Figure 3.7 – Odds ratio of transcription factor binding in individual developmental and aging clusters. Transcription factor ChIP-seq datasets have been obtained from modENCODE and modERN projects (Kudron *et al.*, 2018; ModENCODE, 2011). Relative transcription factor gene expression in individual tissues is also represented (using data from Cao *et al.*, 2017). Note the correspondence between Transcription factor binding enriched in clusters of tissue-related promoters and their pattern of tissue-specific expression (*e.g.* XND-1 is enriched at promoters of G1, G2, G3 and G4 gonad-related clusters and is also expressed specifically in Gonad).

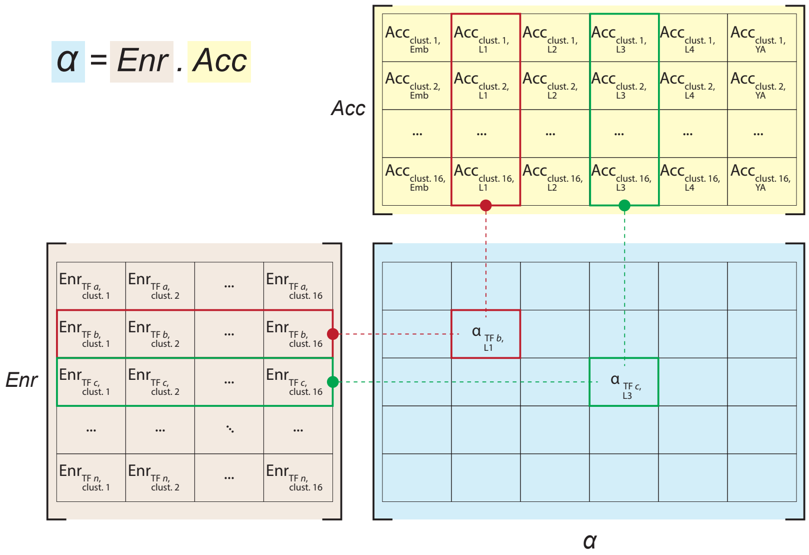


FIGURE 3.8 – Temporal activity of transcription factors defined as the matrix product between the matrix of binding enrichment (for each transcription factor in each cluster) and the matrix of average accessibility (for each promoter cluster at each developmental timepoint)

$$\alpha_{XND-1,L1} = \sum_{i=1}^{16} \left(Enr_{XND-1,cluster.i} \times Acc_{cluster.i,L1} \right)$$

Thus, the temporal activity of any transcription factor at any given time during development is defined by the matrix product $\alpha = Enr \cdot Acc$ (Figure 3.8). When z-scored by rows, the resulting matrix of temporal activity scores indicates to which extent a factor binds to accessible promoters at each developmental stage, for each transcription factor (Figure 3.9). Thus, the temporal activity metric aims to describe more directly the binding patterns of a transcription factor over time, compared to using its levels of expression.

This temporal activity metric shows expected results, such as the germline-specific transcription factors being mostly active at the young adult stage (*e.g.* HIM-1, XND-1) or transcription factors involved in cell fate specification mostly active at the embryo stage (*e.g.* SPTF-1). Interestingly, unusual patterns of transcription factor temporal activity can also be detected. For example, NHR-25 is a transcription factor involved in the patterning of the hypodermis during embryogenesis, but also regulates a network of genes involved in larva-to-adult transition (Chen *et al.*, 2004; Hada *et al.*, 2010). The activity of NHR-25 at these two different strategic time points of *C. elegans* development is clearly detected in its temporal activity (Figure 3.9).

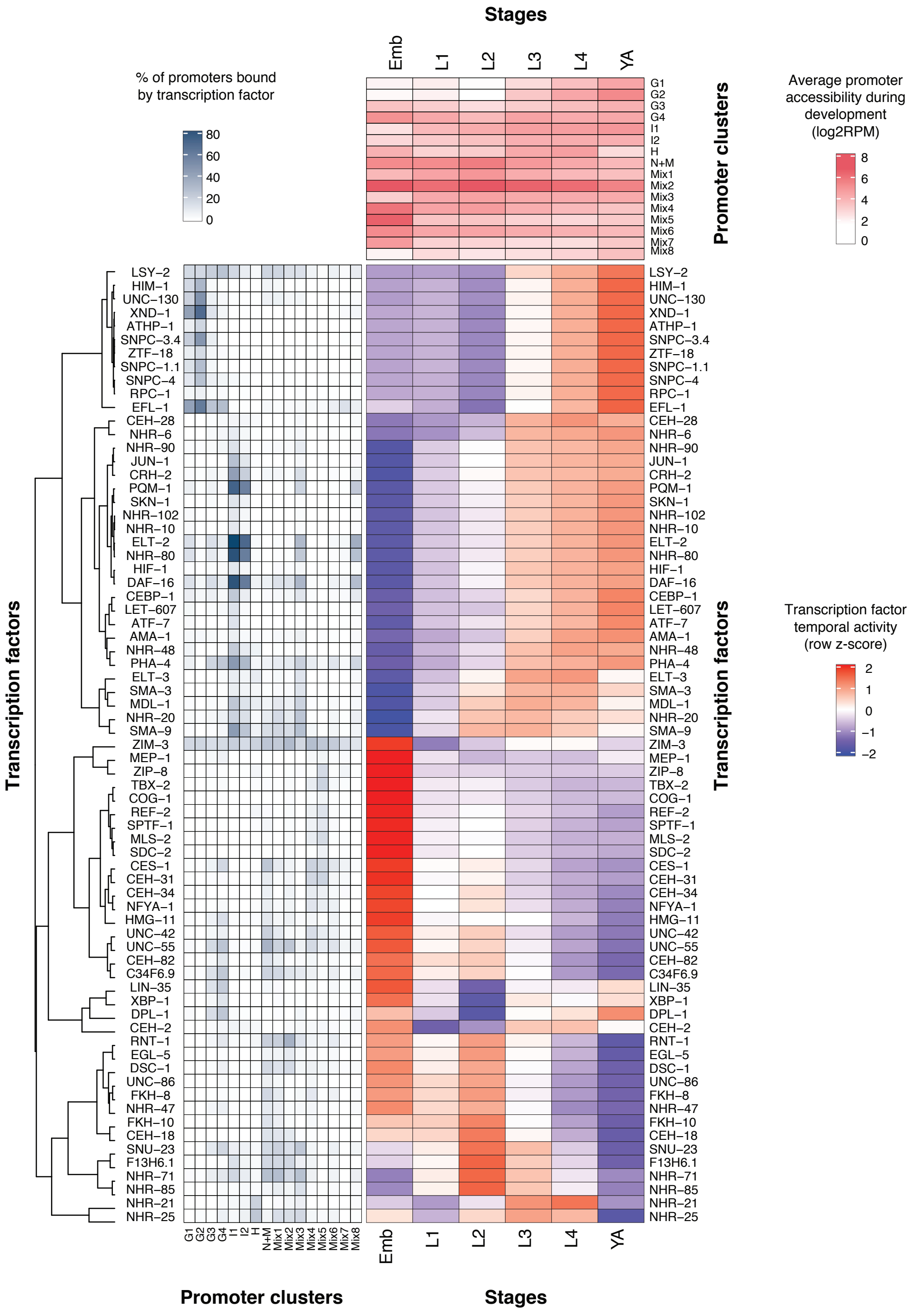


FIGURE 3.9 – Temporal activity of transcription factors during development. For each factor, its temporal activity at any given developmental stage is the product of its binding pattern across all the promoter clusters by the average promoter accessibility in each cluster during development. The resulting matrix product is scaled by rows.

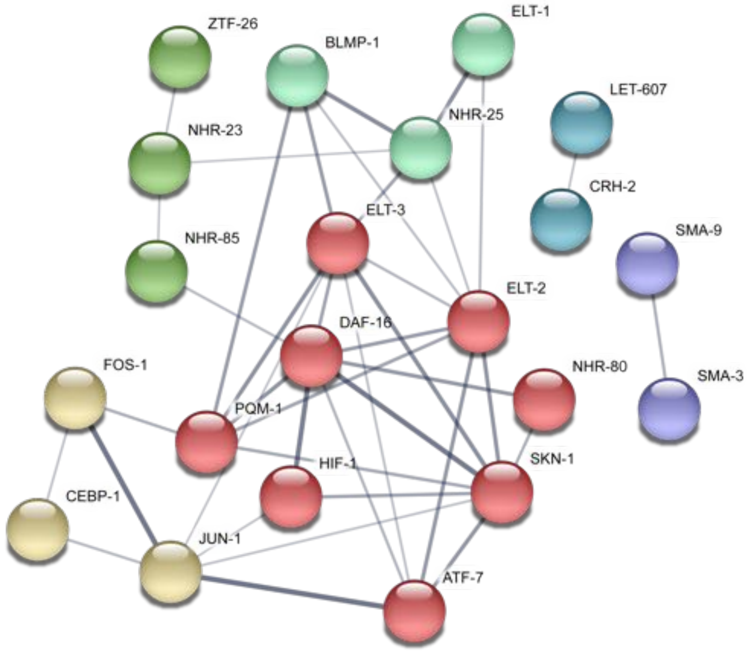


FIGURE 3.10 – Network of protein-protein interactions between transcription factors involved in aging. The interaction network is built using STRING database (Szklarczyk *et al.*, 2019). Colors indicate groups of transcription factors preferentially interacting with each other within the network. Line width indicates the confidence of the interaction.

3.2.2 Transcription factors associated with aging promoter clusters

I also evaluated transcription factor binding enrichment at aging promoter clusters. DAF-16/FoxO, a master regulator of aging (Lin *et al.*, 2001), preferentially binds promoters in aging clusters I, I+H [1], I+H [2] and Mix1 (Figure 3.7). Consistent with a prominent role of DAF-16 in the intestine (Kaplan and Baugh, 2016), these clusters are enriched for promoters associated with intestinal genes (Figure 3.6). The binding enrichment patterns of five other TFs implicated in aging (DVE-1, NHR-80, ELT-2, FOS-1 and PQM-1, Folick *et al.*, 2015; Goudeau *et al.*, 2011; Mann *et al.*, 2016; Tepper *et al.*, 2013; Tian *et al.*, 2016; Uno *et al.*, 2013) are similar to DAF-16, indicating that they bind to similar sets of promoters associated with intestinal genes and potentially function in complex(es) (Figure 3.7). This is consistent with the protein-protein interactions documented between these factors, with DAF-16 at the center of a network of proteins all physically or functionally interacting (Figure 3.10).

Importantly, promoters bound by these factors are characterized by an overall

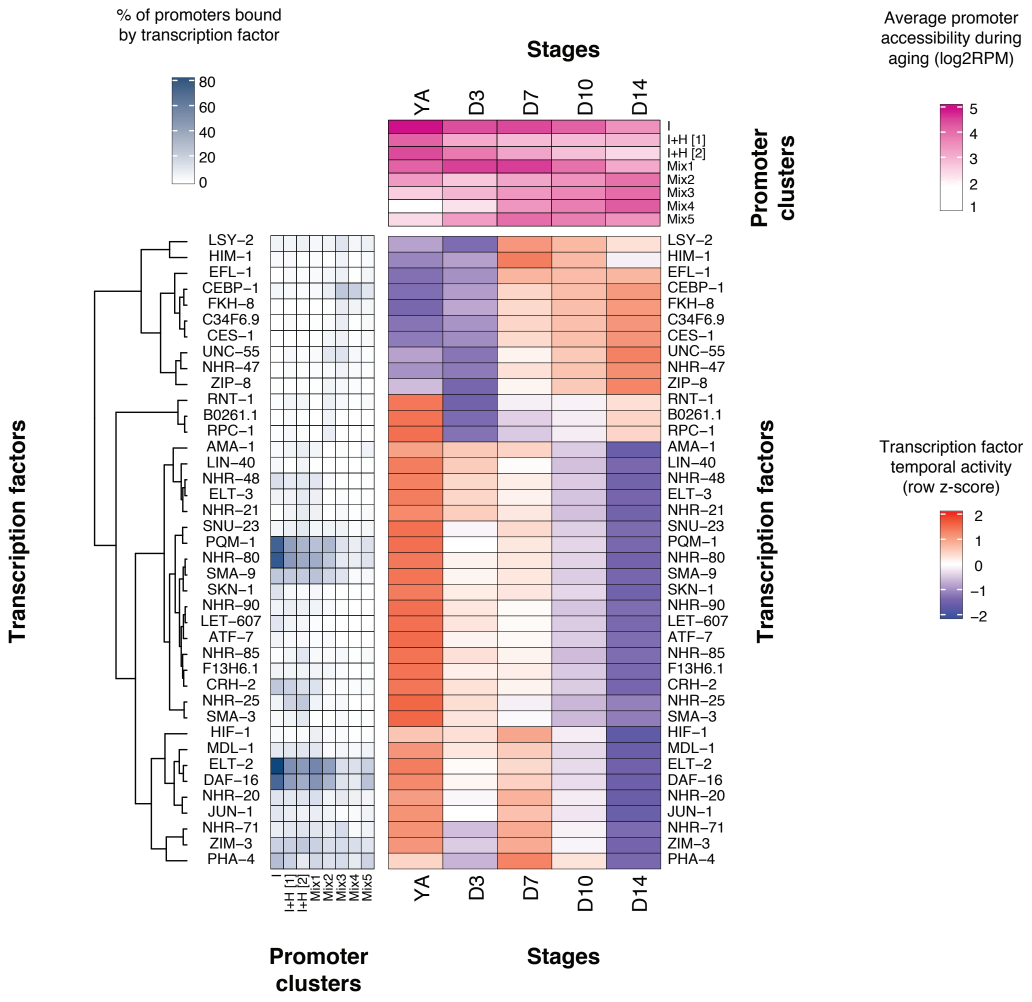


FIGURE 3.11 – Temporal activity of transcription factors during aging. For each factor, its temporal activity at any given time during aging is the product of its binding pattern across all the promoter clusters by the average promoter accessibility in each cluster during aging. The resulting matrix product is scaled by rows.

decrease of accessibility during aging (Figure 3.6). This suggests that these promoters, accessible during development of the worm and bound by TFs, undergo local chromatin remodeling during cellular senescence. This in turn may prevent key transcription factors such as DAF-16 from binding to these promoters, thus triggering tissue-wide defects accumulating during aging (Wolkow *et al.*, 2017).

Finally, I summarized the temporal activity of all the transcription factors involved in aging using the same approach as described above (Figure 3.11). This revealed a major switch of transcription factor temporal activity during aging. Around D7 (day 7), many transcription factors involved in physiological regulation of hypodermis and intestinal genes (*e.g.* ELT-2, DAF-16, SKN-1) stop being active. At the same time, an increased binding of other transcription factors normally not active in adult worms (*e.g.* ZIP-8) is observed (Figure 3.11).

These observations suggest that D7, seven days after having reached adulthood, is a pivotal time point in aging; it stands as the transition between two developmental stages regulated by different sets of transcription factors. This approach points out to transcription factor candidates which could be involved in gene regulation during aging before or after D7.

3.2.3 Unravelling new putative roles for uncharacterized transcription factors

The aforementioned results also revealed cluster-specific associations for uncharacterized transcription factors, such as ZTF-18 and ATHP-1 with germ line promoter clusters and CRH-2 with the intestinal clusters (Figure 3.7 on page 71). Particularly, CEBP-1 binding is enriched in aging promoter clusters Mix3 and Mix4 and is characterized by a readily increasing temporal activity during aging starting at D7 (Figure 3.11). This suggests a potential role of CEBP-1 in activating a subset of genes involved in cellular senescence in *C. elegans*, as it is the case for its homologue CEBP-β in mouse (Sandhir and Berman, 2010).

3.3 Discussion

In Chapter 3, I identified and characterized clusters of promoters temporally coordinated during *C. elegans* development and aging. This showed that the overwhelming majority of promoters (75%) are regulated during development. I characterized different dynamics of coordinated chromatin accessibility during development, including monotonously decreasing or increasing accessibility over time, but also oscillating accessibility. To get an insight on their function, I have correlated each cluster of promoters with orthogonal data, such as single-cell-derived tissue-specific gene expression or gene ontology terms. I observed expected results, such as the oscillating promoters being associated with molting process (cluster H in Figure 3.3), a circadian process (Turek and Bringmann, 2014). New insights also emerged from this analysis, for example the existence of sets of promoters characterized by the same trend of accessibility changes but with different amplitudes, which could be further studied in the future.

Chromatin accessibility changes in aging have never been studied before, and *C. elegans* stands as the ideal model system to do so. I found that a small set of promoters (13%) is dynamically accessible during aging in *C. elegans*. These promoters are largely associated with intestinal and hypodermal functions, confirming the central role of these two tissues in aging (Gelino *et al.*, 2016; Herndon *et al.*, 2002; McGee *et al.*, 2011).

Interestingly, the set of dynamic promoters in development and in aging largely overlap, and ~ 20% of all the annotated promoters do not show any significant accessibility change throughout *C. elegans* life. It is conceivable that specific mechanisms ensure the constant accessibility of these persistent promoters over time. In the future, investigation of this set of promoters could shed light on some of these mechanisms.

Using modENCODE/modERN public database, I also identified transcription factors preferentially binding at each cluster of promoters. This confirmed and extended the current knowledge of transcription factor binding during development and aging, as well as the biological functions regulated by these transcription factors. I also defined the "temporal activity" metric, to provide an alternative metric of the activity of a given transcription factor across development and aging. This approach integrates the transcription factor binding patterns measured by ChIP-seq and the dynamic accessibility of promoters bound by these transcription factor. Thus, this metric aims at describing the binding patterns of a transcription factor over time more directly, rather than relying on its levels of expression. However, for each transcription factor, its ChIP-seq profile is usually derived from the developmental stage where the factor is most expressed (Gerstein *et al.*, 2010). Thus, an important limitation of this approach is the assumption that the sites bound by each factor do not significantly vary during development.

This analysis is the first step in understanding how regulatory elements contribute to gene regulation. However, a precise annotation of tissue-specific regulatory elements is still lacking. This is a major hurdle which prevents from investigating the mechanisms of tissue-specific regulation.

Chapter 4

Optimizing tissue-specific chromatin profiling in *C. elegans*

In Chapter 3, I characterized temporal patterns of chromatin accessibility during development and aging. However, to understand developmental regulation of cell-type specific gene expression, comprehensive annotation of regulatory element activities in different cells is needed.

To profile chromatin accessibility in specific cells, it is necessary to have a reliable method to isolate material of interest. In Chapter 4, I present the work I conducted in the first half of my PhD to optimize an experimental procedure to isolate nuclei from individual tissues in *C. elegans*.

Collaboration note: Michael Chesney designed and cloned the StrepTag::GFP::UNC-83 reporter and generated two of the five reporter strains. Chiara Cerrato helped inject constructs in the EG6699 strain. Rhys McDonough helped generate CUT&RUN libraries.

4.1 Optimizing an approach to isolate material from individual tissues

4.1.1 Existing methods to isolate tissue-specific material

Several different approaches have been used to isolate biological material from individual tissues in *C. elegans* (Table 1.3 on page 42). Micro-dissection has been successfully used to obtain material specifically from germline or intestine (Diag *et al.*, 2018; McGhee *et al.*, 2007; Ortiz *et al.*, 2014; West *et al.*, 2018). This allowed the characterization of germline gene expression in different subset of germline cells (*e.g.* germline meiotic cells, oocytes, sperm) or the identification of alternative poly-adenylation usage. However, the tissues obtained using micro-dissection are not completely pure. For example, the somatic gonad is also extracted when the germline is isolated by micro-dissection. It would also be feasible to isolate neurons, muscle or hypodermis by micro-dissection, but this would only yield low amounts of material. RNA tagging by an engineered RNA-binding protein has also been used to isolate RNA from specific tissues or cell types, but this approach is limited to transcriptome analysis and cannot be used to assay chromatin features such as its accessibility (Blazie *et al.*, 2015; Ma *et al.*, 2016). Isolation of tagged nuclei by affinity purification (INTACT) has been used in worm to isolate muscle nuclei in young adults (Steiner *et al.*, 2012) and seems to be a promising technique. However, it relies on antibody-covered magnetic beads to separate nuclei by immunoaffinity, resulting in nuclei clumping. This would not hinder standard RNA-seq and/or ATAC-seq procedures but it would exclude the possibility of using single-cell approaches on these isolated nuclei, as these methods require high-quality single-nucleus suspensions.

FACS-based methods have been used to sort specific cells, sometimes generating high-quality RNA extracts but other times only allowing for 3'-end-seq (see Table 1.3 on page 42). Yet, full-length RNA-seq is crucial to investigate many aspects of gene expression (*e.g.* alternative exon splicing) and more importantly to annotate regulatory elements as promoters. Moreover, because of the worm thick cuticle, cells embedded in highly connected complex tissues such as neurons can be difficult to isolate in larvae and adult worms. Rather than trying to isolate specific cells, I sought to optimize a nuclear sorting approach inspired from the standard fluorescence-activated cell sorting (FACS) technique.

4.1.2 Optimizing a FACS-based sorting procedure

4.1.2.1 Designing a fluorescent reporter for nuclei sorting

To sort a population of tissue-specific nuclei (*e.g.* all muscle nuclei) by FACS, I needed a method to fluorescently label nuclei of interest. I initially decided to use strains where a GFP reporter would be specifically expressed in nuclei from individual tissues. I took advantage of the large collection of strains generated in the lab to investigate which type of reporter construct would yield the best results.

Chromatin-bound or not? Different GFP reporter strains have been generated for promoter- and reporter-assay (*e.g.* Jänes *et al.*, 2018), however, these generally relied on a GFP protein (26.9 kDa) being fused to HIS-58, the histone H2B protein. Conceptually, this results in a ~25% increase of the total molecular weight of the histone octamer constituting the nucleosome protein core. Though this has been widely used to mark lineage-specific cells, the impact of this histone tagging on chromatin organization and nucleosome remodeling dynamics has never been formally investigated. Since the study of chromatin accessibility and nucleosome organization is at the core of my project, I decided to exclude chimeric histone-GFP reporters to avoid any potential technical artifact. Instead, I focused on non-chromatin-bound GFP reporters.

Freely diffusing or localized? An important feature of a fluorescent reporter used for sorting is its brightness. By microscopy, the NLS::GFP::LacZ reporter (a GFP fused to a nuclear localization signal and to a LacZ protein) shows strong fluorescence in the PD4251 strain where it is expressed under the control of the muscle-specific *myo-3* promoter (Timmons *et al.*, 2003). In fact, its expression is so high that it is visible even under the dissecting microscope. Yet, even though the GFP signal is robustly detected in living worms, I found that it was largely lost during nuclei preparation from frozen worms, and could not be detected by FACS (not shown). One possible explanation is that the small chimeric protein leaks out during nuclei preparation, potentially due to increased permeability of the nuclear envelope once nuclei are isolated. Another explanation is that the fluorescent protein is located within the nucleus and thus its signal cannot be readily detected by FACS. Thus, I decided to rely on a chimeric protein anchored to the outer side of the nuclear envelope. This presents three advantages: (i) it directly exposes the fluorescent protein to the FACS sensors, without nuclear envelope in-between, (ii) it traps it within the phospholipid bilayer, preventing it from diffusing away during nuclei preparation and finally (iii) it limits the interactions between the fluorescent reporter and chromatin and their potential side effects.

To target a GFP reporter to the nuclear envelope, I used a construct generated by Michael Chesney in the lab (inspired from Henry *et al.*, 2012), consisting of the GFP protein fused to a Streptavidin tag and to the C-terminal sequence of UNC-83 protein. UNC-83 is a conserved major constituent of the nuclear envelope and its C-terminal end is located on the outer nuclear membrane (McGee *et al.*, 2006). Upon integration in *C. elegans* genome, this chimeric protein is localized to the nuclear envelope (see 4.1.2.2 and Figure 4.3). I also generated an alternative tag consisting of the NPP-9 protein fused to GFP (previously described in Steiner *et al.*, 2012) and used it to label neuron-specific nuclei.

Array or single-copy transgene integration? Finally, the method by which a reporter is integrated to the worm genome is important. For example, the NLS::GFP::LacZ reporter in the PD4251 strain is integrated in *C. elegans* genome as an array and thus expressed at very high levels. While this seems ideal for increased brightness, transgenes in integrated arrays are known to show variable transformation efficiencies and to be silenced over generations in germline (Mello *et al.*, 1991). For these reasons, I decided to favor targeted single-copy reporter integrations over arrays, to ensure reproducible and heritable labelling of individual tissues. Targeted single-copy reporter integrations can be achieved by using mos1-mediated Single Copy Insertion (mosSCI) (Frøkjaer-Jensen *et al.*, 2008).

Which promoters to control reporter expression? To isolate nuclei from individual tissues, I need to express the fluorescent reporter in each specific tissue separately. Many reporter assays have been performed in worm and regulatory regions driving different patterns of tissue-specific expression have already been identified. Among others, the *mex-5* promoter drives reporter expression in germline (Merritt *et al.*, 2008); the *myo-3* promoter is typically used to drive expression of reporters in all muscle cells (Fox *et al.*, 2007); *npa-1* promoter has been used for reporter expression in intestine (Segref *et al.*, 2010); *dpy-7* promoter has been used for reporter expression in hypodermis (Gilleard and McGhee, 2001); *egl-21* promoter has been used for reporter expression in all neurons (Jacob and Kaplan, 2003).

I decided to use these promoters to drive a tissue-specific expression of my reporter. These promoters present two major advantages:

1. The five associated genes have been shown to be expressed across their entire

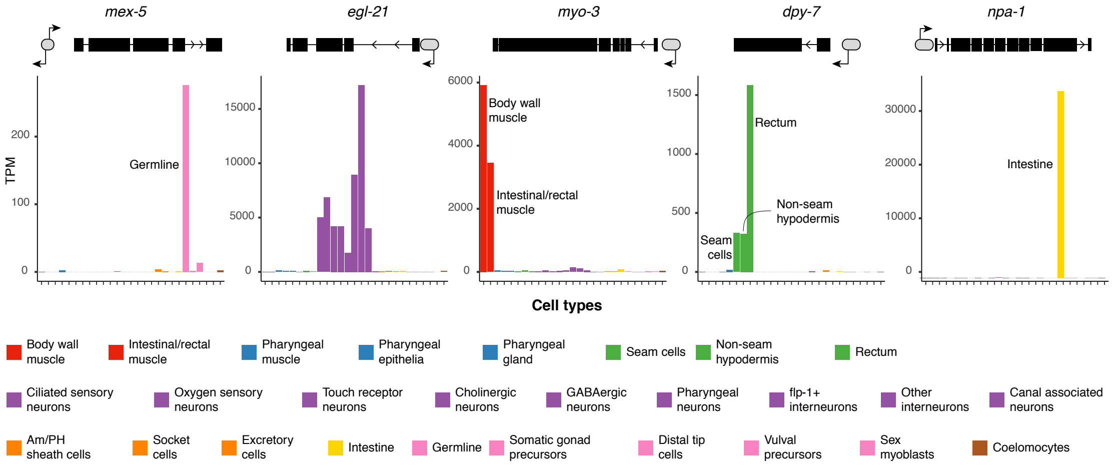


FIGURE 4.1 – Expression of tissue-specific genes in different cell types. The promoters used to create the reporter strains are obtained from these genes (data from Cao *et al.*, 2017).

tissue using single-cell RNA-seq (*e.g.* *egl-21* is expressed across all types of neurons, Figure 4.1 and Cao *et al.*, 2017).

2. The regulatory regions associated to these genes are relatively simple and promoters have been annotated to specific genomic loci (Figure 4.1).

4.1.2.2 Creating tissue-specific reporter strains

I decided to rely on Gateway cloning technology to generate the constructs expressing the chimeric fluorescent marker under the control of the different tissue-specific promoters. In Gateway cloning, three "slots" are stitched to each other using specific recombinases (Figure 4.2A). In my case, the three slots were:

1. The promoter controlling the expression of the reporter in the appropriate tissue (slot 1);

2. The GFP coding sequence, fused at the N-terminus to a Streptavidin Tag and at the C-terminus to the first 290 amino-acids of UNC-83 (slot 2);

3. *tbb-2* 3' untranslated region (UTR), which is required for proper expression of the construct (Merritt *et al.*, 2008; Zeiser *et al.*, 2011, slot 3).

These three "slots" were integrated in a mosSCI-compatible backbone vector and injected into the EG6699 strain (Frøkjaer-Jensen *et al.*, 2008). In total, five strains were generated, each one of them expressing the final GFP reporter under the control of a tissue-specific promoter (Figure 4.2B and Table 4.1). I validated the

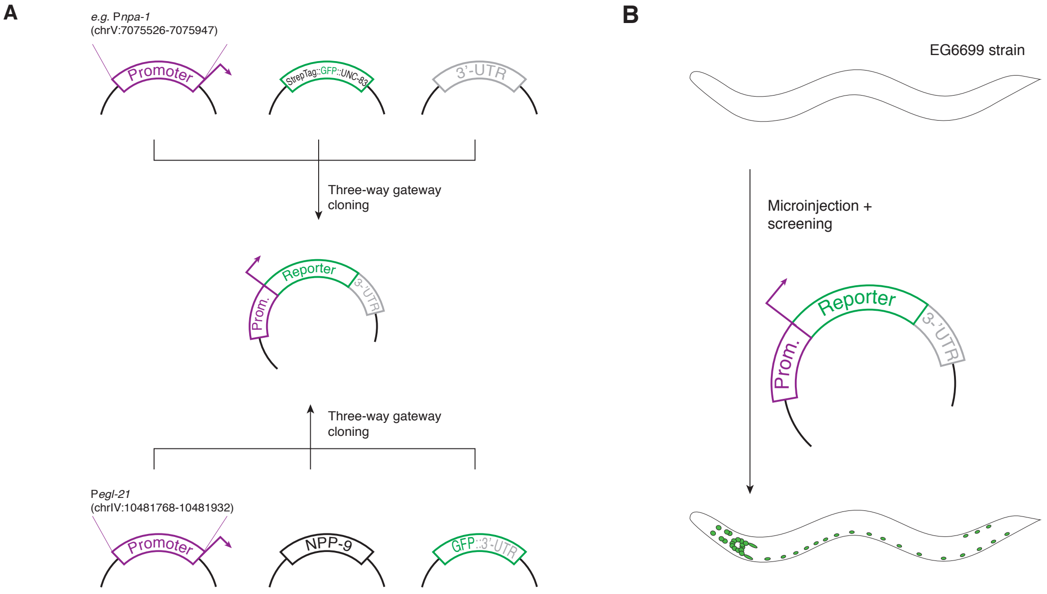


Figure 4.2 – Gateway strategy used to create tissue-specific reporter strains. **A-** A three-way gateway cloning strategy is used to insert a promoter, a fluorescent reporter coding sequence and a 3'-UTR into a destination vector. **B-** The resulting construct can be injected into the EG6699 strain to integrate into the ttTi5605 Mos1 allele on chrII as a single copy.

Table 4.1 – Tissue-specific reporters used in this study. The promoter controls the expression of the reporter in a single tissue. Spatio-temporal expression of each reporter has been established by visual observation of the reporter strain. The tissue in parenthesis is where GFP fluoresence is observed.

Tissue marked (Strain ID)	**Promoter**	**Reporter**	**Expressed**
Germline (JA1616)	*mex-5* (chrIV:13353242-13353729)	StrepTag::GFP ::UNC-83	Emb: - L1: - L3: +/- (germline) YA: + (germline)
Neurons (JA1816)	*egl-21* (chrIV:10481768-10481932)	NPP-9::GFP	Emb: + (neurons) L1: + (neurons) L3: + (neurons) YA: + (neurons)
Muscle (JA1585)	*myo-3* (chrV:12234302-12236686)	StrepTag::GFP ::UNC-83	Emb: + (muscle) L1: + (muscle) L3: + (muscle) YA: + (muscle)
Hypodermis (JA1815)	*dpy-7* (chrX:7537794-7538688)	StrepTag::GFP ::UNC-83	Emb: + (hypodermis) L1: + (hypodermis) L3: + (hypodermis) YA: + (hypodermis)
Intestine (JA1817)	*npa-1* (chrV:7075526-7075947)	StrepTag::GFP ::UNC-83	Emb: + (intestine) L1: + (intestine) L3: + (intestine) YA: + (intestine)

reporter integration in the five strains and assessed its spatiotemporal pattern of expression by microscopy (Figure 4.3 and Table 4.1).

4.1.2.3 Sorting nuclei from individually labelled tissues

Once I generated the five tissue-specific reporter strains, I started optimizing the tissue-specific nuclear sorting procedure. I could reproducibly obtain suspensions of intact and well-separated nuclei with visible GFP fluorescence in subsets of nuclei. However, when running a nuclei preparation through a cell sorter, no GFP+ sub-population could be detected (data not shown). It is known that the GFP protein does not perform well for fluorescence-activated sorting, compared to other fluorescent proteins or fluorophores. For instance, phycoerythrin (PE) is a fluorescent protein characterized by a great absorption coefficient and an almost perfect quantum efficiency, and can be directly conjugated to antibodies. I decided to enhance fluorescence by immunostaining the GFP protein with a PE-conjugated α-GFP antibody (4.4). This strategy resulted in a clear separation of two nuclei populations, one PE- (*i.e.* GFP-) and one PE+ (*i.e.* GFP+) (4.4B).

I also initially struggled to recover sorted nuclei. After sorting 500,000 nuclei, pelleting and resuspending them in a small volume, very few nuclei were detected by microscopy but a lot of debris and burst nuclei were visible. Biological material is often prone to degradation after sorting, so I optimized the handling of the nuclei post-sorting to prevent them from bursting. After many tests, I found that sorting nuclei into a 15mL polystyrene Falcon tube pre-coated and pre-filled with 500uL of nuclei buffer enriched 0.025% IGEPAL CA-630 (a mild non-ionic non-denaturing detergent) ensured that the nuclei remained intact and did not adhere to the sides of the collection tube.

The optimized experimental procedure can now be used to isolate specific populations of nuclei labelled by a GFP reporter targeted to the nuclear envelope. Importantly, this sorting approach yields highly pure populations: more than 95% of the sorted nuclei are GFP+ when the purity is assessed post-sorting by microscopy and/or by flow cytometry (4.4B).

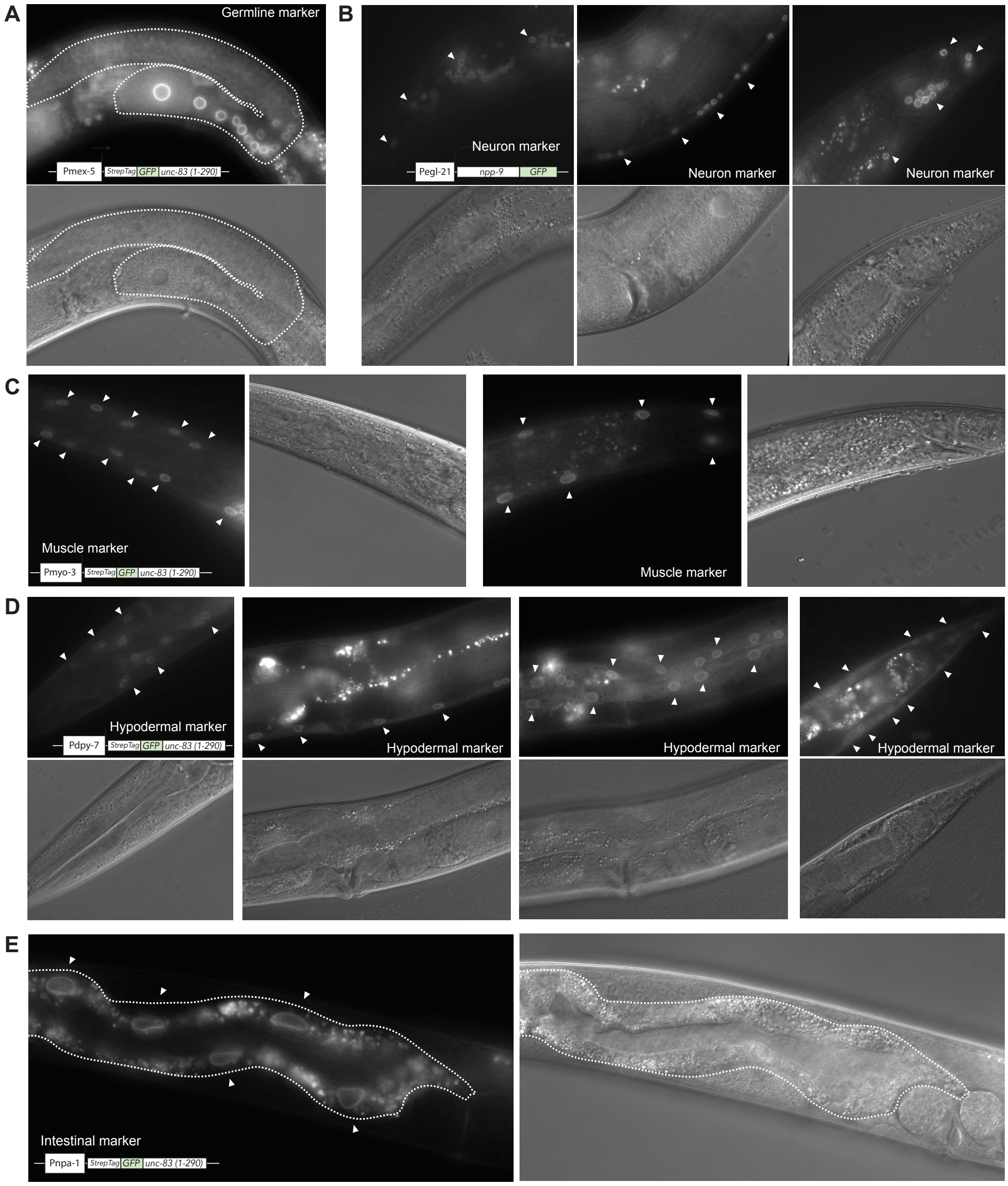


Figure 4.3 – Reporter strains created for this study, to label nuclear envelope of germline nuclei (**A**), neuronal nuclei (head neurons, ventral nerve cord and tail neurons) (**B**), muscle nuclei (anterior and posterior sides) (**C**), hypodermis nuclei (head, ventral hypodermal ridge, seam and tail) (**D**) and intestine nuclei (**E**). For each reporter, the construct used to drive reporter expression is depicted and the resulting GFP expression pattern is shown. DIC images are also shown for reference.

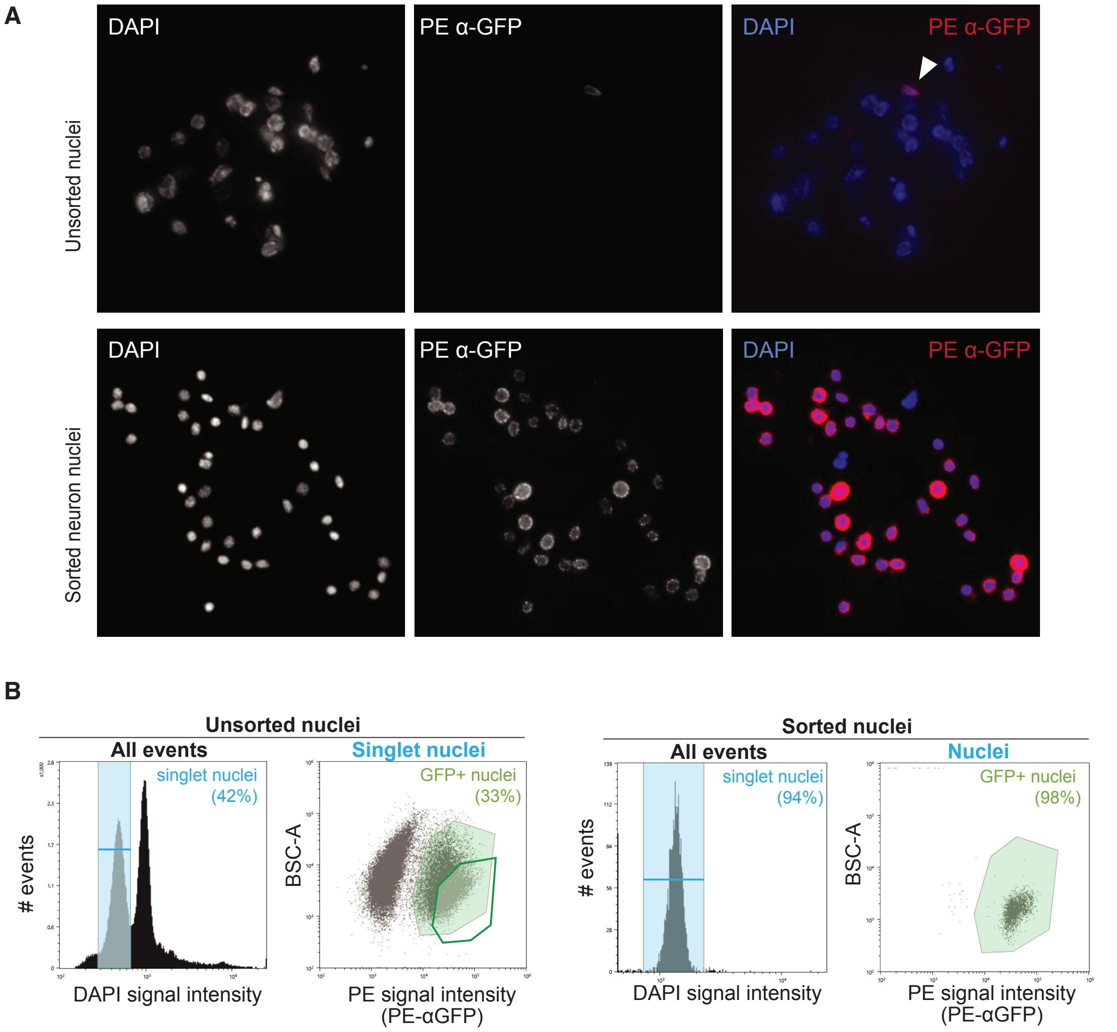


Figure 4.4 – PE-conjugated α-GFP immunostaining of nuclei and sorting strategy. **A-** Nuclei from neuronal reporter strain (P*egl-21*::*npp-9*::GFP::*tbb-2*-3'UTR) immuno-stained with a phycoerythrin (PE) α-GFP antibody, before (top) and after (bottom) nuclei sorting. The arrow points to a single PE+ nucleus. **B-** top: gating strategy to isolate PE+ (*i.e.* GFP+) nuclei from a nuclear preparation. Single nuclei are gated (shaded blue area) and GFP+ nuclei (green shaded area) are readily separated from GFP- nuclei. Here, the gate used to sort GFP+ nuclei is the thick-lined green gate (no shading). Bottom: flow cytometry recording of sorted nuclei to estimate the purity of GFP+ nuclei after sorting.

4.2 Optimizing downstream genome-wide profiling assays

Thanks to this optimized sorting procedure, I was able to sort tissue-specific nuclei based on their GFP signal and efficiently recover them for downstream procedures. Still, the sorting throughput remained generally low (~ 100,000 to 500,000 in an hour of sorting). On the other hand, genome-wide experiments such as RNA-seq and/or ATAC-seq are typically performed using several million nuclei. Thus, it was crucial to ensure these assays could work using lower inputs of sorted nuclei.

4.2.1 Nuclear transcription: RNA-seq

RNA-seq is a technique widely used to profile genome-wide gene expression in populations of cells. However, in *C. elegans*, ~ 70% of the mature cytoplasmic transcripts are trans-spliced: their 5' end (the "outron") is spliced out and replaced by a splice leader sequence (Allen *et al.*, 2011). The detection of outrons is important to accurately map transcription initiation events and to annotate promoter activity for regulatory elements (1.2.3.1 on page 31 and Chen *et al.*, 2013; Jänes *et al.*, 2018). By sorting nuclei rather than cells, the RNA extracts are enriched in nuclear immature transcripts with their outron still unspliced.

Intact RNA is required to generate high-quality RNA-seq libraries. However, RNA is readily degraded if buffers contain even traces of RNase and can also be chemically fragmented when heated in presence of divalent metal cations. Thus, the nuclei sorting procedure can potentially lead to RNA degradation at many steps. I extracted RNA from nuclei immediately after their isolation from whole worms, after immunostaining with a PE-conjugated α-GFP antibody and finally after sorting (Figure 4.5A). This revealed that RNA was largely degraded as soon as nuclei were immunostained. I confirmed this by performing a stepwise RNA extraction (Figure 4.5B). The first steps of nuclei isolation do not have a major impact on RNA integrity, but incubation of nuclei in staining buffer for as short as five minutes was enough to largely degrade RNA. I then incubated purified intact RNA in different solutions for one hour, to test which one would cause RNA degradation (Figure 4.5C). As expected, staining buffer completely degrades pure RNA, even in presence of RNase inhibitors. The staining buffer recipe is composed

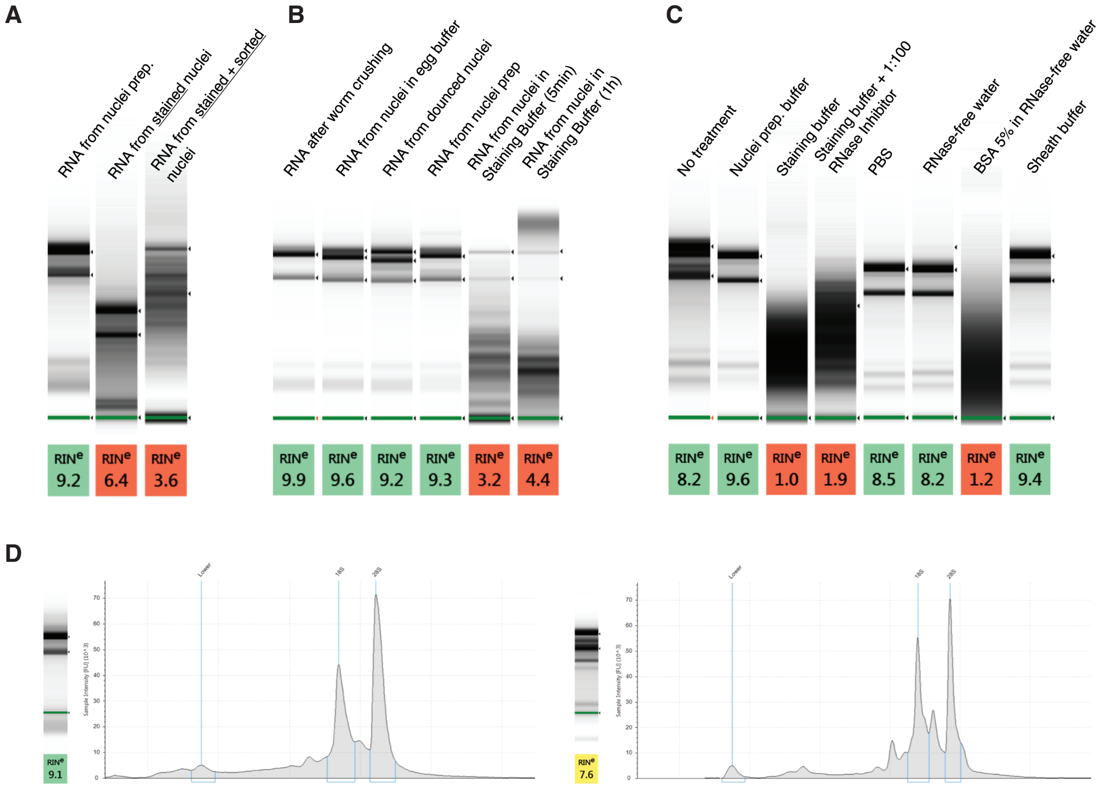


Figure 4.5 – RNA integrity during nuclei sorting procedure. **A-** RNA integrity before nuclei staining, after nuclei staining and after nuclei sorting. **B-** RNA integrity after each step of nuclei isolation. **C-** RNA integrity after one hour of incubation in the indicated solutions. **D-** RNA integrity from RNA prepared from simple nuclei extraction (left) and from nuclei after sorting (right). In this experiment, nuclei have been stained in their extraction buffer.

of 5% BSA and 1mM EDTA in PBS (Bonn *et al.*, 2012). I found that RNA incubated with 5% BSA in RNase-free water was readily degraded (Figure 4.5C). From these results, I concluded that the BSA (supposedly RNase-free, according to the manufacturer) was contaminated with RNase. I then experimented staining in different other buffers and realized that immunostaining with the PE-conjugated α-GFP antibody was possible in the standard nuclei extraction buffer without added BSA, thus preserving RNA integrity.

Eventually, optimized immunostaining and sorting procedures yield mostly intact RNA (Figure 4.5D), which can then be processed into a library for high-throughput sequencing.

4.2.2 Chromatin accessibility: ATAC-seq

ATAC-seq is one of the most straightforward genome-wide techniques to study chromatin accessibility (see 1.1.3.2). Even though the assay was originally designed for cultured cell lines (Buenrostro *et al.*, 2013), it has since then been used in many organisms using both cells and isolated nuclei. In worm, it has been performed using one million nuclei from standard *C. elegans* nuclear preparations (Jänes *et al.*, 2018). To test whether the procedure would yield high-quality results using a lower number of sorted nuclei, I sorted different amounts of nuclei and performed ATAC-seq on each individual sorted population (Figure 4.6).

ATAC-seq performed using 500,000 nuclei yields a very satisfactory chromatin accessibility profile, comparable to the high-quality reference accessibility track from Jänes *et al.* (2018) (Figure 4.6A). Importantly, ATAC-seq performed with decreasing amounts of sorted nuclei also showed good enrichment of accessibility over the background, though generally slightly lower than that of ATAC-seq done with 500,000 nuclei (Figure 4.6B). Moreover, the peaks annotated with MACS2 using ATAC-seq results from 500,000 unsorted nuclei or ten times less sorted nuclei show a very good overlap (Figure 4.6C).

These results suggest that ATAC-seq can be performed using sorted nuclei with fewer nuclei than initially assumed and still result in high-quality chromatin accessibility tracks. In the following ATAC-seq experiments, I typically used between 25,000 and 50,000 sorted nuclei to perform ATAC-seq.

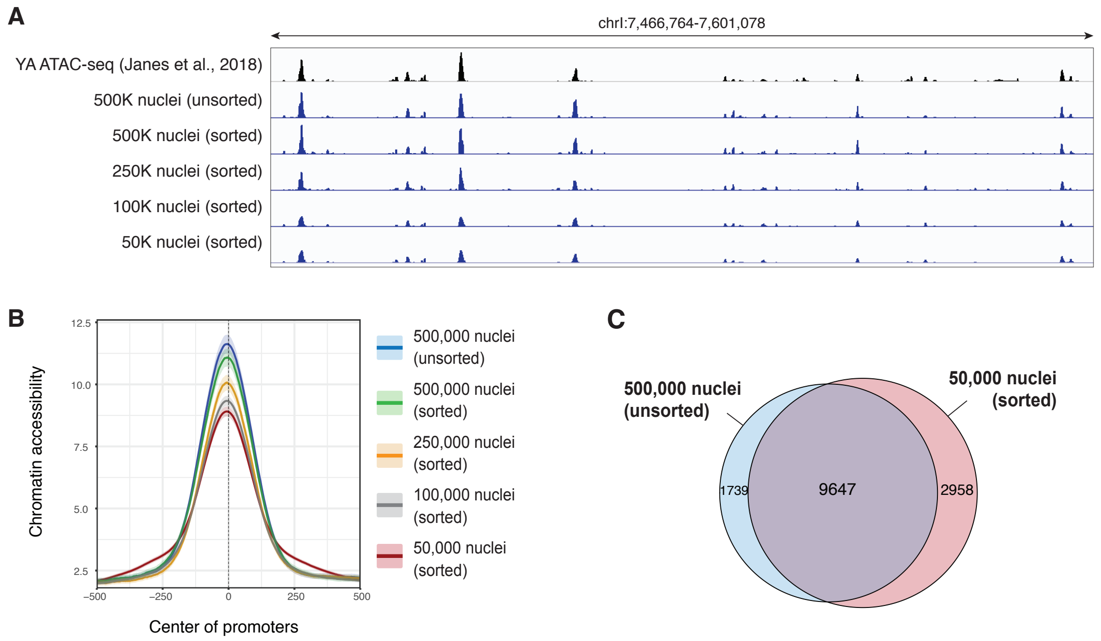


FIGURE 4.6 – Results of ATAC-seq performed on different amounts of nuclei. **A-** Snapshot of chromatin accessibility tracks from reference ATAC-seq track (in black, from Jänes *et al.*, 2018) or performed in 500K unsorted nuclei or 500K, 250K, 100K and 50K sorted nuclei. The scale is the same across the tracks. **B-** Aggregate profiles over annotated *C. elegans* promoters. **C-** Euler diagram representing overlap between sets of peaks obtained from ATAC-seq results from 500K unsorted nuclei (left) or 50K sorted nuclei (right).

4.2.3 Histone modifications: ChIP-seq and variants

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a powerful method traditionally used to profile histone modifications or transcription factor binding landscapes. It has been widely used in *C. elegans* to understand the mechanisms underlying gene regulation (*e.g.* Latorre *et al.*, 2015; McMurchy *et al.*, 2015), yet only one study so far has reported tissue-specific ChIP-seq profiling in worm (Steiner *et al.*, 2012). ChIP-seq experiments usually require several million nuclei to prepare chromatin extracts for immunoprecipitation, which cannot be easily obtained using sorting-based approaches. However, recent advances have lowered the amount of input needed to profile histone modifications (*e.g.* Brind'Amour *et al.*, 2015; Skene and Henikoff, 2017; Valensisi *et al.*, 2015). cChIP-seq is an straightforward adaptation of the standard ChIP-seq protocol. By adding recombinant histones harboring the modification of interest as a carrier (*e.g.* recombinant H3 with H3K4me3 modification), one can lower the chromatin input to as low as tens of nanograms, rather than micrograms (Valensisi *et al.*, 2015). Alternatively, nuclease-based alternatives such as the 'Cleavage Under Targets and Release Using Nuclease' (CUT&RUN) method are also emerging (Skene and Henikoff, 2017). In CUT&RUN, an epitope of interest (*e.g.* H3K4me3) is marked by an antibody in intact unfixed permeabilized nuclei, then a chimeric protein A fused to MNase is added to the nuclei. This leads to controlled DNA cleavage upstream and downstream of the epitope and the cut DNA is then released and processed into a sequencing library. I decided to test whether cChIP-seq and CUT&RUN would work for low inputs of sorted tissue-specific nuclei. I profiled H3K4me3, H3K36me3 and H3K27me3 histone marks on sorted nuclei by cChIP-seq (neuron nuclei) and CUT&RUN (muscle nuclei) and compared the results to gold-standard ChIP-seq.

Both CUT&RUN and low-input ChIP-seq methods yielded decent results for H3K4me3 profiling (Figure 4.7A), with an important difference. H3K4me3 enrichment profile from CUT&RUN is more focused at promoters than that from low-input ChIP-seq (Figure 4.7A-B). This could be due to the different methods used to achieve DNA fragmentation (physical fragmentation in ChIP-seq and enzymatic cleavage in CUT&RUN), or to the fact that CUT&RUN can lead to artifactual signal at accessible chromatin loci (Skene and Henikoff, 2017).

Previous work in *C. elegans* showed that H3K36me3 and H3K27me3 histone modification landscapes are respectively segregated and form active or regulated domains (Evans *et al.*, 2016). Borders between each type of domain are characterized by a switch between these histone modification patterns. Such switch between HK36me3 and H3K27me3 can be observed from either CUT&RUN or low-input ChIP-seq, suggesting that the profiling of these modifications is possible in sorted tissue-specific nuclei. In these test experiments, H3K27me3 profile seems better by CUT&RUN than by cChIP-seq, while H3K36me3 profile seems better by cChIP-seq than by CUT&RUN (Figure 4.7B-C). However, further experiments are needed to confirm these observations.

Active and regulated domains have been annotated using mixed-tissue whole-organism ChIP-seq datasets, thus the variability of these chromatin domains in individual tissues is largely unstudied. Notably, it is still unknown whether regulated domains, which contain tissue-specific genes, harbor H3K27me3 or H3K36me3 histone modifications in the nuclei from the tissue(s) where the genes they contain are expressed. I focused on three neighboring genes on chrII, which are characterized by different patterns of expression: *C03H5.5* is specifically expressed in muscle

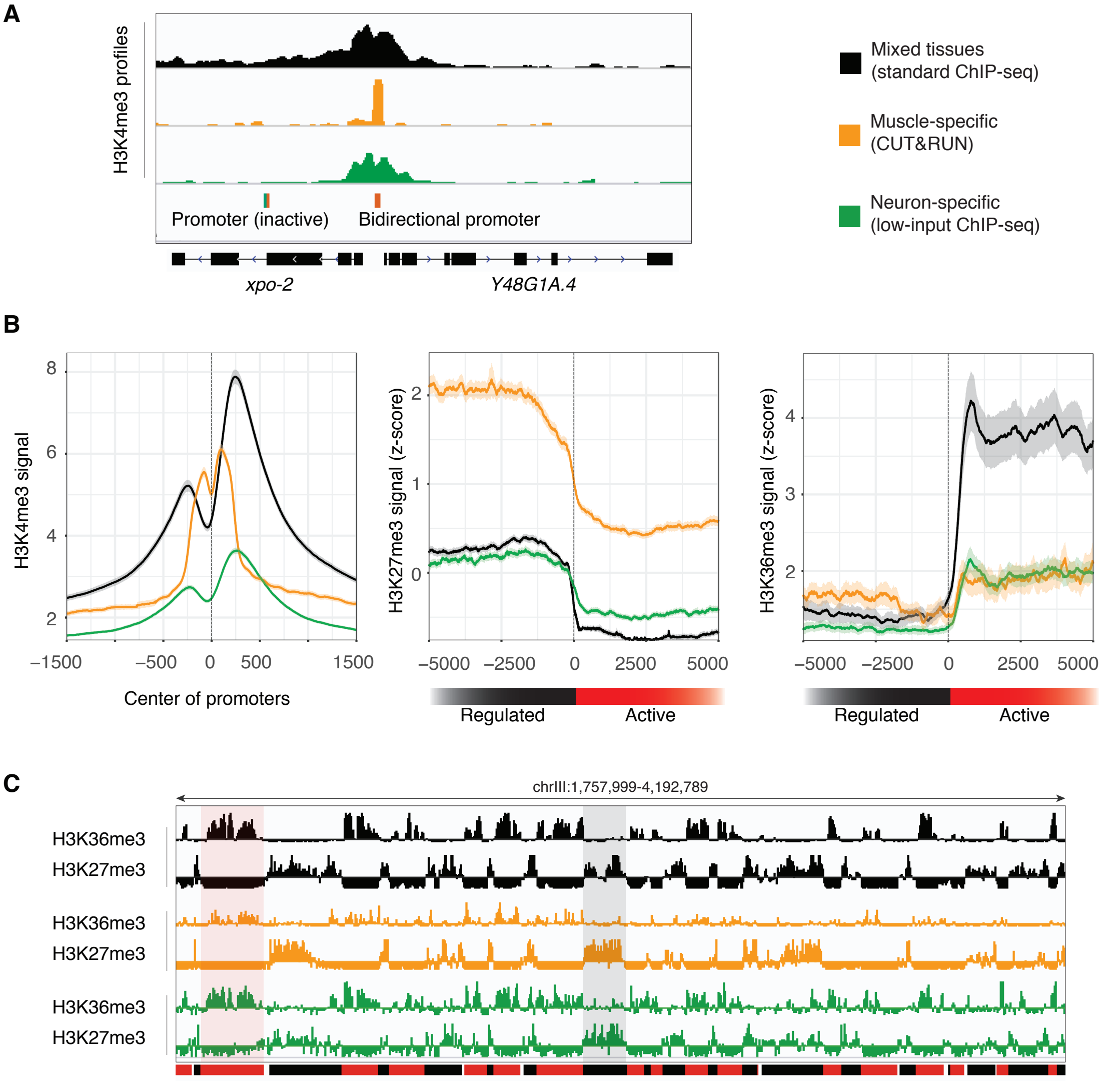


FIGURE 4.7 – Results of histone modifications profiling performed using different methods. **A-** H3K4me3 profiling by reference ChIP-seq protocol in whole organisms (in black, from Jänes *et al.*, 2018), by CUT&RUN in muscle-specific sorted nuclei (in orange) and by low-input ChIP-seq in neuron-specific sorted nuclei (in green). The scale is the same across the tracks. **B-** Aggregate plots of H3K4me3 signal (left), H3K27me3 signal (middle) and H3K36me3 signal (right). **C-** H3K36me3 and H3K27me3 profiling in the arm region of chrIII. Chromatin domains (Evans *et al.*, 2016) are shown in black (regulated domains) and red (active domains), and two domains are highlighted by the transparent rectangles. The scale is the same across the six tracks.

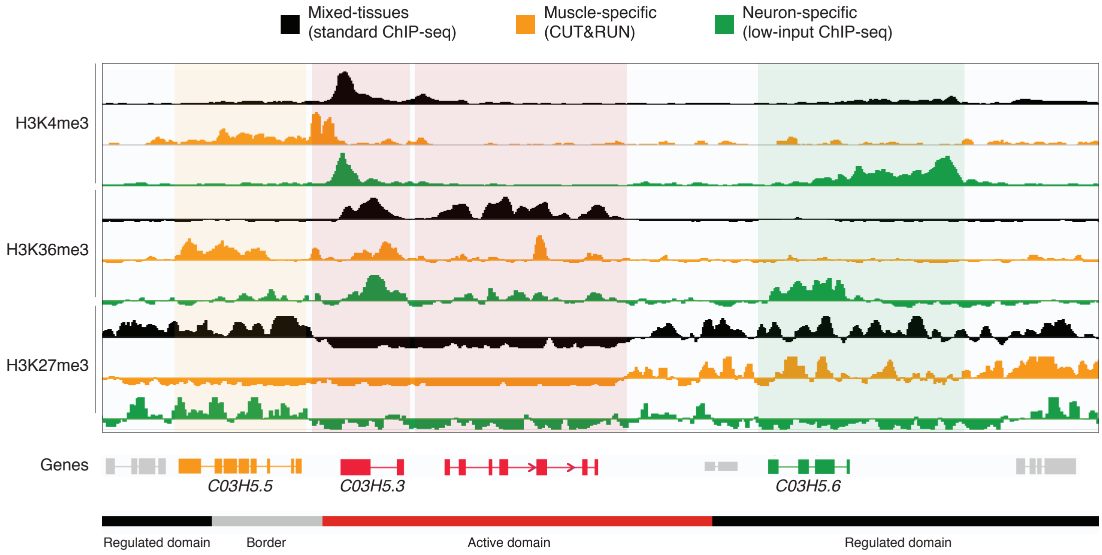


Figure 4.8 – Histone modifications profiles in muscle and neuron nuclei. Three histone modifications are profiled: H3K4me3, H3K36me3 and H3K27me3. Modifications are profiled in muscle nuclei by CUT&RUN (in orange) and in neuron nuclei by low-input ChIP-seq (in green). Reference ChIP-seq tracks from mixed-tissues whole-organism worms are also displayed (in black). Chromatin domains (Active: red; Regulated: black; borders: gray) are also shown at the bottom. The orange-shaded area represents a muscle-specific locus, the green-shaded area represents a neuron-specific locus and the red-shaded area represent ubiquitous loci. For each histone modification, the vertical scale is the same across the three tracks.

cells, *C03H5.3* is ubiquitously expressed and *C03H5.6* is only expressed in neurons (Figure 4.8). H3K4me3, H3K36me3 and H3K27me3 histone modification profiles from mixed-tissue whole-organism samples suggest that only *C03H5.3* (located in an active domain) harbors H3K36me3 modifications whereas the two other genes appear covered by H3K27me3. Yet, when looking at tissue-specific histone modification profiles, it clearly appears that indeed, muscle-specific *C03H5.5* locus is enriched for H3K36me3 over its gene body in muscle nuclei but retains H3K27me3 in neurons. Inversely, neuron-specific *C3H05.6* harbors H3K36me3 over its gene body in neuron nuclei but is marked by H3K27me3 in muscle nuclei (Figure 4.8). Though based on isolated examples, these preliminary observations suggest that some tissue-specific genes indeed harbor tissue-specific active H3K36me3 mark in the tissue(s) where they are expressed, contrary to what was previously suggested (Evans *et al.*, 2016; Pérez-Lluch *et al.*, 2015). More generally, they confirm that the tissue-specific histone modification profiling after nuclei sorting is suitable to investigate the tissue-specific mechanisms of gene regulation.

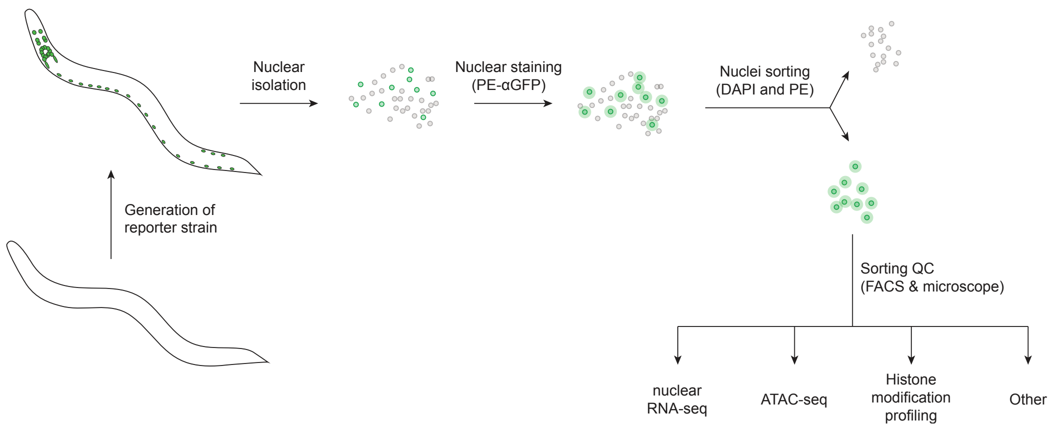

Nuclear isolation
Nuclear staining (PE-αGFP)
Nuclei sorting (DAPI and PE)
Generation of reporter strain
Sorting QC (FACS & microscope)
nuclear RNA-seq
ATAC-seq
Histone modification profiling
Other

Figure 4.9 – Procedure to perform genome-wide assays in individual tissues in *C. elegans*. Note that different promoters can be used to label specific populations of cells. Most genome-wide assays require nuclear material and should be compatible with sorted nuclei.

4.3 Discussion

Compared to other model organisms, obtaining biological material from individual tissues *C. elegans* is challenging, notably due to its small size. Because of this, few tissue-specific chromatin profiling experiments have been conducted yet and the study of mechanisms of tissue-specific gene regulation in *C. elegans* has been hampered. Thus, being able to isolate nuclei from individual tissues is key to study tissue-specific gene regulation. In this chapter, I described the development of a method to isolate highly pure nuclei from individual tissues by nuclear sorting, a necessary prerequisite for obtaining tissue-specific data. With the help of others, I have compared and optimized different types of fluorescent reporters, streamlined the sorting procedure and adapted downstream genome-wide assays to work with low material inputs. The resulting workflow is highly modular (Figure 4.9). For example, someone studying a specific tissue (*e.g.* the pharynx) would only need to clone a new promoter upstream of the existing reporter by Gateway and inject this construct to generate a new tissue-reporter strain, and would subsequently be able to perform all the genome-wide assays previously described.

Compared to other methods of tissue-specific profiling, the sorting-based approach retains several advantages. For example, RNA-tagging is the method generally used to profile tissue-specific transcriptomes, but this approach does not have the same versatility as the sorting-based approach. In this case, the experimenter is limited to studying gene expression and cannot profile chromatin accessibility, for instance. Compared to the INTACT nuclei purification method (Steiner *et al.*, 2012), the purity and the number of sorted nuclei can be finely adjusted during the sorting, and the detection of other parameters such as the diploidy or the size of the nuclei can be used to further select specific populations of nuclei. However, the sorting method remains a low-throughput, laborious task, even with optimized conditions. Obtaining 100,000 nuclei can take between 30 minutes and 2 hours, depending on the percentage of nuclei to isolate among the nuclei preparation. Along with the required immunostaining prior to sorting (~ 1h30 minimum), this may represent a significant increase of the time required to perform an assay, especially compared to other approaches like INTACT nuclei purification, which is more "instantaneous". This needs to be taken into account when planning an experiment.

I also optimized several genomic assays, which can now be performed using low amounts of sorted tissue-specific nuclei. These assays can be used to study the gene regulatory organization (addressed in Chapter 5 and Chapter 6), but also widen the possibilities to tackle other biological questions, such as transcript alternative splicing in different tissues (Gracida *et al.*, 2016), the difference of transcription factor binding patterns across tissues (Reinke *et al.*, 2018) or the tissue-specific chromatin integrity throughout aging (Sen *et al.*, 2016).

An important point to address is the relevance of using a FACS sorting approach, which can only generate populations of nuclei from bulk tissues, when other approaches such as single-cell methods are rapidly emerging. A sorting-based approach still retains several important benefits compared to single-cell approaches. It allows investigation of the chromatin organization by histone modification and/or transcription factor profiling, which is currently impossible or still in early development in single cells. It also remains more cost-effective than single-cell approaches. Finally, the reproducibility and the significance of the results can be experimentally assessed by directly comparing replicates. Batch effects, cell variability and cell annotation algorithms still pose significant challenges for evaluating data reproducibility in single-cell assays (Yuan *et al.*, 2017). However, a sort-based approach is limited to the study of bulk tissues and patterns in small sub-populations cannot be easily identified. On the contrary, results from single-cell methods are much more granular and can lead to a better understanding of the

the variability and heterogeneity of a feature measured in individual cells. Overall, single-cell techniques and sorting-dependent assays are complementary and, in the future, the combined use of the two approaches should help to further improve our understanding of tissue-specific mechanisms of gene regulation (see Chapter 7).

Chapter 5

Gene regulatory architectures in adult *C. elegans*

In Chapter 4, I presented the experimental procedure I developed to isolate nuclei from individual tissues and perform genome-wide assays. In Chapter 5, I describe the gene expression and chromatin accessibility datasets generated from the five main tissues of the worm (germline, neurons, muscle, hypodermis and intestine), at the young adult stage. I then classify accessible sites and genes according to their tissue-specific patterns of accessibility or expression and use the data to characterize distinctive features of gene regulatory architecture of each gene class.

Collaboration note: Yan Dong generated nuclear RNA sequencing libraries using RNA from sorted tissue-specific nuclei which I isolated. Jürgen Jänes used the datasets I generated to identify and functionally annotate accessible chromatin loci, using a pipeline he developed during his PhD (Jänes et al.*, 2018).*

5.1 Profiling tissue-specific transcriptomes and chromatin accessibility

With (i) five different tissue-specific reporter strains, (ii) an optimized sorting procedure and (iii) genome-wide assays adapted for small amounts of sorted nuclei, I was able to profile chromatin accessibility and gene expression across the five main tissues of the worm. I focused on young adult worms, where somatic cells are terminally differentiated and germline has reached maturity.

Table 5.1 – Percentages of paired-end reads mapping to exons, introns and elsewhere, for each replicate of nuclear RNA-seq.

	% of paired-end reads in		
Sample	**exons**	**introns**	**others**
Germline (YA) rep1	82.95	14.113	2.934
Germline (YA) rep2	89.61	9.026	1.368
Neurons (YA) rep1	86.9	12.659	0.445
Neurons (YA) rep2	87.57	11.915	0.511
Muscle (YA) rep1	87.82	12.18	0.003
Muscle (YA) rep2	87.4	11.719	0.877
Hypod. (YA) rep1	73.59	21.738	4.667
Hypod. (YA) rep2	73.88	20.797	5.323
Intest. (YA) rep1	68.34	25.397	6.263
Intest. (YA) rep2	79.2	17.347	3.451

5.1.1 Tissue-specific gene expression

I profiled the nuclear transcriptome of five main tissues of the adult worm: germline, neurons, muscles, hypodermis and intestine. For each sample, I sorted at least one million nuclei and obtained between 10 and 100 ng of nuclear RNA. Libraries were generated in duplicates for each tissue and sequenced in paired-end mode, eventually obtaining ten datasets. I aligned and filtered the sequencing results following a standard procedure described in Jänes *et al.* (2018). Importantly, I obtained the raw expression counts by counting overlapping RNA-seq fragments at the gene level rather than at the transcript level, as immature nuclear transcripts were sequenced and a significant portion of the fragments map to introns (Table 5.1). After transforming the raw counts into "Transcripts Per Million" (TPM), I compared all the samples to each other. I observed well-correlated results between duplicates, suggesting that the nuclei sorting followed by RNA-seq is highly reproducible (Figure 5.1A). Moreover, when browsing the different nuclear RNA-seq tracks at known tissue-specific loci, I could observe expected patterns of expression in each tissue (Figure 5.2A).

An important step in annotating promoters in *C. elegans* relies on the detection of outrons, and previous reports from our group showed that profiling of nuclear RNA, enriched for immature unspliced transcripts, efficiently captures outron signals (1.2.3.1 on page 31 and Chen *et al.*, 2013; Jänes *et al.*, 2018). I browsed my tissue-specific nuclear RNA-seq profiles and could indeed observe clear transcription elongation, originating from known promoters (Figure 5.2A, B).

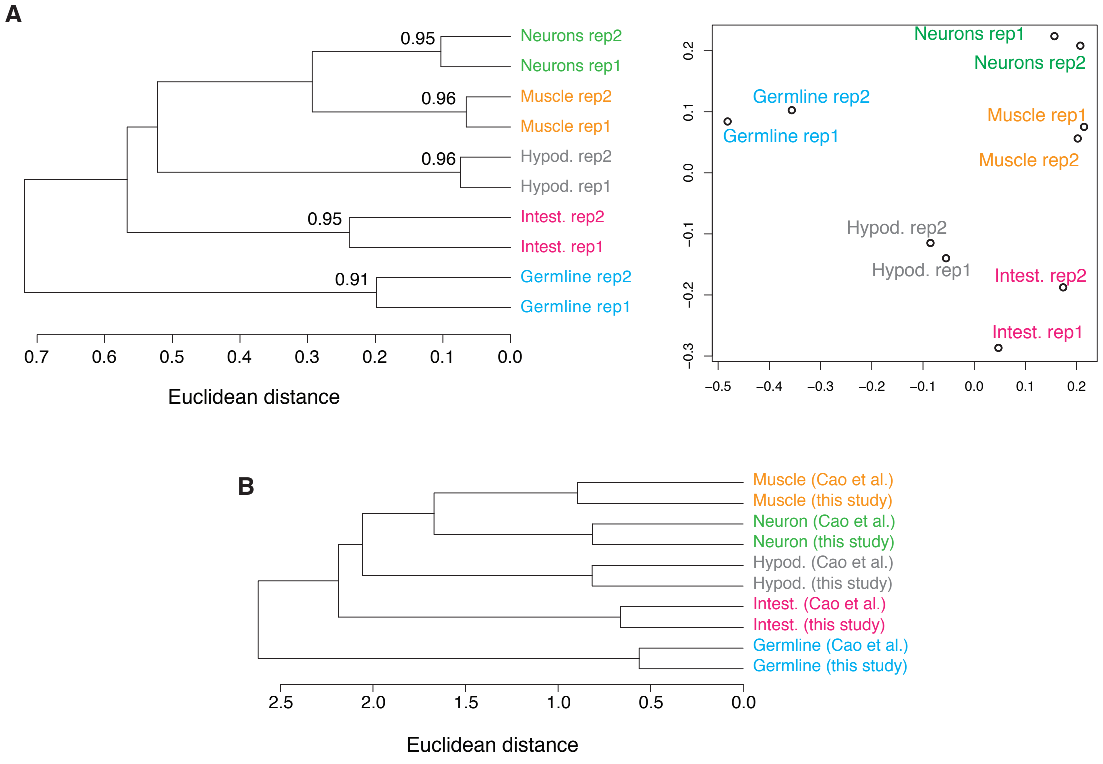


Figure 5.1 – Correlations between tissue-specific nuclear RNA-seq samples. **A-** Euclidean distances between all the tissue-specific nuclear RNA-seq replicates (left). The values on the distance tree are the Pearson correlation scores between two replicates of the same tissue. A Principal Component Analysis (PCA) also shows spatial grouping of duplicates (right). **B-** Distance tree between the tissue-specific gene expression values measured in YA worms (this study) or in L2 worms by single-cell RNA-seq (Cao *et al.*, 2017).

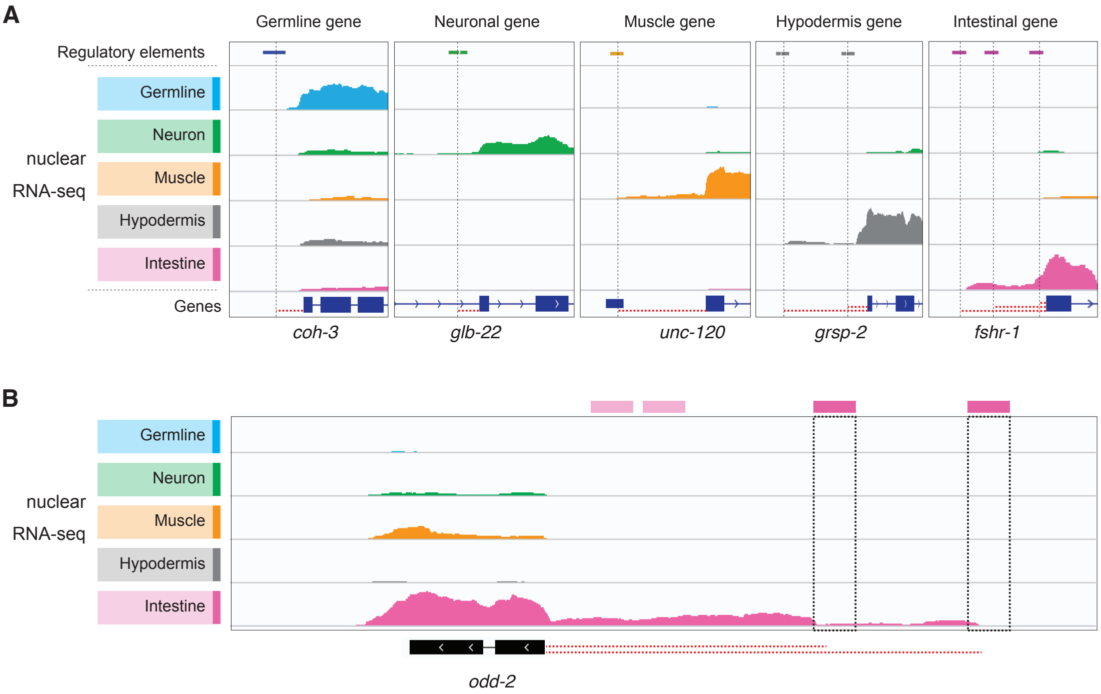


FIGURE 5.2 – Tissue-specific nuclear RNA-seq signals. **A-** Nuclear RNA-seq tracks at five different tissue-specific loci. The annotated promoters are indicated by vertical dotted lines. The scale is the same across the tracks and is adjusted at each locus. **B-** Focus on *odd-2*, an intestine-specific gene. Promoters (solid-colored) and putative enhancers (transparent) are represented on top of the tracks. The outron signals between the promoters and the splice acceptor site (represented by red dotted lines) are readily detected. Note the increase of transcription at each of the two promoters. The scale is the same across the tracks.

I also compared gene expression values derived from tissue-specific nuclear RNA-seq from young adult worms with aggregated gene expression values obtained from single-cell RNA-seq performed in worms at the L2 stage (Cao *et al.*, 2017). I observed that each adult tissue-specific dataset clusters with L2 data from the corresponding tissue. This suggests that my tissue-specific RNA-seq datasets are of high quality (Figure 5.1B).

Together, these observations suggest that the tissue-specific nuclear RNA-seq datasets I generated after nuclei sorting are suitable for genome-wide analysis of tissue-specific gene expression. Moreover, these nuclear RNA-seq datasets constitute an important asset to complete the annotation of promoters in *C. elegans* (see 5.1.2.2 below).

5.1.2 Tissue-specific chromatin accessibility landscapes

5.1.2.1 Quantifying tissue-specific accessibility of regulatory elements

To profile chromatin accessibility in each tissue of the adult nematode, I also performed tissue-specific ATAC-seq experiments. For each sample, I typically used between 25,000 and 50,000 sorted nuclei. I generated two replicates for each tissue and sequenced the libraries in both single-end and paired-end mode (see Chapter 5 for the analysis of paired-end datasets). After transforming raw counts into “Reads Per Million” (RPM), I compared all the samples to each other and observed well-correlated results between replicates, suggesting that the nuclei sorting followed by ATAC-seq is highly reproducible (Figure 5.3).

I then processed the ATAC-seq sequencing datasets into genome-wide browsable tracks representing tissue-specific chromatin accessibility and inspected the tracks at known loci to estimate their quality. HLH-1 is a known transcription factor specifically expressed in muscle cells. The muscle-specific nuclear RNA-seq profile indeed shows transcription signal at the *hlh-1* locus, extending from its promoter approximately 500 bp upstream (Figure 5.4A). The location and tissue-specific activity of this promoter region has been experimentally determined previously (Krause *et al.*, 1994). Importantly, I could observe a specific increase of chromatin accessibility at the expected *hlh-1* promoter region in muscle nuclei, but not in other tissues (Figure 5.4A). Two other regulatory elements present in the first intron of *hlh-1* gene also exhibit muscle-specific accessibility, as expected from previous

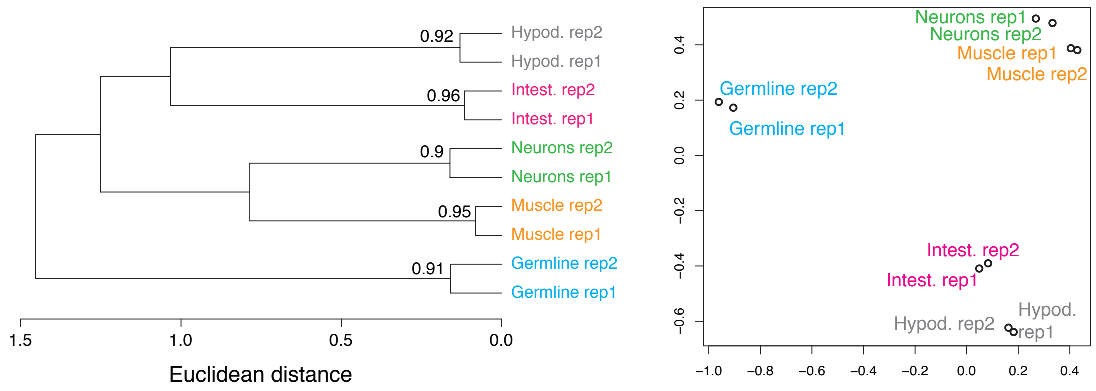


FIGURE 5.3 – Correlations between tissue-specific ATAC-seq samples. Euclidean distances between all the tissue-specific ATAC-seq replicates (left). The values on the distance tree are the Pearson correlation scores between two replicates of the same tissue. A Principal Component Analysis (PCA) also shows spatial grouping of duplicates (right).

promoter dissection studies (Krause *et al.*, 1994). BED-3 is a transcription factor specifically involved in molting regulation and reporter assays have shown that *bed-3* promoter drives transcription specifically in hypodermal cells (Jänes *et al.*, 2018). In my datasets, I observed that, indeed, the promoter associated to *bed-3* is only accessible in hypodermal nuclei. Here again, two putative *bed-3* enhancers are accessible specifically in hypodermis, similar to the *bed-3* promoter. More generally, the comparison of the two types of datasets at known tissue-specific loci reveals a strong concordance between chromatin accessibility signals and tissue-specific patterns of gene expression (Figure 5.4C).

These observations suggest that the tissue-specific ATAC-seq datasets generated after nuclei sorting are suitable for genome-wide analysis of tissue-specific gene regulation. Importantly, the ATAC-seq and nuclear RNA-seq profiles match each other, with tissue-specific transcription initiation readily detected at promoters active in the corresponding tissue(s).

5.1.2.2 Completing the functional annotation of accessible loci in *C. elegans*

Transcriptome and chromatin accessibility profiles throughout *C. elegans* life previously allowed to (i) map and (ii) functionally annotate 42,245 chromatin loci (see 1.2.3.1 on page 31 and Jänes *et al.*, 2018). With the new tissue-specific datasets I generated, additional accessible loci could potentially be identified and annotated.

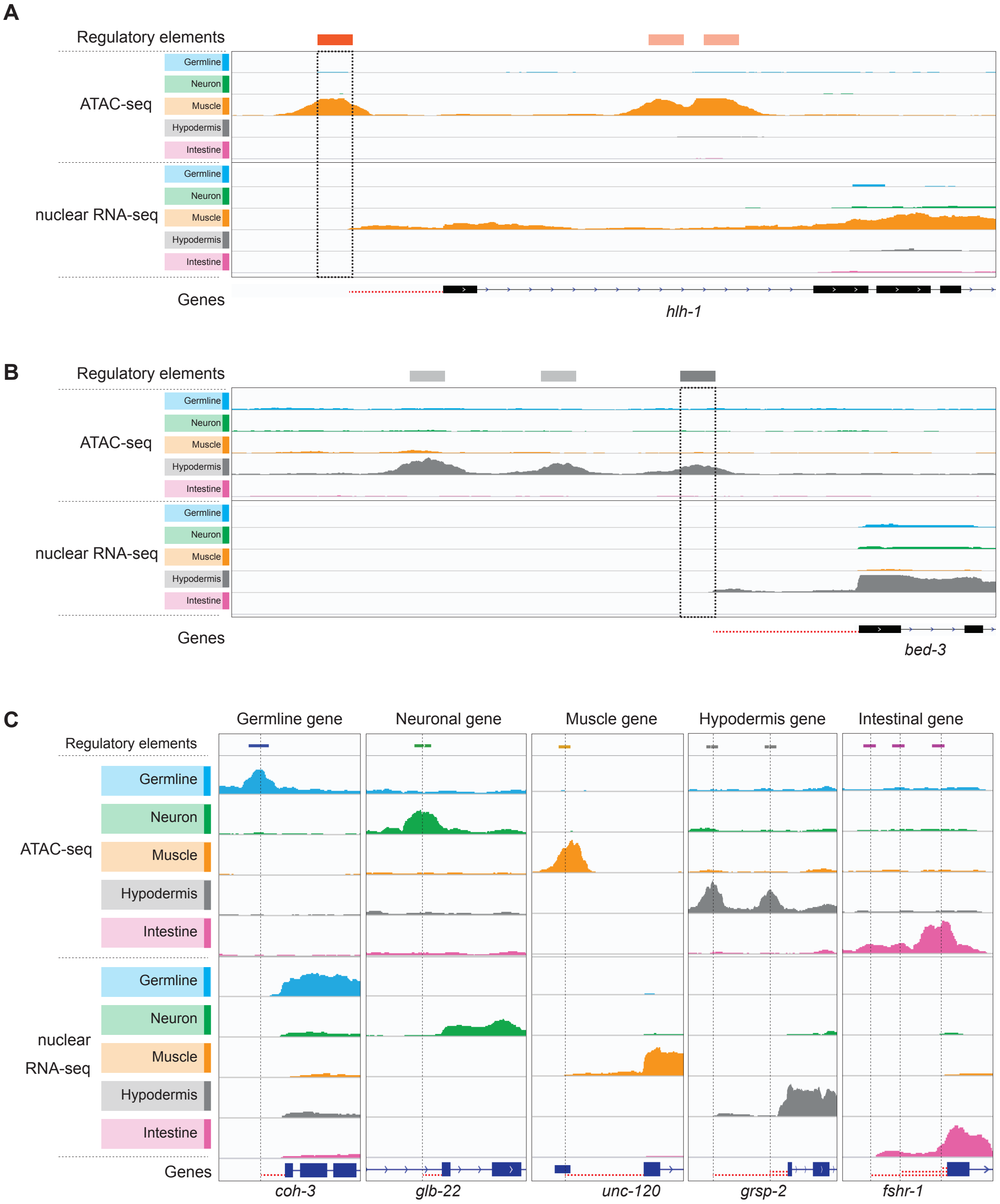


Figure 5.4 – Tissue-specific ATAC-seq signals and accordance with tissue-specific ATAC-seq signals. **A-** *hlh-1*, a known muscle-specific locus. **B-** *bed-3*, a known hypodermis locus. Promoters (solid-colored) and putative enhancers (transparent) are represented on top of the tracks. **C-** Accordance between ATAC-seq and RNA-seq signals at different known tissue-specific loci. The red dotted lines represent trans-spliced outrons. In each panel of this figures, the scale is the same across the five ATAC-seq tracks and across the five RNA-seq tracks, and is adjusted at each locus.

TABLE 5.2 – Original and new accessible sites. The original accessible sites have been mapped and annotated in Jänes *et al.* (2018). The new annotated sites have been obtained by including the tissue-specific ATAC-seq and RNA-seq datasets to the original ones and re-running the annotation pipeline.

Accessible sites	**# of original sites**	**# of new sites**	**Increase of # of sites**
Forward promoters	5,863	384	+6.5 %
Reverse promoters	5,757	352	+6.1 %
Bi-directional promoters	1,976	31	+1.6 %
Putative enhancers	19,231	2,218	+11.5 %
Non-coding RNA	824	20	+2.4 %
Pseudogene promoters	291	23	+7.9 %
Unknown promoters	1,791	91	+5.1 %
Other elements	6,512	2,150	+33.0 %

Jürgen Jänes first integrated my tissue-specific ATAC-seq datasets with the previously generated developmental and aging datasets and ran the mapping pipeline he had developed during his PhD. This led to the identification of 5,269 additional chromatin accessible sites in *C. elegans* genome, for a total of 47,514 sites. He then integrated my tissue-specific nuclear RNA-seq datasets with the previously generated developmental and aging datasets and ran his annotation pipeline. This let to the functional annotation of 3,119 of the 5,269 new loci, with notably 767 new promoters and 2,218 new putative enhancers (Table 5.2). For instance, a new promoter was annotated at the *shk-1* locus, on top of the three already annotated ones (Figure 5.5). At this promoter, transcription clearly initiates specifically in neurons, whereas the three promoters previously annotated are characterized by either muscle-specific or muscle and neuron-specific transcription initiation. Of note, the other 2,150 new loci do not overlap any transcriptional signals and remain uncharacterized.

Thus, the tissue-specific datasets that I generated provided useful information to improve the annotation of regulatory elements in *C. elegans* genome.

5.2 Classification of regulatory elements and genes

Mechanisms of tissue-specific gene regulation remain poorly investigated in *C. elegans*. With the tissue-specific RNA-seq and ATAC-seq datasets I generated, I aimed at annotating gene and regulatory element usage in the five main tissues of the nematode. This would help investigating how patterns of tissue-specific gene

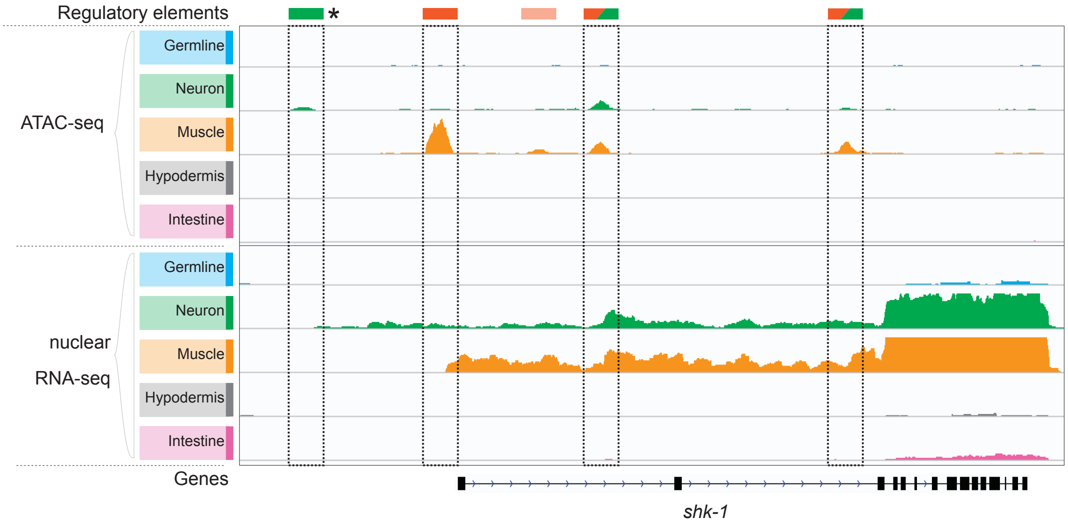


FIGURE 5.5 – Example of a newly annotated accessible site at the *shk-1* locus. Promoters (solid-colored) and putative enhancers (transparent) are represented on top of the tracks. The new promoter (labelled with an asterisk) has been annotated using the tissue-specific datasets. The color of the bars indicate in which tissue(s) the site is accessible (orange: muscle; green: neurons). Both tissue-specific ATAC-seq and nuclear RNA-seq profiles are displayed. The scale is the same across the five ATAC-seq tracks and across the five RNA-seq tracks.

expression are obtained.

5.2.1 Diversity of tissue-specific patterns of gene expression

5.2.1.1 Classification of genes into ubiquitous and tissue-specific classes

I leveraged the tissue-specific nuclear RNA-seq datasets generated in young adult worms to classify each of the 20,222 protein-coding genes annotated in *C. elegans* genome into different classes. I compared expression changes between all possible pairs of tissues ($\binom{5}{2} = 10$ pairs) using DESeq2 (Love *et al.* (2014)). Genes with an increase or decrease of expression between two tissues higher than 3-fold and with an adjusted p-value < 0.01 were considered significantly differently expressed (DE). Of note, within the 20,222 protein-coding genes, many are not expected to be expressed in young adults. Thus, 5,575 genes (28%) had an expression lower than 5 TPM across all of the five tissues in young adult and were annotated as inactive. In the remaining 14,647 protein-coding genes, the tissue specificity was determined according to the following successive rules:

1. *Genes specifically active in a single tissue:* genes (i) significantly DE between

the first and second most expressed tissue and (ii) not significantly DE between the second and the third most expressed tissues.

2. *Genes restricted to two tissue*s: genes (i) significantly DE between the second and the third most expressed tissues and (ii) not significantly DE between the third and the fourth most expressed tissues.

3. *Genes restricted to three tissues:* genes (i) significantly DE between the third and the fourth most expressed tissues and (ii) not significantly DE between the fourth and the fifth most expressed tissues.

4. *Genes restricted to four tissues:* genes significantly DE between the fourth and the fifth most expressed tissues.

5. *Ubiquitous-biased genes:* genes (i) significantly DE between any other pair of tissues (e.g. first and fourth most expressed tissue) and (ii) detected across all tissues (RPM $>$ 5 in all tissues).

6. *Ubiquitous-uniform (also simply referred to as "uniform") genes:* genes (i) not significantly DE between any pair of tissues and (ii) detected across all tissues (RPM $>$ 5 in all tissues).

7. *Unclassified genes:* genes with expression $<$ 5 RPM in some tissues and not significantly DE could not be confidently classified (n = 2,346).

Using this approach, almost half of the classified genes (48%) were ubiquitously expressed (Figure 5.6, 28% uniformly expressed and 20% with biased ubiquitous expression). The rest of the classified genes were either expressed in a single tissue (32%) or a subset of tissues (20%).

I observed that the nuclear RNA datasets have minor contamination, likely originating from bulk cytoplasmic RNA released during nuclear isolation. This resulted in tissue-specific genes with high expression (*e.g.* muscle myosin gene *unc-54*) being classified as "ubiquitous-biased". From this point forward, when studying ubiquitous genes and chromatin loci, I specifically focus on the "ubiquitous-uniform" class and for simplicity refer to them as "ubiquitous"

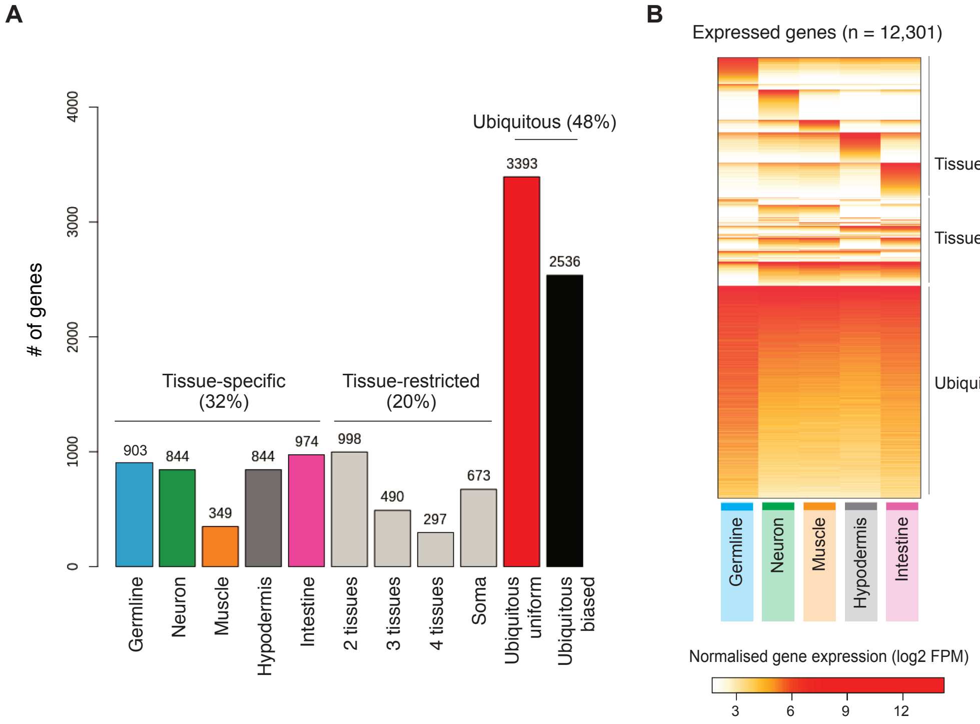


FIGURE 5.6 – Classification of expressed genes. **A-** Distribution of expressed protein-coding genes in tissue-specific, tissue-restricted and ubiquitous classes. The 27 tissue-restricted classes have been merged into "2-tissues", "3-tissues" and "4-tissues" for simplicity. **B-** Heatmap of gene expression values in each tissue dataset. The genes are ordered by their class.

5.2.1.2 Validation of the tissue-specific and ubiquitous gene sets

Gene Ontology term enrichment analysis is an approach traditionally used to assess the biological meaningfulness of gene sets. I performed GO analyses on the main classes of genes and observed an enrichment of terms relevant to each set of genes (*e.g.* synaptic vesicle transport in neuron-specific gene set, contractile fibers in muscle-specific gene set or cuticle in hypodermis-specific gene set, Figure 5.7). This suggested that the different sets of genes obtained by classification of their expression indeed reflect functional groups of genes involved in tissue-specific biological processes.

A useful metric to assess variability of expression across different conditions is the Coefficient of Variation (CV, Gerstein *et al.*, 2014). Genes expressed in a single tissue have a high CV whereas those expressed in multiple tissues have a lower CV. I measured CV values in each of my gene sets. As expected, tissue-specific genes have a high CV and ubiquitous genes have a very low CV (Figure 5.8). Moreover, genes restricted to two, three or four tissues are characterized by decreasing CV

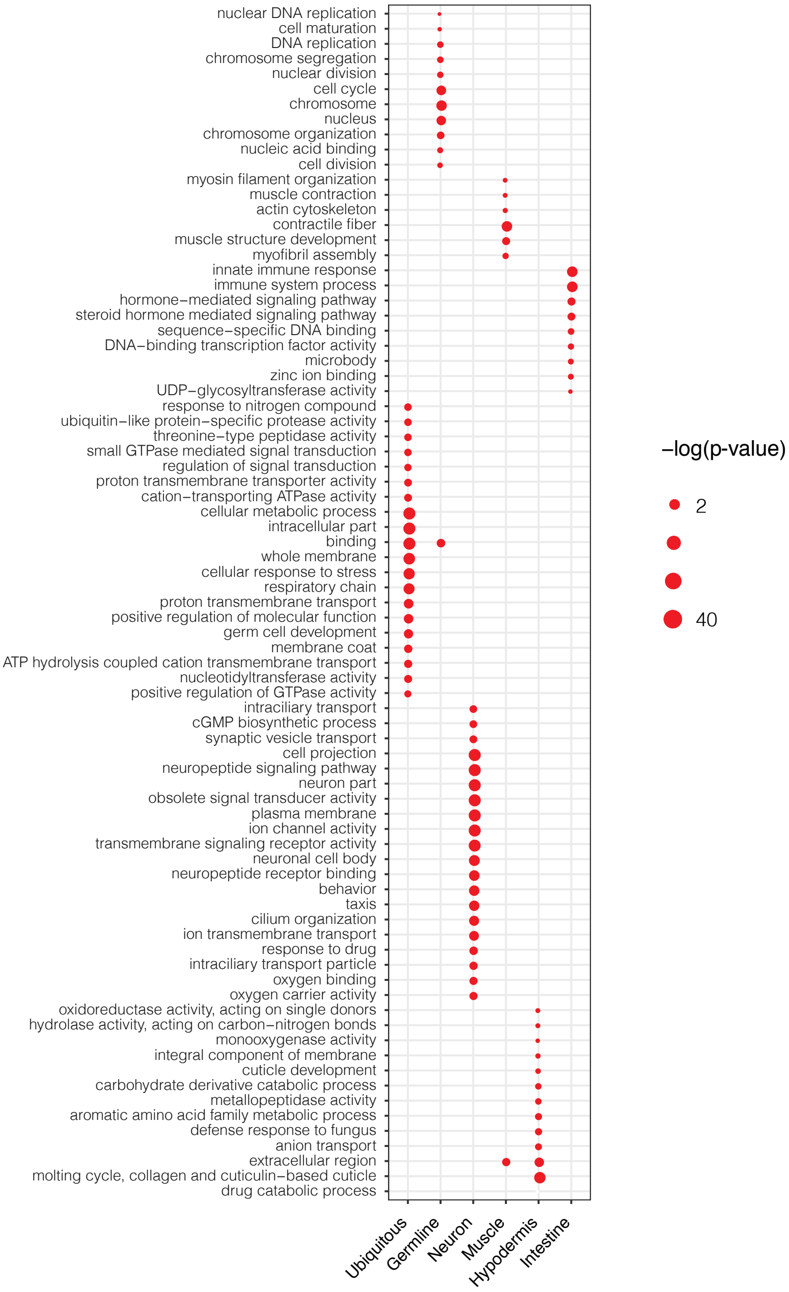


FIGURE 5.7 – GO terms enriched in the main classes of genes.

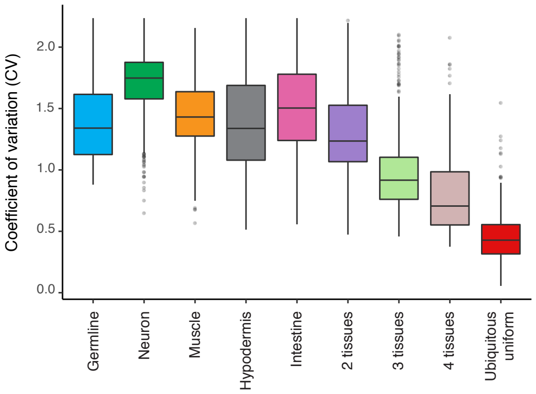


Figure 5.8 – Coefficient of Variation (CV) of expression across tissues for each gene expression class. The 27 tissue-restricted classes have been merged into "2-tissues", "3-tissues" and "4-tissues" for simplicity.

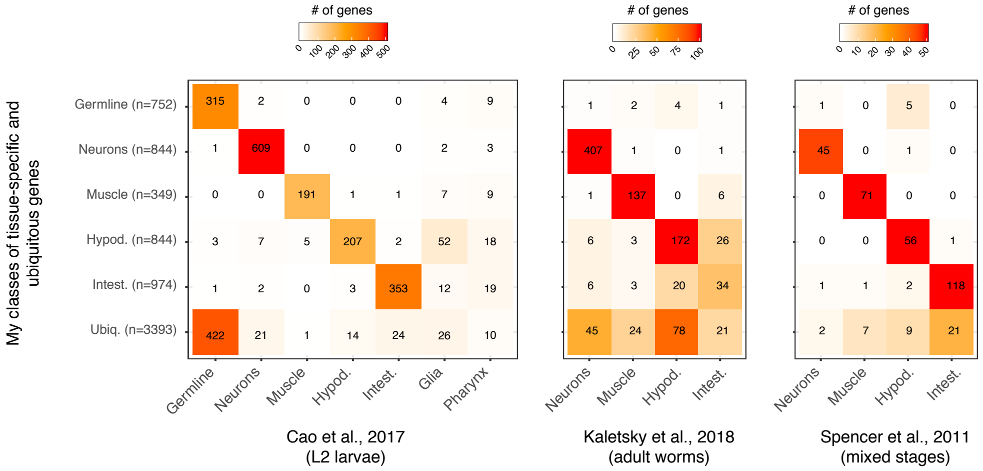


Figure 5.9 – Comparison of gene expression classes with previously published gene sets from Cao *et al.*, 2017; Kaletsky *et al.*, 2018; Spencer *et al.*, 2011. Only my main tissue-specific and ubiquitous gene classes are displayed. Note that none of these studies formally define a class of ubiquitous genes.

values lower than those of tissue-specific tissues but higher than those of ubiquitous genes. This suggests that the classification method used to generate these classes accurately identifies genes expressed across subsets of tissues.

Several tissue-specific transcriptomes have already been generated in *C. elegans* using different approaches (see Table 1.3 on page 42). To validate my classification method, I compared my sets of gene annotations to those from previous studies which annotated gene expression across multiple tissues (Cao *et al.*, 2017; Kaletsky *et al.*, 2018; Spencer *et al.*, 2011). This revealed a good intersection between my datasets and previously published tissue-specific annotations (Figure 5.9). Importantly, I profiled gene expression in most of the cells from the adult worm, where the germline is developed and functional and all the other tissues are terminally

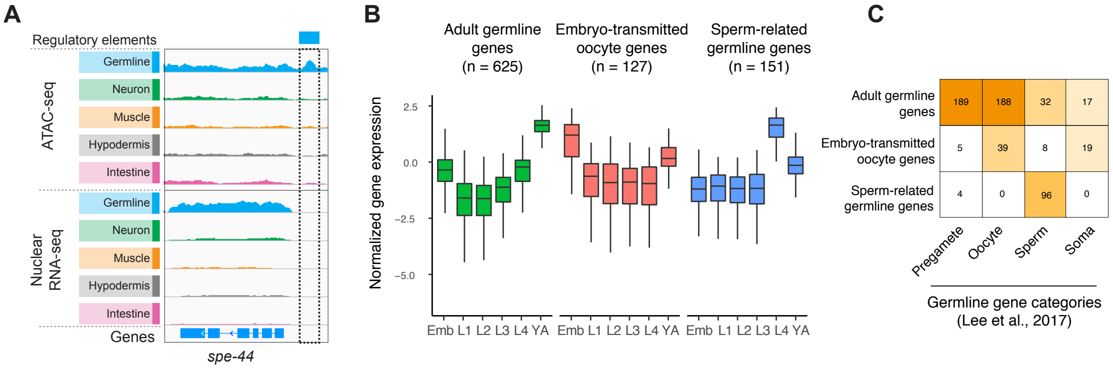


FIGURE 5.10 – Classification of germline genes in subcategories. **A-** Example of *spe-44*, a known sperm-specific gene detected in my germline ATAC-seq dataset generated at the YA stage. **B-** Three germline gene subcategories based on temporal gene expression. **C-** Intersection of the three germline gene subcategories with sperm-specific, pregamete-specific, oocyte-specific and soma genes annotated in Lee *et al.* (2017). Only the relevant germline genes are shown here.

differentiated, and I was thus able to define both tissue-specific and ubiquitous sets of genes.

When investigating the results of my gene classification method, I observed that the germline set contained several sperm-related genes (*e.g. spe-44*, Figure 5.10A). Sperm-specific genes are activated during spermatogenesis at the L4 stage and are largely down-regulated in young adults, but may still be detected in my datasets as some L4 worms are present in the collection of adult worms. Using transcriptome profiles generated across development (Jänes *et al.*, 2018), I could subdivide my original set of 903 germline genes into three subcategories: 625 adult germline genes (whose expression peak at YA stage), 127 germline genes enriched in oocytes and inherited in embryos (whose expression peak in embryos) and 151 sperm genes (whose expression peak at L4 stage, Figure 5.10B). These subcategories significantly overlap with previously annotated pregamete, oocyte and sperm genes (Figure 5.10C and Lee *et al.*, 2017). In subsequent analysis, the set of germline-specific genes only refers to those highly expressed in adult germline (*i.e.* the 625 adult germline genes and the 127 oocyte genes).

5.2.2 Diversity of tissue-specific patterns of chromatin accessibility

I leveraged the tissue-specific ATAC-seq datasets generated in young adult worms to classify the 47,514 accessible loci annotated in *C. elegans* genome into different classes. I used the rules described in the previous section, with a threshold of 8 RPM. I successfully classified 25,205 accessible sites into five tissue-specific classes (germline, neuron, muscle, hypodermis or intestine), 27 tissue-restricted classes (*e.g.* sites active in both neurons and muscles, or in all somatic tissues but not in germline, etc) and two ubiquitous classes (containing sites with either uniform or non-uniform accessibility across tissues) (Figure 5.11). Interestingly, chromatin accessibility is overall largely tissue-specific (Figure 5.11A). A majority of the classified sites (56%) are only accessible in a single tissue, with 22% other sites having tissue-restricted and only 22% showing ubiquitous accessibility. Many tissue-specific regulatory elements have been observed from bulk tissue-specific ATAC-seq in fly (Liu *et al.*, 2019a) and single-cell ATAC-seq studies in fly and mouse (Cusanovich *et al.*, 2018a,b), but they did not always represent the major part of all the regulatory elements (10 to 40% of the accessible loci are reported to be differently accessible across tissues of the adult mouse, Cusanovich *et al.*, 2018b; Liu *et al.*, 2019a). This could originate from the higher complexity of mammalian tissues or from the highly heterogenous nature of single-cell datasets, making the identification of tissue-specific accessibility much more complex.

The remaining 22,309 (47%) sites were not classified, either because they are not detected in my datasets or not accessible enough to confidently classify them. They could represent genomic loci accessible in early developmental stages but inactive in young adult. To check this, I looked at their accessibility across development. While the successfully classified sites are generally highly accessible at the YA stage (when tissue-specific accessibility is assessed), the unclassified sites have a decreased accessibility after the embryonic stage (Figure 5.12). Thus, these sites are preferentially accessible in embryos.

Using a nuclei sorting approach presents several limitations. Fist, intra-tissue heterogeneity cannot be determined. For instance, gene expression in different types of neurons is controlled by specific regulatory elements but isolating the whole population of neuron nuclei does not allow to distinguish or estimate this diversity. For instance, *unc-30* is a neuronal gene highly expressed specifically in GABAergic neurons; pan-neuron tissue-specific datasets cannot resolve this intra-tissue specificity (Figure 5.13). Moreover, lowly expressed (accessible) genes (chromatin) may be under the threshold of detection of ATAC-seq and RNA-seq. For example, *unc-55* is another neuron-specific gene, expressed in the same

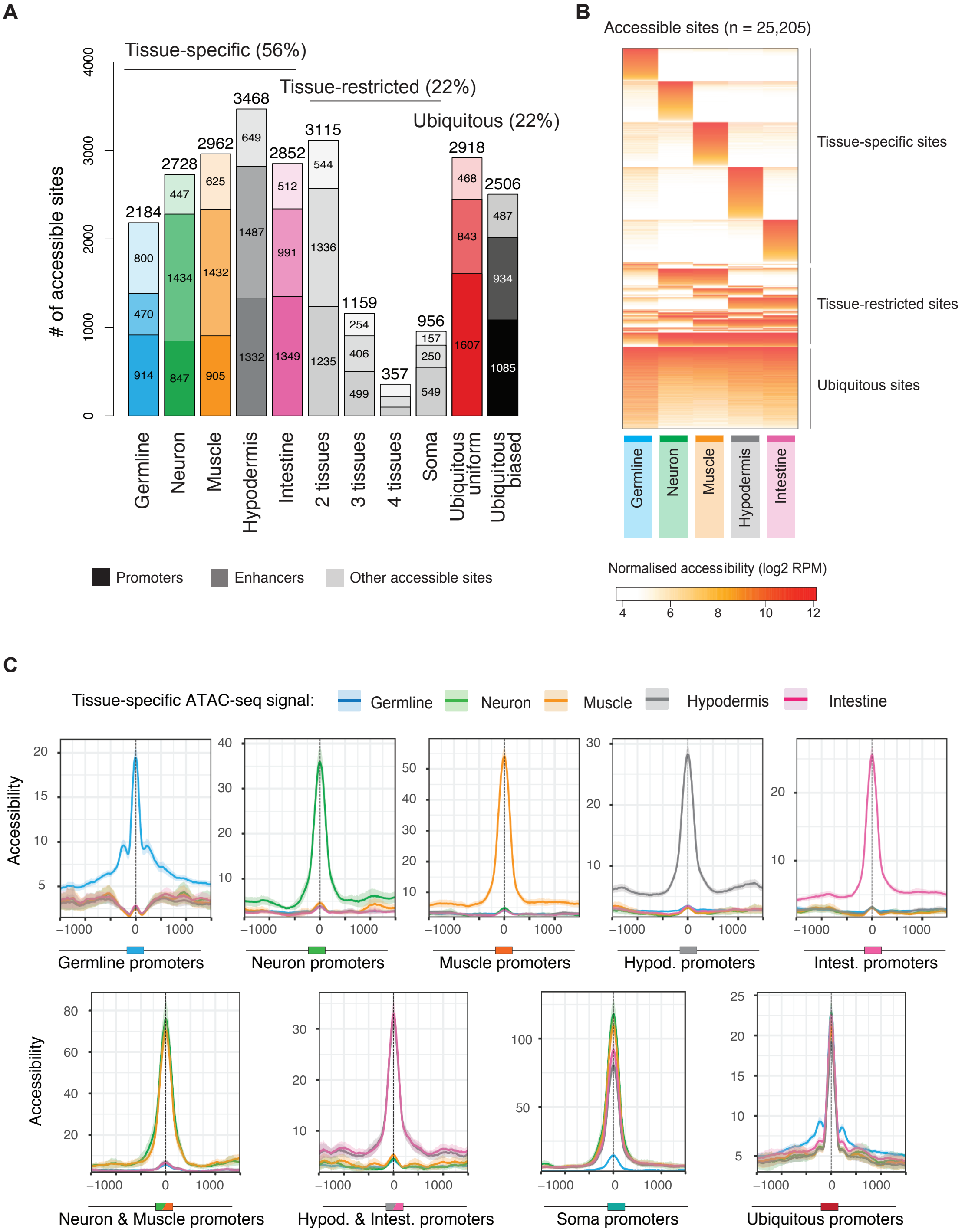


Figure 5.11 – Classification of accessible chromatin loci. **A-** Distribution of accessible sites (split into promoters, putative enhancers and others) in tissue-specific, tissue-restricted and ubiquitous classes. The 27 tissue-restricted classes have been merged into "2-tissues", "3-tissues" and "4-tissues" for simplicity. **B-** Heatmap of accessibility values in each tissue dataset. The accessible sites are ordered by their class. **C-** Tissue-specific accessibility signal over different sets of promoters.

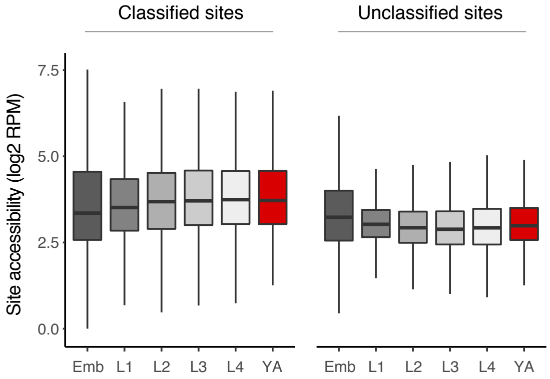


FIGURE 5.12 – Comparison of developmental accessibility for loci successfully classified into tissue-specific, tissue-restricted or ubiquitous classes (left) or for unclassified loci (right). Note the decrease of accessibility during development for the unclassified loci, with accessibility generally lower at the YA stage, compared to that of successfully classified loci.

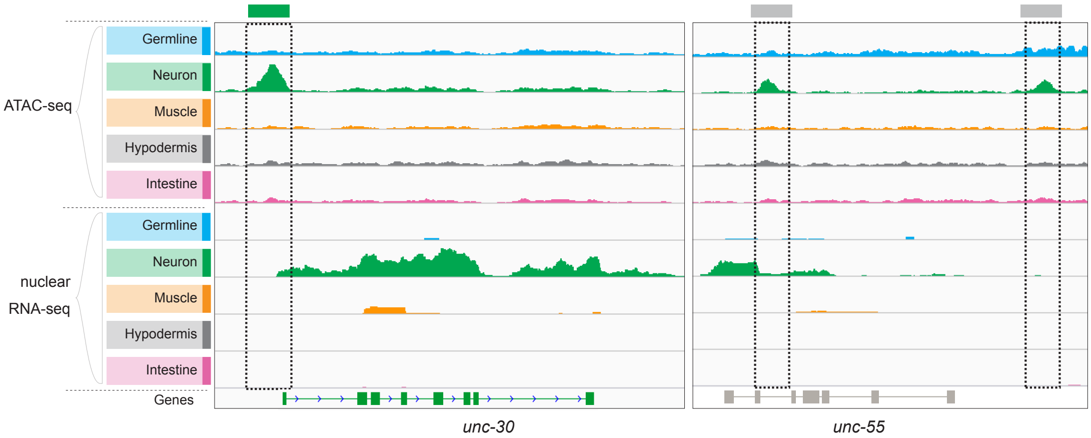


FIGURE 5.13 – Neuron-specific genes and promoters are sometimes missed. *unc-30* (left) and *unc-55* (right) are two genes specifically expressed in GABAergic neurons (the former is highly expressed while the latter is less expressed). *unc-30* promoter (green bar) is accurately classified as a neuron-specific promoter while the two *unc-55* promoters (light gray bars) could not be classified. The scale is the same across the five ATAC-seq tracks and across the five RNA-seq tracks, and the same across the two genomic loci.

GABAergic neurons than *unc-30* but to lesser levels. While the *unc-30* promoter was successfully classified as a neuron-specific promoter, neither of its the two *unc-55* promoters could be classified using tissue-specific ATAC-seq data. When studying specific subsets of cells such as the GABAergic neurons, using a new reporter strain to directly sort nuclei from these cells could overcome these limitations.

5.3 Two distinctive gene regulatory architectures

According to the aforementioned classification, ~ 80% of the open chromatin loci are accessible in a subset of tissues or in a single tissue, compared to only ~ 50% of the genes (Figure 5.6 and Figure 5.11). This difference raised the question of how regulatory elements are combined to regulate gene expression. Do tissue-specific regulatory elements generally control tissue-specific genes and ubiquitous ones control ubiquitous genes? Or is ubiquitous expression achieved by combining different tissue-specific regulatory elements together? What is the transcriptional effect of alternative promoters? I used the classifications described hereabove to investigate the organization of tissue-specific and ubiquitous genes and to decipher their regulatory grammars.

5.3.1 Functional groups of ubiquitous genes are differently structured

I first focused on dissecting the structural organization of ubiquitous genes. I observed that among the ubiquitous genes with annotated promoter(s), 45% of them have at least two promoters, which is significantly more than for other classes of gene (Figure 5.14A, 1.53-fold higher, p-value 2.75e-17). I wondered whether structural features were specific to ubiquitous genes with one or three or more promoters. I found that ubiquitous genes with only one promoter had fewer enhancers than those with alternative promoters (Figure 5.14B). The promoters associated with one-promoter ubiquitous genes are more often bidirectional while those associated with ubiquitous genes with alternative promoters are generally unidirectional (Figure 5.14C). Finally, one-promoter ubiquitous genes have fewer and shorter introns than those with alternative promoters (Figure 5.14D-E).

Overall, this revealed extensive structural differences between ubiquitous genes

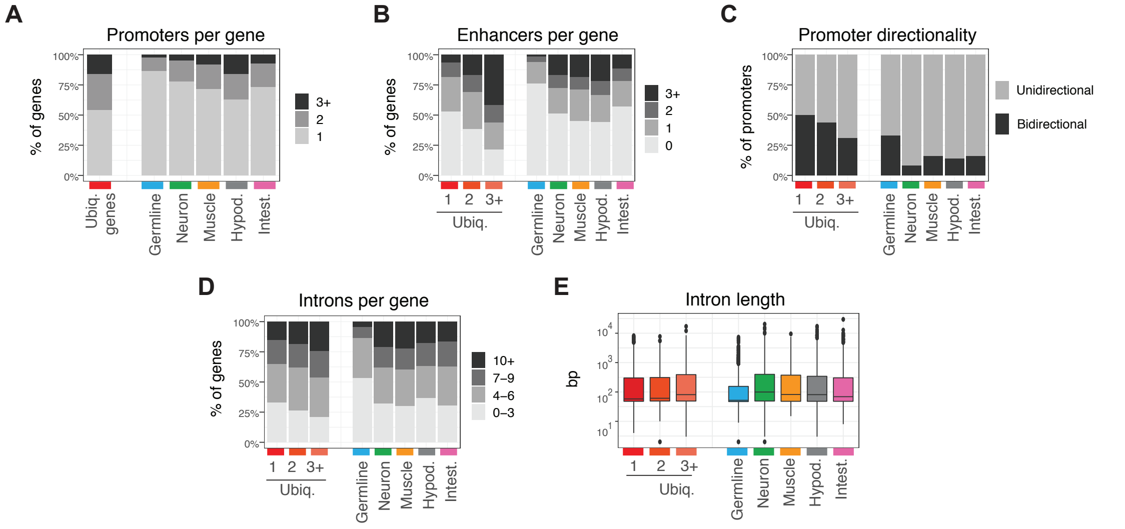


Figure 5.14 – Structural characteristics of ubiquitous, germline-specific or somatic-tissue-specific genes. Ubiquitous genes are split into genes with one, two or three or more promoters. **A-** Number of promoters per gene. **B-** Number of enhancers per gene. **C-** Directionality of the promoters associated to each gene. **D-** Number of introns per gene. **E-** Length of introns. Only the genes with at least one classified promoter are considered.

controlled by a single promoters and those controlled by alternative promoters. This prompted me to investigate the function of ubiquitous genes with only one promoter, two promoters or three or more promoters. I performed GO enrichment analysis for each set and found that ubiquitous genes with a single promoter are enriched for basal cellular processes such as RNA processing, peptide transport and ribonucleoprotein complex biogenesis. In contrast, ubiquitous genes with three or more alternative promoters are involved in more complex processes such as embryo development, regulation of signaling and cellular response to stress (Figure 5.15).

5.3.2 Germline-specific and soma-restricted genes have distinctive regulatory structure

5.3.2.1 Gene structure is associated with tissue-specific patterns of expression

I then investigated the structure of tissue-specific genes and found striking differences in the organization of germline genes compared to those restricted to somatic tissues. Germline genes very rarely have alternative promoters or associated enhancers (Figure 5.14A-B), their promoters are often bidirectional (Figure 5.14C) and they have few short introns (Figure 5.14D-E). In contrast, somatic-tissue-specific

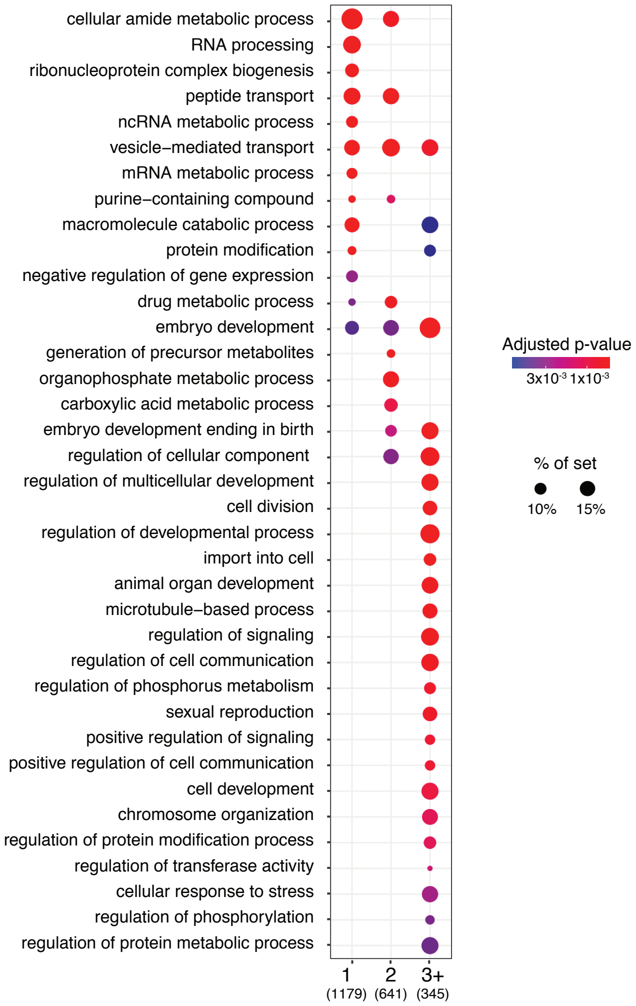


Figure 5.15 – GO enrichment analysis of ubiquitous genes with one, two or three or more promoters

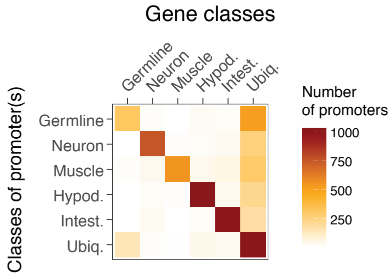


FIGURE 5.16 – Classes of promoters associated with genes of each expression class. Only the major classes of genes and promoters are displayed.

genes have more alternative promoters, are frequently associated to several enhancers, their promoters are largely unidirectional and they have many more and longer introns (Figure 5.14). Thus, germline genes resemble ubiquitous genes with one promoter while somatic-tissue-specific genes resemble ubiquitous genes with multiple promoters.

5.3.2.2 Germline and ubiquitous promoters of germline-specific genes

As expected, most tissue-specific genes are associated with promoter(s) specifically active in the corresponding tissue (Figure 5.16). However, I observed that a group of genes with germline-specific expression have ubiquitously accessible promoters (Figure 5.16, bottom left corner). To investigate whether these genes have different cellular functions than germline genes with germline promoters, I performed GO term enrichment analyses. This revealed that the germline genes with only germline promoters are involved in gamete generation and reproduction, while those with only ubiquitous promoters are involved in cell division and maintenance of DNA integrity (Figure 5.17A). Besides, I found that many of the ubiquitous promoters associated with germline-specific genes are targets of Rb/DREAM, a transcriptional repressor complex (13-fold enrichment, p-value = 5e-13, Figure 5.17B). This supports a model whereby a group of genes with ubiquitously active promoters are predominantly expressed in the germline at the young adult stage via their silencing by the DREAM complex in somatic tissues (Petrella *et al.*, 2011; Wu *et al.*, 2012).

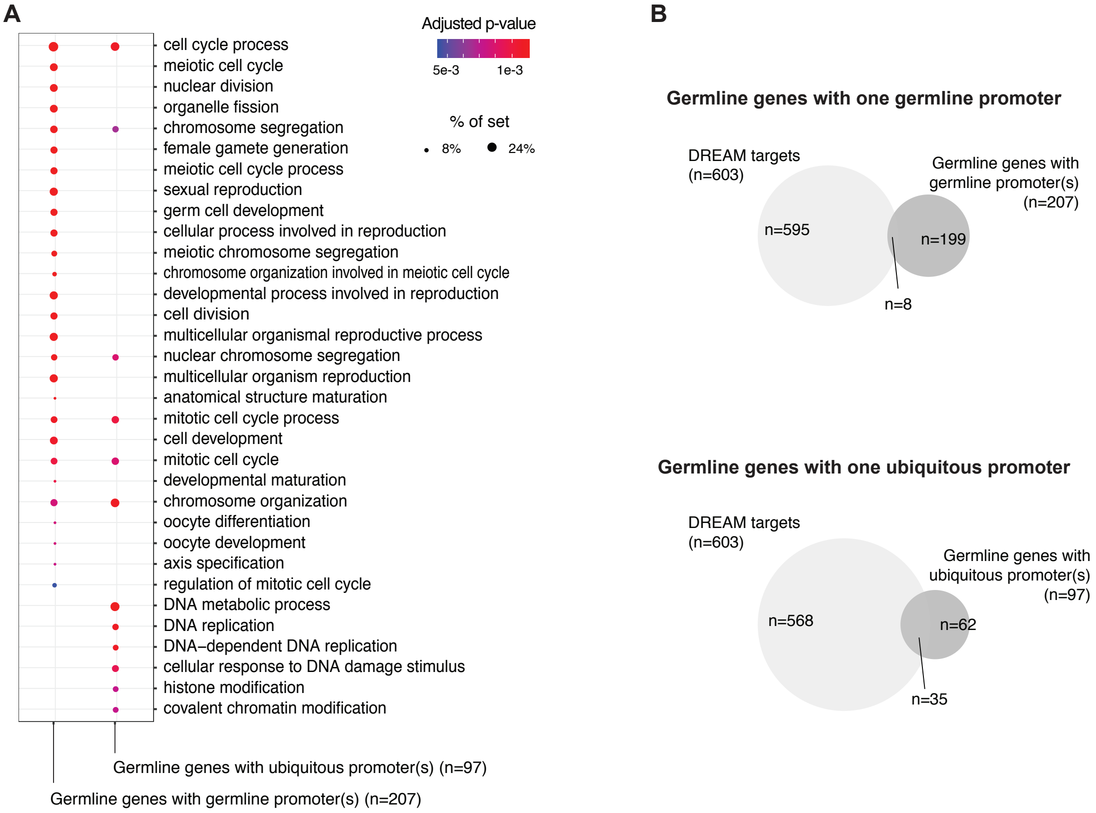


Figure 5.17 – GO enrichment analysis of germline-specific genes with germline or ubiquitous promoters. **A-** GO enrichment analysis of germline-specific genes with only germline promoters or only ubiquitous promoters. **B-** Intersection of the two sets of germline genes with known targets of the DREAM complex (Latorre *et al.*, 2015).

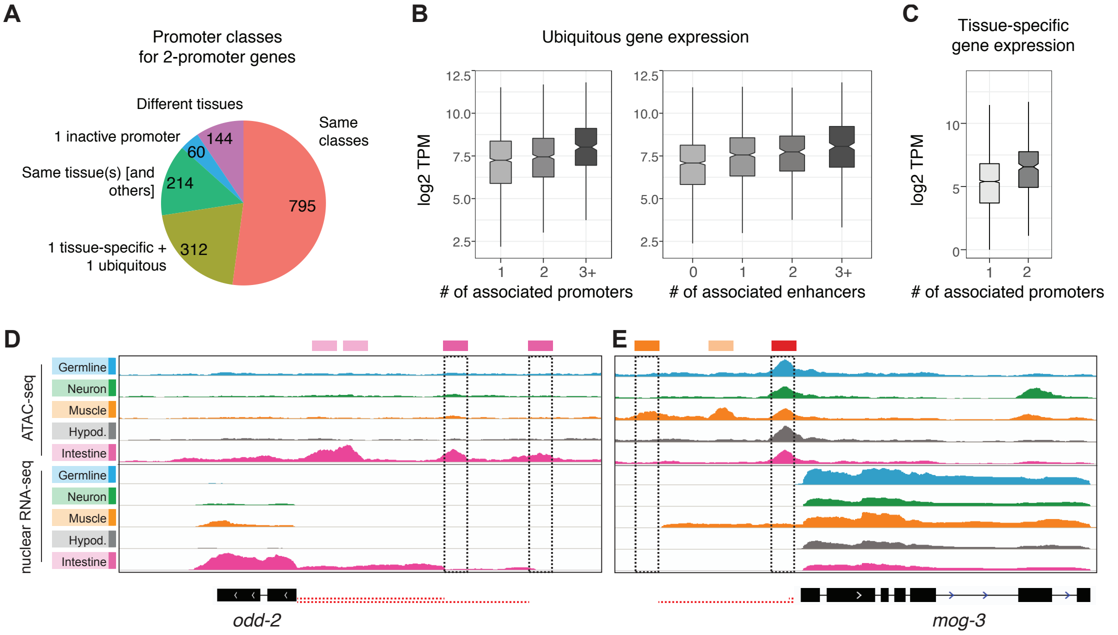


Figure 5.18 – Alternative promoters and regulation of gene expression. **A-** Concordance of promoter classes for genes with two promoters. **B-** Gene expression levels in whole young adults for ubiquitous genes with one, two, or three or more promoters (left), and zero, one, two, or three or more enhancers (right). **C-** Gene expression levels of tissue-specific genes with one promoter or two promoters specifically active in the same tissue. **D-** Example of a tissue-specific gene with multiple tissue-specific promoters (here *odd-2*, an intestine gene with two intestine-specific promoters). **E-** Example of a ubiquitous gene with a ubiquitous promoter and a tissue-specific promoter (here *mog-3*, with one ubiquitous and one muscle-specific promoter). Promoters (solid-colored) and putative enhancers (transparent) are represented on top of the tracks. The red dotted lines represent trans-spliced outrons. The scale is the same across the five ATAC-seq tracks and across the five RNA-seq tracks, and is adjusted at each locus.

5.3.3 The role of alternative promoters

Despite some variability across gene sets, 30% of all genes with at least one annotated promoter have alternative promoters. I wondered whether alternative promoters would rather be redundant or complementary (*i.e.* active in the same set of tissues or in different tissues). I observed that more than half of the genes with two alternative promoters have their two promoters in the same class, whereas less than 10% had two alternative promoters active in different non-overlapping sets of tissues (Figure 5.18A).

To understand the transcriptional impact of multiple promoters active in the same tissue(s), I examined the relationship between the number of regulatory elements and gene expression level. Among ubiquitously expressed genes, I found that the number of promoters and enhancers is positively correlated with gene expression (Figure 5.18B). Similarly, tissue-specific genes with two tissue specific promoters have higher gene expression levels than those with only one (Figure 5.18C). At individual loci, the nuclear RNA-seq signal originating from the two alternative promoters highlights the respective contribution of each promoter (*e.g.* at *odd-2* locus, Figure 5.18D).

Furthermore, I also noted that 15% of the ubiquitously expressed genes with two promoters have one tissue-specific promoter in addition to a ubiquitously active one. In such context, the additional tissue-specific promoter can lead to an increased expression of a ubiquitous gene in the corresponding tissue (*e.g.* *mog-3* in muscle nuclei, Figure 5.18E).

Taken together, these observations suggest that alternative promoters primarily play an additive role in the regulation of expression levels, rather than in increasing the number of tissues in which genes are expressed. In the case of ubiquitous genes, this can lead to increased expression in individual tissues.

5.3.4 Two emerging classes of gene regulatory architecture

Overall, these observations suggest that both fundamental cellular processes and germline functions are encoded by genes with a basic structure resembling that of simpler organisms such as yeast, with few regulatory elements and few introns. On the contrary, genes involved in developmental and tissue-specific processes have a more complex structure, with many regulatory elements and longer introns. For these genes, additional enhancers and alternative promoters may finely regulate their levels of transcription.

These results also suggest that an important role of alternative promoters is to increase gene expression rather than being necessary for its expression *per se*. This could explain some cases where the deletion of an individual regulatory element does not have an obvious effect on gene expression, despite the regulatory element having transcriptional activity in transgenic assays (Catarino and Stark, 2018; Dukler *et al.*, 2016).

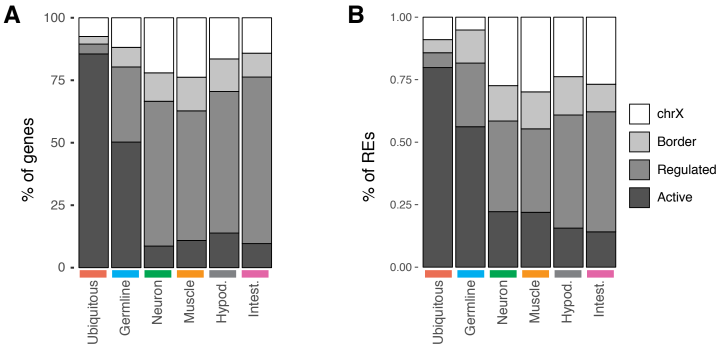


FIGURE 5.19 – Distribution of genes and regulatory elements in different types of chromatin domains. **A-** Distribution of genes in active domains, regulated domains, borders between domains or in chromosome X, for each gene expression class. **B-** Same for regulatory elements. Domains are obtained from Evans *et al.*, 2016.

5.4 Spatial organization of genes in the nucleus

I have shown that ubiquitous, germline and somatic genes are differently structured in terms of local regulatory architecture. At a larger scale, chromatin is spatially organized in 3D (see 1.2.2.2 and Appendix Chapter A). Here, I investigate how genes and regulatory elements are spatially organized in nuclei from individual tissues.

5.4.1 Genes and regulatory elements are segregated in clusters along the chromosomes

In *C. elegans* and across metazoans, chromatin is typically segmented in active or regulated domains (Carelli *et al.*, 2017). In *C. elegans*, “active” domains, delimited by H3K36me3 histone modifications, are enriched for ubiquitously expressed and germline-specific genes, whereas “regulated” domains, delimited by H3K27me3 histone modifications, are enriched for spatially or temporally regulated genes (Evans *et al.*, 2016; Gaydos *et al.*, 2012). Using my gene expression classes, I confirmed that indeed ubiquitous genes are largely segregated in active domains while nearly 90% of the soma-restricted genes are located in regulated domains, domain borders or chromosome X (Figure 5.19A) . However, only 50% of the germline-specific genes were located in the active domains, with the rest generally located in regulated domains.

For the first time, I could also assess the segregation of regulatory elements into active or regulated domains. As expected from the distribution of tissue-specific and ubiquitous genes, the distribution of the ubiquitous or tissue-specific regulatory elements recapitulates that of genes. Ubiquitous regulatory elements are almost entirely found within active domains, germline-specific regulatory elements are enriched in active domains but some are also found over regulated domains, and soma-restricted regulatory elements are largely restricted to regulated domains or chromosome X (Figure 5.19B).

5.4.2 Inferring 3D interactions in individual tissues of young adult worms

The previous observation suggests that based on their tissue-specificity, genes and regulatory elements are linearly segregated along the chromosomes into active and regulated domains. It is now well-established that chromatin spatially folds to acquire a 3D architecture within each nucleus (Beagan and Phillips-Cremins, 2020; Serizay and Ahringer, 2018). I wondered whether ubiquitous and tissue-specific regulatory elements also spatially segregate.

To answer this question, I first needed to identify chromatin interactions occurring in individual tissues at the young adult stage. Our lab developed a method to identify interactions between regulatory elements at a very fine scale ($<$ 500 bp) in *C. elegans* called ARC-C (Accessible Region Chromosome Conformation Capture, Huang *et al.*, 2018). Chromatin is cross-linked and *in situ* digested by DNase at low concentration, preferentially cutting at accessible chromatin loci. DNA ends in spatial proximity are then ligated to each other and finally, the Tn5 transposase is used to capture accessible chromatin and convert it into a sequencing library. After sequencing, chimeric DNA fragments are identified and used to construct chromosome-wide contact maps and to call significant interactions. ARC-C has been performed to profile spatial interactions in nuclei from whole *C. elegans* larvae at the L3 stage. It captured more than twelve million chimeric DNA fragments and identified $\sim$ 33,000 chromatin interactions linking two chromatin accessible sites. Analysis of this set of interactions from whole animals at the L3 stage has shed light on fundamental aspects of the chromatin organization (Huang *et al.*, 2018).

I sought to leverage this set of ~33,000 chromatin interactions (hereafter referred to as *whole larvae L3* interactions) to infer the chromatin interactions in nuclei from each tissue in young adults.

- First, I removed 8,547 interactions identified in larvae where at least one anchor regulatory element was inactive in young adults. This identified 24,741 interactions that are anchored to two regulatory elements both active in YA (this set of interactions is referred to as *pseudo-YA* interactions hereafter) (Figure 5.20A, 1$^{\mathrm{st}}$ step).

- Then, I defined sets of inferred interactions in germline, neuron, hypodermis or intestine nuclei. To do this, for each tissue, I took the subset of interactions from the *pseudo-YA* set of interaction which were anchored at both ends to a regulatory element active in the tissue of interest (Figure 5.20A, 2$^{\mathrm{nd}}$ step). For instance, an interaction anchored to a germline-specific promoter on both ends belongs to the set of germline interactions, whereas an interaction anchored to a ubiquitous promoter on one end and to a promoter active in both neurons and muscles on the other end belongs to both the neuron and muscle interaction sets (see interaction ‡ in Figure 5.20A).

In total, 24,152 interactions were *coherent* and assigned to at least one of the five different sets of inferred interactions. The remaining 589 interactions (2.4% of all the *pseudo-YA* interactions) represent *incoherent* interactions, *i.e.* interactions that are anchored to two regulatory elements which are not active in the same tissues (for instance one germline and one neuron regulatory element).

This approach has two major limitations. First, the original set of interactions is obtained from whole larvae ARC-C. This implies that interactions occurring in a limited number of cells (*e.g.* interactions restricted to a subset of neurons) may be missed. Second, it presumes that the accessibility of regulatory elements in each tissue (*i.e.* their tissue annotations) does not change at all between the L3 stage and the young adult stage. To check to which extent a change in tissue-specific accessibility would affect the resulting sets of interactions, I randomly shuffled the tissue annotations for different proportions of regulatory elements in adults. 2.4% new *incoherent* interactions correspond to a change of tissue annotation for ~6% of the regulatory elements. This suggests that only a small number of regulatory elements change accessibility between the L3 stage and the young adult stage, and that the five resulting different sets of inferred interactions are generally accurate.

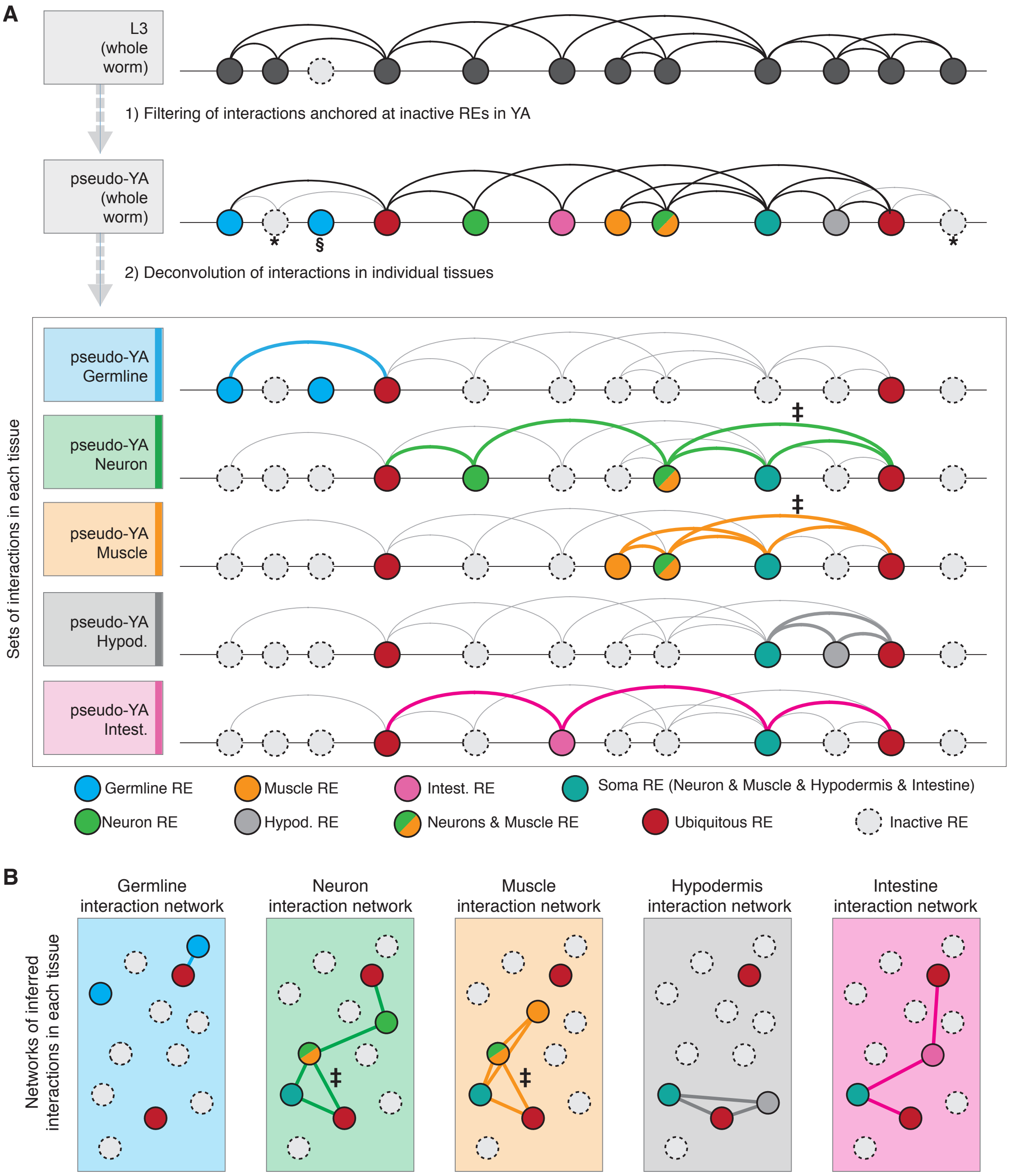

A
L3 (whole worm)
1) Filtering of interactions anchored at inactive REs in YA
pseudo-YA (whole worm)
*
§
*
2) Deconvolution of interactions in individual tissues
Sets of interactions in each tissue
pseudo-YA Germline
pseudo-YA Neuron
‡
pseudo-YA Muscle
‡
pseudo-YA Hypod.
pseudo-YA Intest.
Germline RE
Muscle RE
Intest. RE
Soma RE (Neuron & Muscle & Hypodermis & Intestine)
Neuron RE
Hypod. RE
Neurons & Muscle RE
Ubiquitous RE
Inactive RE
B
Networks of inferred interactions in each tissue
Germline interaction network
Neuron interaction network
‡
Muscle interaction network
‡
Hypodermis interaction network
Intestine interaction network

Figure 5.20 – Inferring chromatin interactions in each tissue at the YA stage. **A-** 33,288 interactions have been mapped in whole worms at the L3 stage. 8,547 interactions anchored to a chromatin locus inactive in YA are discarded (* in the figure). The remaining 24,741 link accessible sites all active in young adult. 24,152 of them are linking two regulatory elements both accessible in at least one tissue. Some specific cases exist: for instance, an interaction between a locus accessible in both neurons and muscle and a ubiquitous locus (accessible in all tissues) ends up in both sets of neuron interactions and muscle interactions (‡ in the figure). Of note, some elements are accessible in YA but not in L3 and thus do not have any interaction with other elements (§ in the figure). **B-** Networks of inferred interactions in each tissue.

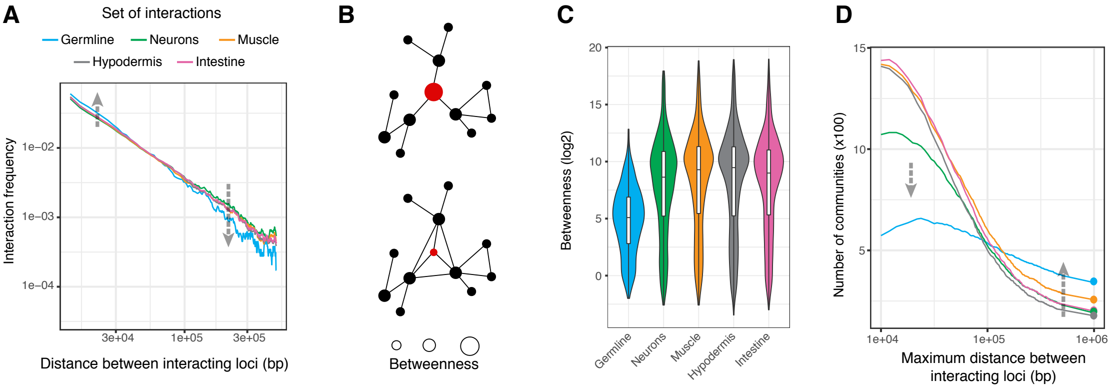


FIGURE 5.21 – Features of inferred interaction networks in individual tissues. **A-** Distance-dependent interaction frequency plot for interaction networks inferred in individual tissues. The x-axis shows the distance between two accessible sites (in base pairs) and the y-axis represents at which frequency two accessible sites separated by this distance interact. **B-** Representation of the notion of betweenness. In the upper network, the red node is central to the network: many of the shortest paths between any two points go through it. Thus, it has a high betweenness. In the lower network, none of the shortest paths between any two points go through the red nod, which has a lower betweenness. **C-** Betweenness of accessible sites in each inferred interaction network. **D-** Distance-dependent number of communities for interaction networks inferred in individual tissues. For each network, the number of communities found in absence of interactions whose distance exceed X base pairs is computed, with X varying between 10 kb and 1 Mb. The total number of communities in each network (found by using a distance cutoff of 1 Mb) are displayed by dots. In **A** and **D**, the same color code is used and the grey dashed arrows indicate the shift between curves from somatic networks or from germline network.

In the future, chromatin spatial organization could be interrogated directly from nuclei of individual tissues to mitigate these limitations.

5.4.3 Differences in chromatin interaction networks in germline or somatic tissues

To better understand the spatial organization of chromatin in nuclei from each tissue, I generated networks of inferred interactions in each tissue (Figure 5.20B). I further investigated the features of the five tissue networks of inferred interactions using approaches derived from the network theory (Ma'ayan, 2011).

I first asked whether the genomic distances separating pairs of interacting loci differed between tissues. I found that shorter interactions (< 100 kb) were enriched in germline nuclei compared to somatic nuclei whereas longer interactions (> 100 kb) were depleted (Figure 5.21A, p-value = 6.761e-18). This suggests that within germline nuclei, interacting elements are generally close to each other along the

Table 5.3 – Communities in the interaction networks inferred in individual tissues

	Number of communities	Average number of sites / community	Number of orphans (% of sites accessible in the tissue)
Germline interaction network	341	7	6,002 / 8,541 (70%)
Neuron interaction network	191	28	5,892 / 11,244 (52%)
Muscle interaction network	256	27	5,597 / 12,631 (44%)
Hypodermis interaction network	177	40	5,304 / 12,347 (43%)
Intestine interaction network	200	35	4,488 / 11.510 (39%)

chromosomes, whereas there are longer-range chromatin interactions in nuclei from somatic tissues.

To further explore 3D organization, I then defined communities in each interaction network. Within a network, a community is a group of accessible sites which are internally densely connected, *i.e.* a group of accessible sites interacting more with each other than with the rest of the network (Porter *et al.*, 2009). I found that there were more communities in the germline interaction network than in the somatic ones (Table 5.3). Importantly, communities in germline interaction network contain fewer accessible sites on average, compared to those in each of the four somatic interaction networks (Table 5.3). This suggests that within germline nuclei, interacting elements are organized in many small communities.

I also measured the betweenness for each accessible site, in each network. For a given node v, its betweenness score $g(v)$ is given by the expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{st}$is the number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those paths that pass through v (Girvan and Newman, 2002). The betweenness score gives an indication on how essential a given node is for the overall structure of the surrounding network. Essential central sites have a high betweenness score while non-essential sites have a low betweenness score (Figure 5.21B). I found that the betweenness of accessible sites in the somatic inferred interaction networks is generally higher than that of sites in germline inferred interaction networks (Figure 5.21C). This suggests that the organization of the somatic networks relies on key central loci whereas those in the germline network are more interconnected and rely less on local sites.

Finally, I focused on the importance of long-distance interactions in generating communities within the networks inferred in individual tissues. For each tissue interaction network, I simulated the organization of accessible sites into communities in the absence of interactions longer than a certain distance (Figure 5.21D). In the germline interaction network, the number of communities is relatively independent of the presence or absence of long-distance interactions. On the contrary, in somatic interaction networks, the number of individual communities drastically decreases with long-distance interactions.

Taken together, these results suggest that the germline interaction network primarily consists of small uniform communities of regulatory elements (Figure 5.21 and Table 5.3). In contrast, somatic interaction networks seem to rely on long-range interactions to merge individual small communities into larger structured communities (Figure 5.21 and Table 5.3).

5.4.4 Soma promoters could act as physical hubs recruiting somatic transcription factors

5.4.4.1 Soma accessible sites are important interaction hubs in somatic interaction networks

I sought to investigate the factors responsible for the differences observed between germline and somatic regulatory networks. A major difference between accessible chromatin sites in germline or somatic tissues is the presence of ~ 1,000 “soma” sites, accessible in all the somatic tissues but not in germline nuclei. To better understand the different contribution of tissue-specific sites, soma sites and ubiquitous sites to the inferred networks, I calculated the accessible site degrees, for ubiquitous, tissue-specific and somatic sites. The degree of a site is the number of interactions between this site and other sites. This revealed that in each inferred network, 40 to 60% of the ubiquitous sites interact with one to ten other sites (Figure 5.22A). This was also observed for the somatic tissue-specific accessible sites. However, less than 10% of the germline sites are connected to other sites. Strikingly, I observed that in the somatic networks, soma sites are characterized by very high degrees, *i.e.* soma sites have an unusually high number of interactions with other sites (Figure 5.22A). Moreover, the number of interactions for each soma site is positively correlated with their accessibility (Figure 5.22B). These observations suggest that the soma

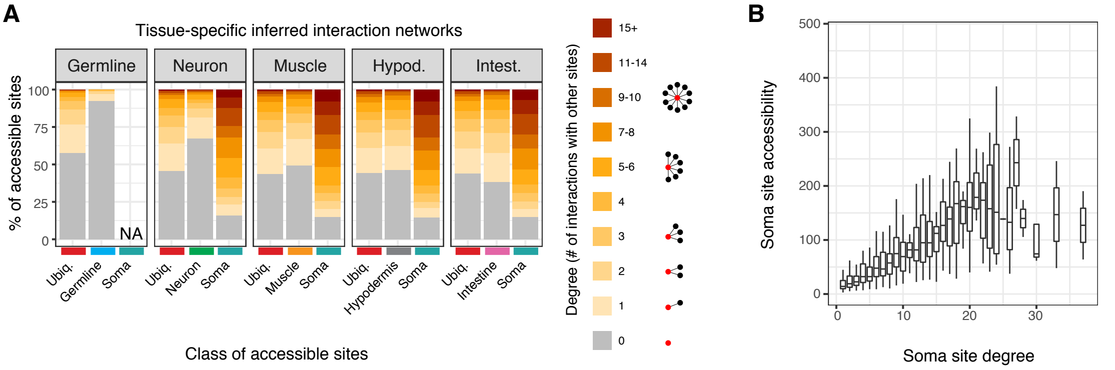


FIGURE 5.22 – Soma sites are interacting with a large number of loci. **A-** Accessible site degree (*i.e.* the number of interactions between a site and other sites), for ubiquitous, tissue-specific or soma sites. The interaction networks inferred in each of the five tissues were individually analyzed. **B-** Correlation between chromatin accessibility and degree of soma sites.

sites, specifically accessible across the four somatic tissues, could play a role of interaction hubs in each somatic tissue inferred network.

To get more insights in the role of soma sites in organizing networks of interactions, I sought to study the organization of the four somatic interaction networks in the absence of these soma sites. To do so, I first computationally masked the ~ 1,000 soma sites from the original networks and simulated the resulting interaction networks in each of the four somatic tissues. I analyzed the structure of the resulting simulated networks using the different metrics previously described (see Figure 5.21). When suppressing the soma sites from the somatic interaction networks, the distance-dependent interaction frequencies in the simulated networks become similar to that of the germline interaction network (Figure 5.23A and insets * and **). Moreover, the betweenness of accessible sites in the simulated somatic interaction networks without the soma sites is reduced and more comparable to that of sites in the germline interaction network (Figure 5.23B). Finally, albeit still fluctuating more than in germline network, the number of communities in the simulated somatic interaction networks without the soma sites is generally less dependent on long-range interactions (Figure 5.23C).

Together, these observations suggest that the soma sites, accessible in nuclei of the four somatic tissues, may play a fundamental role in bringing individual small communities of accessible sites together, in order to form large communities. This data support a model whereby soma sites act as anchoring hubs for long-range interactions, important for the aggregation of accessible sites into large communities.

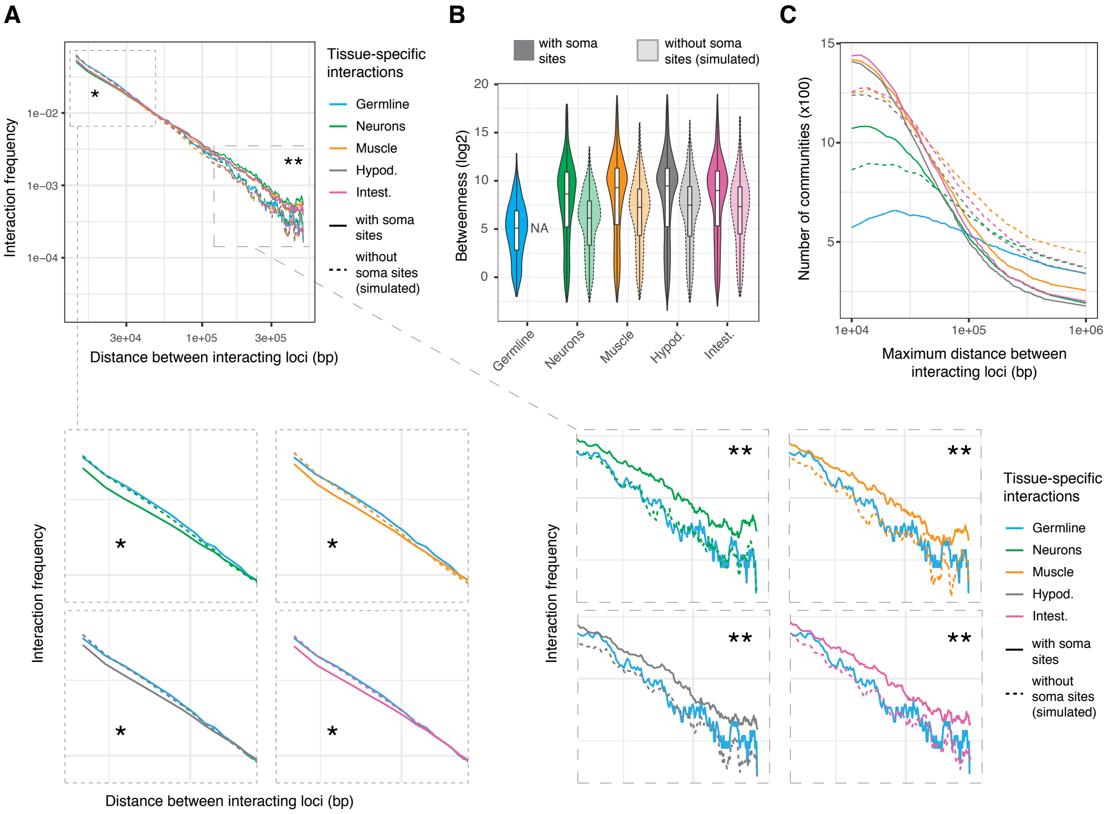


Figure 5.23 – Contribution of the soma sites to the structure of interaction networks. **A-** Distance-dependent interaction frequency plot for interaction networks inferred in individual tissues (as described in Figure 5.21A), with or without soma sites. The dashed lines show the interaction frequencies in the different interaction networks simulated in the absence of soma sites. Close-up insets (* and **) are displayed below the main graph. **B-** Betweenness of accessible sites in each inferred interaction network. The transparent violin plots show the distribution of the betweenness in the somatic interaction networks simulated in the absence of soma sites. **C-** Distance-dependent number of communities in interaction networks inferred in individual tissues (as described in Figure 5.21C). The dashed lines show the number of communities in the different interaction networks simulated in the absence of soma sites. The same code is used in **A** and **C**.

5.4.4.2 Soma sites are bound by many transcription factors

I then sought to investigate which transcription factors were binding to soma sites as well as to other sites. I leveraged the extensive database of ChIP-seq experiments provided by modENCODE and modERN (Kudron *et al.*, 2018; ModENCODE, 2011). I obtained the annotated TF binding sites from the modENCODE/modERN portal and computed the TF binding enrichment score for each TF in each class of tissue-specific, soma (*i.e.* active across the four somatic tissues) and ubiquitous accessible sites (Figure 5.24A). As expected, known tissue-specific transcription factors were found to specifically bind to tissue-specific classes of sites (*e.g.* ELT-2 and DAF-16 enriched in intestine sites, or UNC-55 and UNC-86 enriched in neuron sites, Figure 5.24A). I then clustered the transcription factors according to their binding patterns (Figure 5.24B). This identified several clusters of transcription factors known to functionally interact to co-regulate sets of regulatory elements. For example, DPL-1 and LIN-53 are component of the DREAM complex responsible for silencing reproduction-related genes in somatic tissues in the worm (Petrella *et al.*, 2011; Wu *et al.*, 2012 and Figure 5.24C). SNPC proteins physically interact with UNC-130 and RPC-1 and are responsible for piRNA expression in germline (Kasper *et al.*, 2014 and Figure 5.24C). Finally, DAF-16, ELT-2, PQM-1 and NHR-80 are central factors co-regulating many intestinal genes (McGhee *et al.*, 2009; Tepper *et al.*, 2013 and Figure 5.24C).

Intriguingly, despite their well-known tissue-specific functions, many transcription factors are also found bound to soma sites, active across the four somatic tissues (*e.g.* ELT-2 and DAF-16 in intestine or UNC-55 and UNC-86 in neurons) (Figure 5.24A). This prompted me to investigate the number of transcription factors bound to the different sets of accessible sites. I found that many transcription factors bound to individual soma sites (median of 9, Figure 5.25). Yet, soma sites do not harbor binding sites responsible for the recruitment of many of these transcription factors (data not shown). However, soma sites are characterized by an exceptionally high accessibility, compared to ubiquitous or tissue-specific loci (see Figure 5.11 on page 114).

Together, these results show that many tissue-specific transcription factors can promiscuously bind to soma accessible sites. Previous work identified "high-occupancy target" (HOT) sites in *C. elegans* and other organisms (Chen *et al.*,

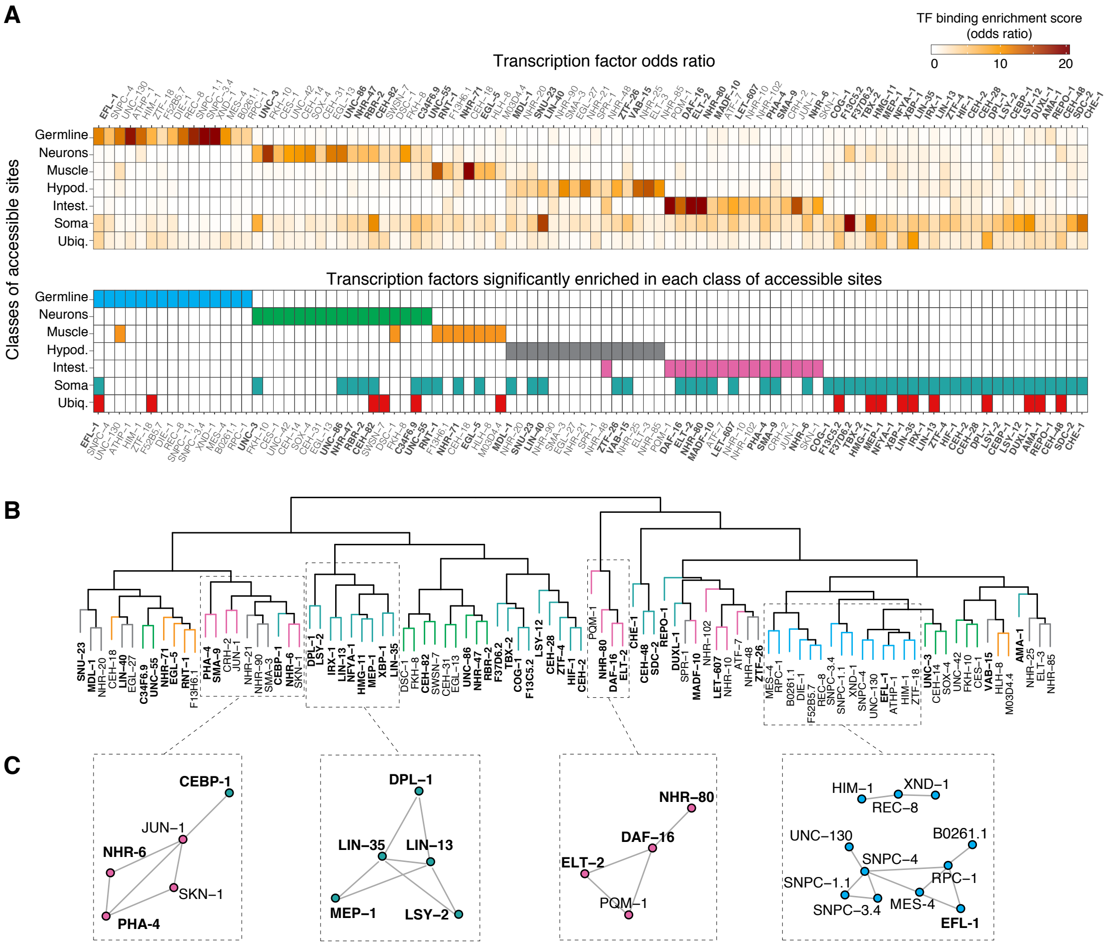


FIGURE 5.24 – Transcription factors enriched in each class of accessible sites. **A-** Heatmap of transcription factor enrichment score (odds ratio) for each transcription factor over tissue-specific, soma (*i.e.* active across the four somatic tissues) and ubiquitous accessible sites. The bottom heatmap is a summary of the transcription factors significantly enriched over each class of accessible sites (enrichment score >= 3, p-value <= 0.05, % of bound accessible sites >= 1%). **B-** Hierarchical clustering of the transcription factor binding patterns. The colors of the branches indicate to which set of accessible sites the corresponding transcription factor preferentially binds. **C-** Examples of functional interaction networks obtained from STRING database based on the hierarchical clustering in **B**. The colors of the nodes indicate to which set of accessible sites the corresponding transcription factor preferentially binds. The transcription factors enriched at soma accessible sites are bolded.

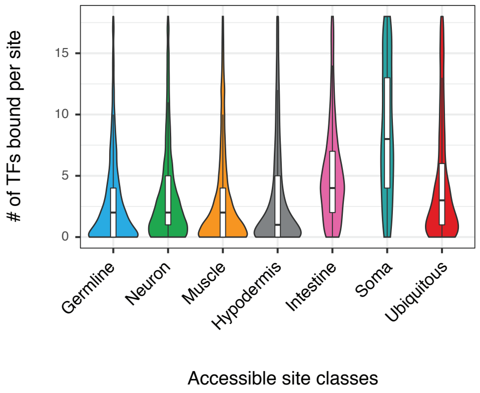


Figure 5.25 – Number of transcription factors bound to individual sites, for each class of accessible site.

2014; Foley and Sidow, 2013; Gerstein *et al.*, 2010; Roy *et al.*, 2010). HOT sites are defined as sites bound by an unusual number of transcription factors but were the TF binding motifs are absent. I investigated whether the soma sites I characterized correspond to previously identified HOT sites in *C. elegans* (Chen *et al.*, 2014). I found that 120 out of the 956 soma sites overlap with HOT sites (120 of the 359 annotated HOT sites, 28-fold enrichment, p-value = 1.8e-111).

5.4.4.3 An emerging role for soma accessible sites

Together, these results suggest that the soma accessible sites play a fundamental role in the spatial architecture of the chromatin and for the recruitment of transcription factors, in somatic tissues. Around 1,000 sites found across the genome are accessible in all the somatic tissues but not in the germline, and they are characterized by a strikingly high accessibility. They appear to act as physical hubs which contribute to spatially aggregate regulatory elements into large communities. They also seem to be able to promiscuously recruit transcription factors regulating tissue-specific and ubiquitous processes, essentially behaving like HOT sites. This suggests a model whereby these soma sites would act as central platforms to locally enrich the concentration of transcription factors and facilitate their transfer to distant accessible sites brought in close proximity.

5.5 Discussion

Determining regulatory architectures responsible for the different gene expression patterns found in multicellular organisms is necessary for understanding how the genome encodes development. For the first time, I comprehensively profiled gene expression and chromatin accessibility in all the main tissues of the adult *C. elegans*. These new datasets allowed to annotate 13% more regulatory elements than the most detailed set of annotations as of today (Jänes *et al.*, 2018). Regulatory element dissection approaches have been historically used to understand how spatiotemporal patterns of gene expression were obtained (Cox and Hirsh, 1985; Krause *et al.*, 1994; Okkema *et al.*, 1993; Rougvie and Ambros, 1995). However, these approaches are laborious and low throughput. The set of annotations provided here is going to be instrumental in helping design experimental procedures. Generally, it is a great resource for the worm community to study gene regulation in *C. elegans*, complementing and improving previously published annotations (Chen *et al.*, 2013; Jänes *et al.*, 2018; Kruesi *et al.*, 2013; Saito *et al.*, 2013).

I classified genes and regulatory elements into ubiquitous, tissue-restricted and tissue-specific sets. This confirmed that regulatory elements are generally highly tissue-specific and suggests that they are even more abundant than previously thought (Cusanovich *et al.*, 2018b; Liu *et al.*, 2019a). Using this classification, I could also directly investigate the relationship between the structure and the biological function of ubiquitous and tissue-specific genes. This revealed that ubiquitous genes involved in basic biological processes and germline-specific genes both have a simple and compact structure, whereas ubiquitous genes involved in developmental processes and somatic-specific genes tend to have a more complex structure (Figure 5.26A).

Importantly, chromatin seems to be differently organized in germline and somatic tissues. In germline nuclei, chromatin may be spatially segregated into small individual communities of regulatory elements. In contrast, regulatory elements located further apart from each other may interact and form larger communities in nuclei from somatic tissues. These dramatic differences in chromatin organization may be linked to the plasticity of the cells in each tissue. Indeed, most somatic cells reach terminal differentiation in embryogenesis, as soon as 500 minutes post fertilization, and then ensure the same biological function throughout *C. elegans* life cycle. On the contrary, the two germ cell progenitors formed early during embryogenesis rapidly go into quiescence, then start proliferating in larval stages, to eventually differentiate into either sperm or oocyte gametes (Kimble and Crittenden, 2005). On top of this, fine regulation of the chromatin is achieved in germline

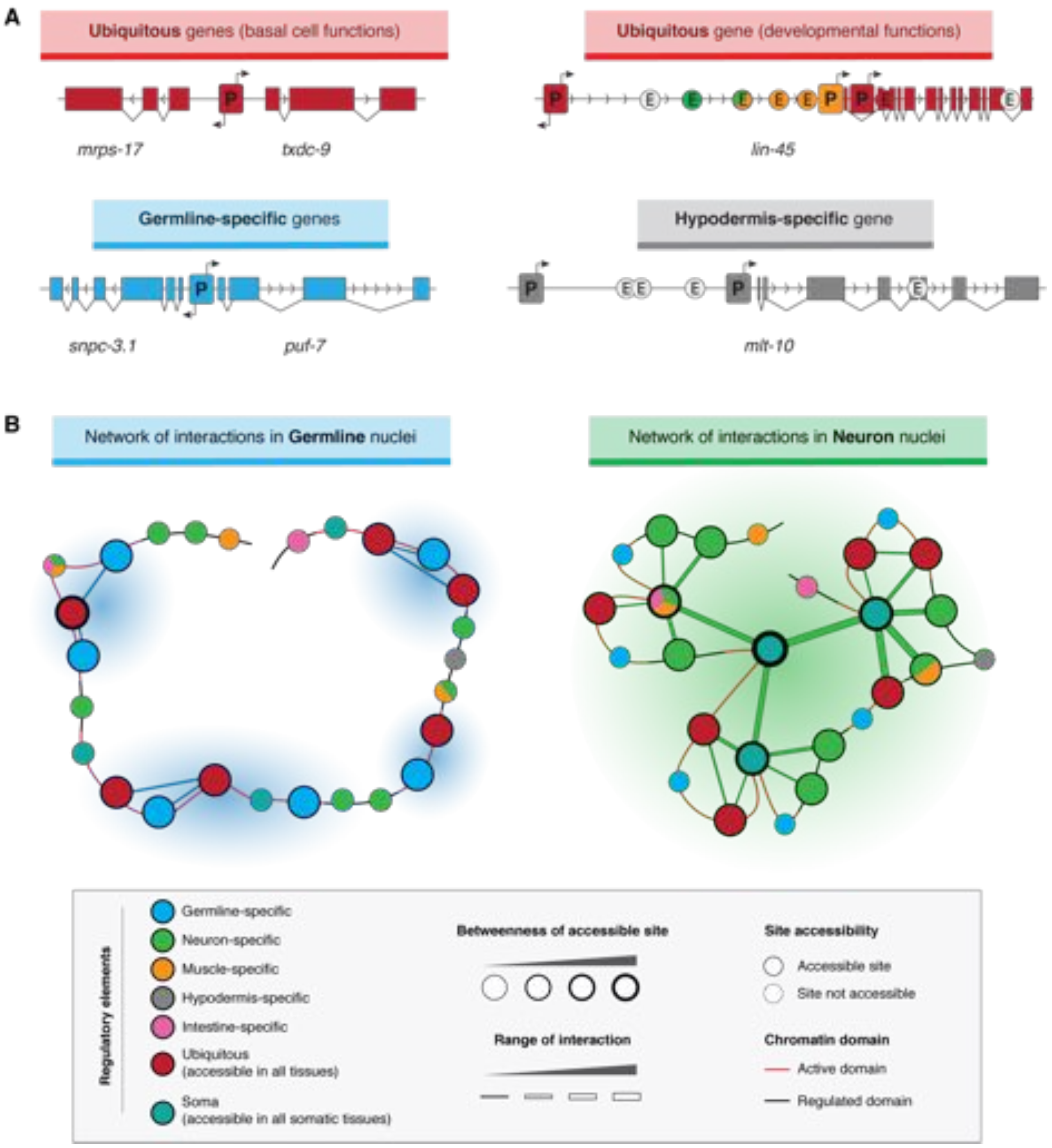


Figure 5.26 – Models of gene regulatory architectures at different scales. **A-** At the scale of individual genes, two distinctive structures are found. Left: examples of the simple regulatory architecture shared by ubiquitous genes (*e.g. mrps-17* and *txdc-9*) and germline-specific genes (*e.g. snpc-3.1* and *puf-7*). Right: examples of more complex architectures found at developmental ubiquitous genes (*e.g. lin*-45) or somatic tissue-specific genes (*e.g. mlt-10*). **B-** At a larger scale, networks of chromatin spatial interactions differ between germline and somatic tissues. The same succession of regulatory elements can either fold into small communities in germline nuclei (left) or into a large organized community in neuron nuclei (right). Central soma accessible sites play an essential role in organizing regulatory elements into a large hierarchical community.

to limit the expression of transgenes (Bagijn *et al.*, 2012) and ensure integrity of germline DNA (Schaner and Kelly, 2006). At each step throughout germline development, important remodeling of histone modifications has been observed *in vivo* (Schaner and Kelly, 2006). Thus, investigating the relationship between these remodeling events and the spatial organization of the chromatin in germline represents an exciting challenge for future studies.

Strikingly, soma sites, defined here as the sites accessible in all four somatic tissues, seem to act as interaction hubs at the center of these large communities (Figure 5.26B). They show high similarity and significantly overlap with previously characterized HOT sites (Chen *et al.*, 2014; Gerstein *et al.*, 2010; Roy *et al.*, 2010). Several different hypotheses could be formulated to explain how soma sites may function as central nodes for large communities. These sites might recruit one or several specific DNA-binding protein(s) not identified yet, which in turn would help generating large multivalent transcription factor complexes, following a model recently proposed to regulated X chromosome folding during the X Chromosome Inactivation (Pandya-Jones *et al.*, 2020). The intrinsic DNA sequence of these soma sites could also have physical properties facilitating the non-specific recruitment of transcription factors (Wreczycka *et al.*, 2019). Eventually, the aggregation of transcription factors could lead to the formation of phase-separated compartments, as suggested in a model proposed for super-enhancers (Hnisz *et al.*, 2017). Models of phase-separated chromatin compartments have already been been proposed to segregate heterochromatin, but such process is usually thought to rely on histone modification and proteins (Feric *et al.*, 2016; Larson *et al.*, 2017; Strom *et al.*, 2017). Here, soma regulatory elements would stand as the hubs favoring nucleation and formation of phase-separated compartments. Overall, these conceptual models open new areas of investigation which could shed light on general principles of chromatin spatial organization in *C. elegans* somatic tissues.

Still, this model of chromatin interactions is based on interaction networks inferred in each tissue from whole organism data and assuming that these interactions do not vary between L3 and YA (Huang *et al.*, 2018). In the future, sets of interactions directly obtained from nuclei of individual tissues would be necessary to confirm these conclusions and go further in the analysis of tissue-specific chromatin interactions.

Chapter 6

Molecular organization of promoters in adult *C. elegans*

In Chapter 5, I charactered the regulatory architectures of ubiquitous and tissue-specific genes and the chromatin interaction networks in nuclei from individual tissues. The tissue-specific differences in these gene regulatory architectures prompted me to investigate whether differences were also present at the level of promoters. In Chapter 6, I study the nucleosome organization at promoters of different classes and investigate the mechanisms underlying their positioning.

6.1 Nucleosomes flank ubiquitous and germline-specific promoters

6.1.1 Patterns of chromatin accessibility at ubiquitous and tissue-specific promoters

When I originally compared the patterns of chromatin accessibility at different classes of promoters, I observed a striking feature only found in germline-specific and ubiquitous promoters (see Figure 5.11 on page 114). At these promoters, the central peak of accessibility is immediately flanked by close neighboring regions of increased accessibility. In contrast, accessibility was greater at somatic promoters at the center of their peak, but no “shoulder” could be detected. I wondered whether differences in ATAC-seq fragment sizes could explain these different signatures. I plotted the distribution of ATAC-seq fragment sizes mapping over promoters of each

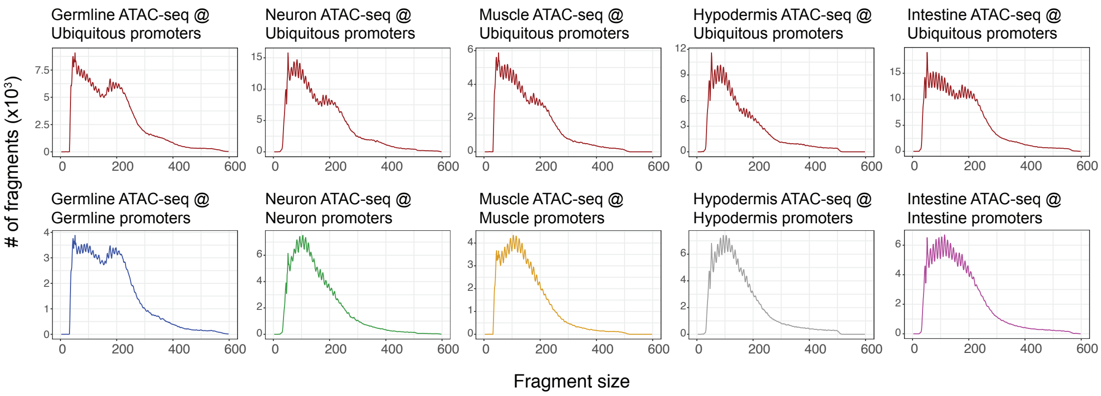


FIGURE 6.1 – Distribution of ATAC-seq fragment sizes from tissue-specific ATAC-seq datasets over different sets of promoters.

class and found that the size of ATAC-seq fragments mapping over ubiquitous or germline-specific promoters followed a multi-modal distribution (Figure 6.1). Both short ATAC-seq fragments (< 100 bp) and longer fragments (~200 bp and ~350 bp) mapped over these loci. Importantly, this was true for ubiquitous promoters across all tissue-specific ATAC-seq datasets, suggesting that this was not a technical artifact associated with the germline ATAC-seq dataset. This was in sharp contrast with the unimodal distribution of ATAC-seq fragment sizes for fragments mapping over somatic-tissue-specific promoters (Figure 6.1). This suggests that the molecular organization of somatic promoters or ubiquitous and germline-specific promoters could be fundamentally different.

A multi-modal fragment size distribution is generally expected from whole ATAC-seq libraries (Buenrostro *et al.*, 2013). Longer ATAC-seq sequenced fragments (> 150 bp) typically represent nucleosome-spanning fragments (Buenrostro *et al.*, 2013; Schep *et al.*, 2015). Thus, the location of these long ATAC-seq fragments relative to the center of regulatory elements is indicative of where and how nucleosomes are positioned. Aggregated ATAC-seq fragment density plots (also known as “V-plots”, Henikoff *et al.*, 2011) can be used to visualize this distribution of fragment lengths relative to the center of different sets of promoters (Figure 6.2, more details in 8.1 on page 173). Stereotypical patterns of fragment density are observed over promoters flanked by consistently aligned -1 and +1 nucleosomes: small fragments mostly overlap the promoter centers while larger fragments are over +1/-1 nucleosomes on either side of the promoters (Figure 6.2A). On the contrary, if the flanking nucleosomes are not consistently aligned relative to promoter centers, fragment density plots do not show any increased fragment density other than that

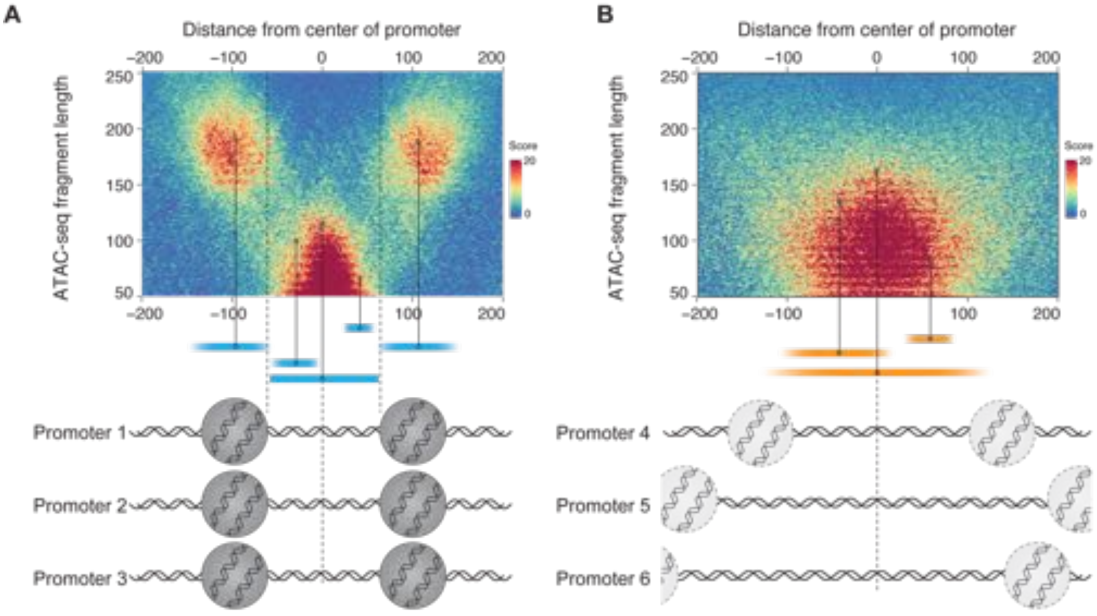


FIGURE 6.2 – Interpretation of two different ATAC-seq fragment density plots. **A-** At promoters flanked by nucleosomes aligned relative to the center of the promoters, short ATAC-seq fragments are enriched close to the center of the promoters while longer fragments are enriched on each side of it. These longer fragments are nucleosome-spanning fragments. **B-** At promoters flanked by weakly positioned nucleosomes which are not aligned relative to the center of the promoters, only the central nucleosome-depleted region is dense. See Figure 8.1 on page 174 for more details.

at the nucleosome depleted regions at promoter centers (Figure 6.2B). For each of the five tissue-specific ATAC-seq datasets, I generated ATAC-seq fragment density plots over ubiquitous and the corresponding tissue-specific promoters. In line with the multimodal distribution observed over ubiquitous promoters (Figure 6.1), a -1 and +1 nucleosome signature is readily apparent over ubiquitous promoters in all tissues (Figure 6.3, top row). The same pattern is also observed over germline-specific promoters in germline ATAC-seq data (Figure 6.3), in agreement with the multimodal distribution of ATAC-seq fragment sizes over germline-specific promoters (Figure 6.1). However, somatic tissue-specific promoters lack this signature of well-positioned +1/-1 nucleosomes (Figure 6.3). To better quantify the differences between each fragment density plots, I devised an approach to estimate a flanking nucleosome enrichment score based on the background distribution of ATAC-seq fragments (Figure 6.4A). The flanking nucleosome enrichment score over ubiquitous promoters is similar across all tissue-specific ATAC-seq datasets, and only the germline-specific promoters have a comparable enrichment score; in contrast, the somatic-tissue-specific promoters have a very low flanking nucleosome enrichment

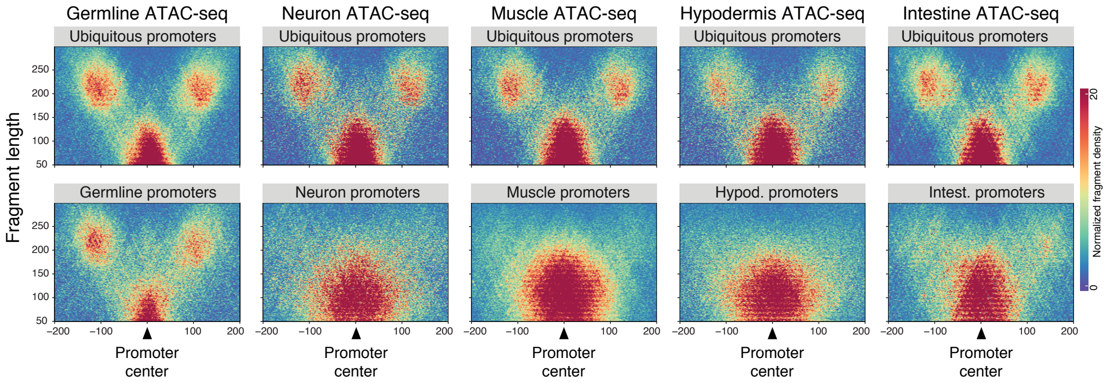


FIGURE 6.3 – ATAC-seq fragment density plots over ubiquitous and tissue-specific promoters.

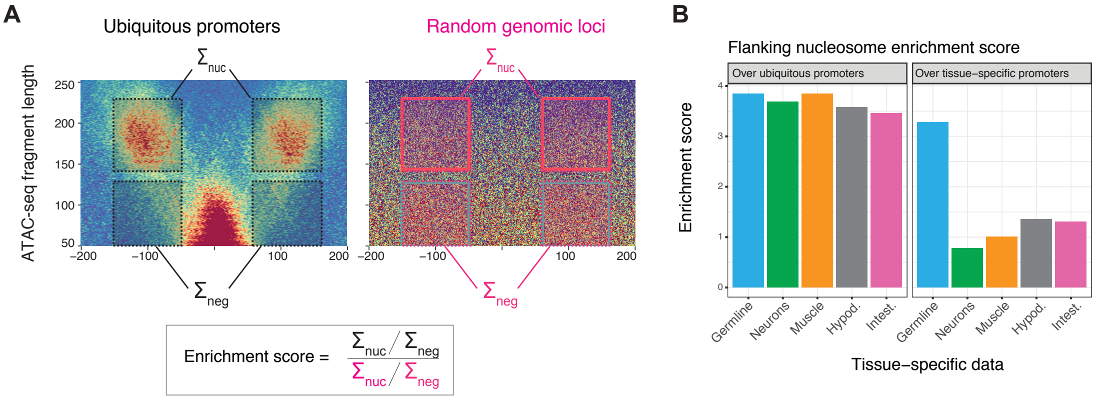


FIGURE 6.4 – Flanking nucleosome enrichment scores at different sets of promoters. **A-** Approach used to compute flanking nucleosomes enrichment score from an ATAC-seq fragment density plot. Left: ATAC-seq fragment density plot computed over a set of promoters; Right: ATAC-seq fragment density plot computed over control intergenic regions. The formula used to compute the resulting flanking nucleosome enrichment score is detailed underneath the two plots. **B-** Flanking nucleosome enrichment scores over ubiquitous and tissue-specific promoters.

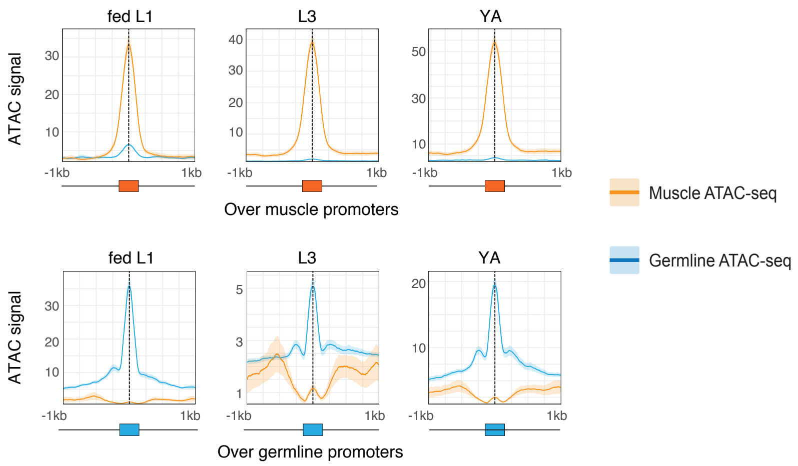


Figure 6.5 – Nucleosomes are flanking germline-specific promoters across *C. elegans* development regardless of gene expression levels. Muscle and germline-specific ATAC-seq aggregate profiles are plotted over muscle and germline-specific promoters, at L1, L3 and young adult (YA) stages. Germline-specific ATAC-seq data at L1 stage comes from Lee *et al.* (2017).

(Figure 6.4B).

Chromatin in germline nuclei adopt a specific conformation during meiosis at the adult developmental stage. To test whether the meiotic chromatin conformation was responsible for the nucleosome positioning I observed, I compared chromatin accessibility in germline and muscle nuclei in adult with two other developmental stages. I produced ATAC-seq datasets from muscle nuclei in L1 and L3 stages and from germline nuclei in L3 stage, and I obtained germline ATAC-seq from L1 stage from Lee *et al.*, 2017. At L1 stage, germline cells are quiescent and at L3 stage, the germline is actively proliferating. I observed consistent ATAC-seq patterns across L1, L3 and adult stages at muscle or germline promoters. Muscle ATAC-seq profiles over muscle promoters consistently showed a single narrow peak of accessibility while germline ATAC-seq profiles showed increased accessibility on the flanking sides of the germline promoters at all stages (Figure 6.5). This suggests that the typical arrangement of flanking nucleosomes is a property of germline promoters throughout post-embryonic development.

6.1.2 Ubiquitous and germline promoters are stereotypically organized

To gain more insight into nucleosome positioning at different sets of promoters, I used nucleoATAC to compute nucleosome occupancy probability profiles from

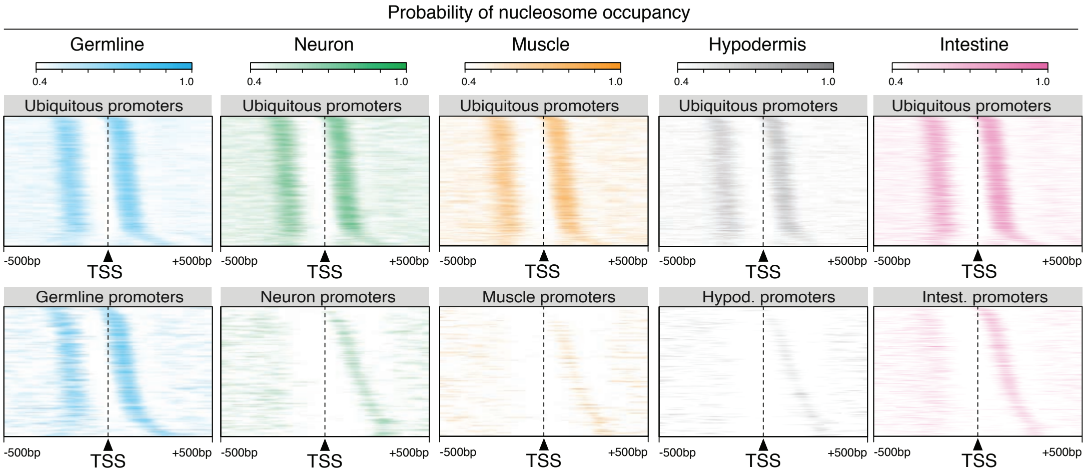


Figure 6.6 – Nucleosome occupancy probability at ubiquitous or tissue-specific promoters. For each tissue ATAC-seq dataset, the nucleosome occupancy probability was computed over ubiquitous and tissue-specific promoters using nucleoATAC as described in Schep *et al.* (2015). Rows are ordered by the distance from the TSS to the +1 nucleosome. TSS annotations were obtained from Jänes *et al.* (2018) and only promoters with experimentally defined forward and reverse TSSs are considered.

ATAC-seq data (Schep *et al.*, 2015). This confirmed the preliminary observations from fragment density plots and revealed that the +1 and -1 nucleosomes are positioned at a relatively consistent distance from the TSS in ubiquitous and germline-specific promoters (Figure 6.6). In contrast, somatic tissue-specific promoters are characterized by lower -1 and +1 nucleosome occupancy and a larger range of nucleosome positions relative to TSSs (Figure 6.6).

Using genome-wide nucleosome occupancy probability profiles in combination with previous mapping of dominant transcription initiation sites (Jänes *et al.*, 2018), I could quantitatively measure parameters of promoter nucleosomal organization (Figure 6.7A). At ubiquitous and germline-specific promoters, I found that the 5' edge of the +1 nucleosome is generally found ~ 20 bp downstream of the TSS (median distances of 22 bp for ubiquitous promoters and 12 bp for germline-specific promoters, Figure 6.7B). In contrast, at somatic tissue-specific promoters, the +1 nucleosomes are more widely distributed downstream of the TSS (Figure 6.7B). Moreover, divergent TSSs are closer to each other in ubiquitous and germline-specific promoters than in somatic tissue-specific promoters and NDRs are narrower in ubiquitous and germline-specific promoters (Figure 6.7B).

Together, these results show that at ubiquitous and germline-specific promoters, well-positioned +1 nucleosomes are reproducibly aligned ~ 20 bp downstream of

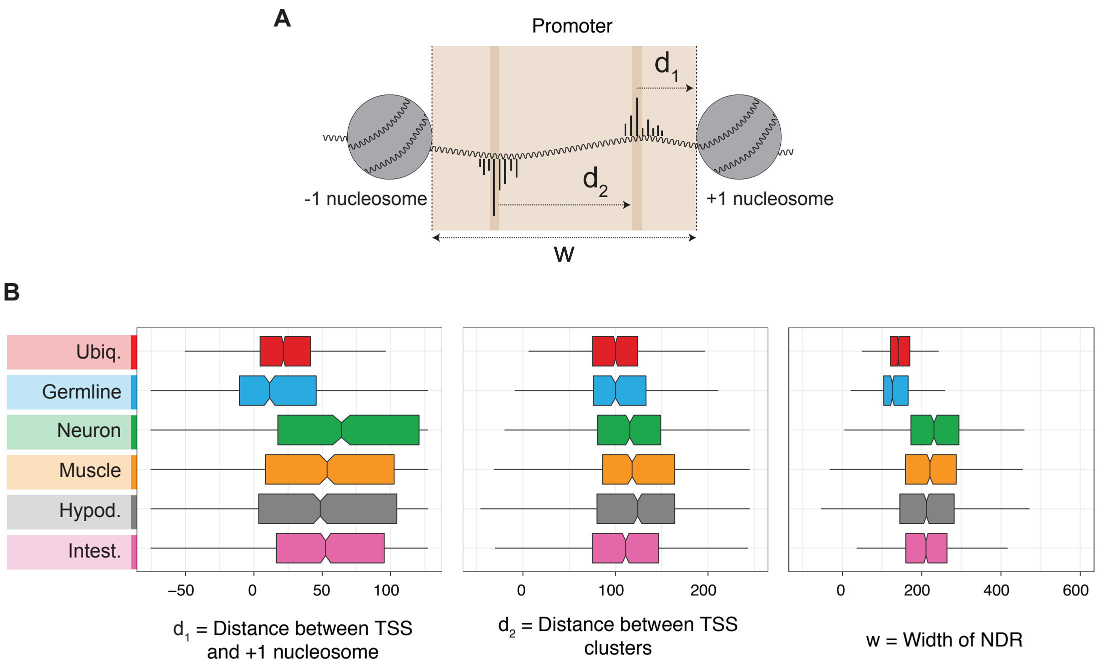


FIGURE 6.7 – Organization of nucleosomes flanking promoters. **A-** Schematic of the distance metrics measured in each set of promoters: d1, distance between the mode TSS and the +1 nucleosome edge; d2, distance between modes of divergent TSSs within the same promoter; w, width of the nucleosome-depleted region (NDR). **B-** d1, d2 and w distance metrics for ubiquitous or tissue-specific promoters. The metrics for ubiquitous promoters were measured using nucleosome occupancy probability track derived from whole young adult ATAC-seq data (Jänes *et al.*, 2018). TSS annotations were obtained from Jänes *et al.* (2018) and only promoters with experimentally defined forward and reverse TSSs are considered.

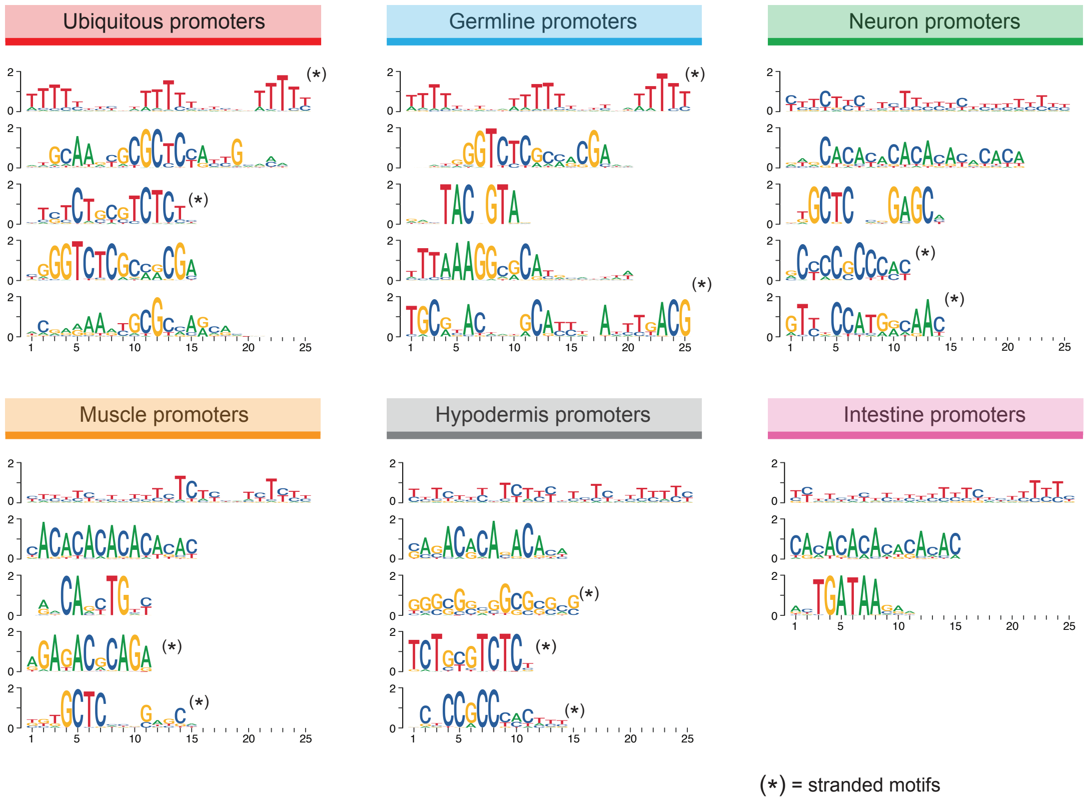


FIGURE 6.8 – Motifs enriched in each set of promoters. Sequences from -75 to +105 bp around the promoter centers were considered. Only the five motifs with the highest enrichment are shown (only three found in intestine promoters).

the TSS, generating a narrow central nucleosome-depleted region. In comparison, somatic tissue-specific promoters have a wider NDR and their flanking nucleosomes are not aligned with their associated TSS.

6.2 Underlying sequence features are contributing to promoter structure

6.2.1 10-bp WW periodicity at germline-active promoters

Some sequences features are thought to influence the positioning of nucleosomes (Struhl and Segal, 2013). To understand whether specific sequence features could be responsible for the differences I observed in ubiquitous and germline promoters compared to somatic-tissue-specific ones, I search for short sequences enriched in each class of promoters. Interestingly, I observed that ubiquitous and germline-specific promoters share a T-rich motif with 10 bp spacing, which was not present at somatic tissue-specific promoters (Figure 6.8). Previous studies have implicated

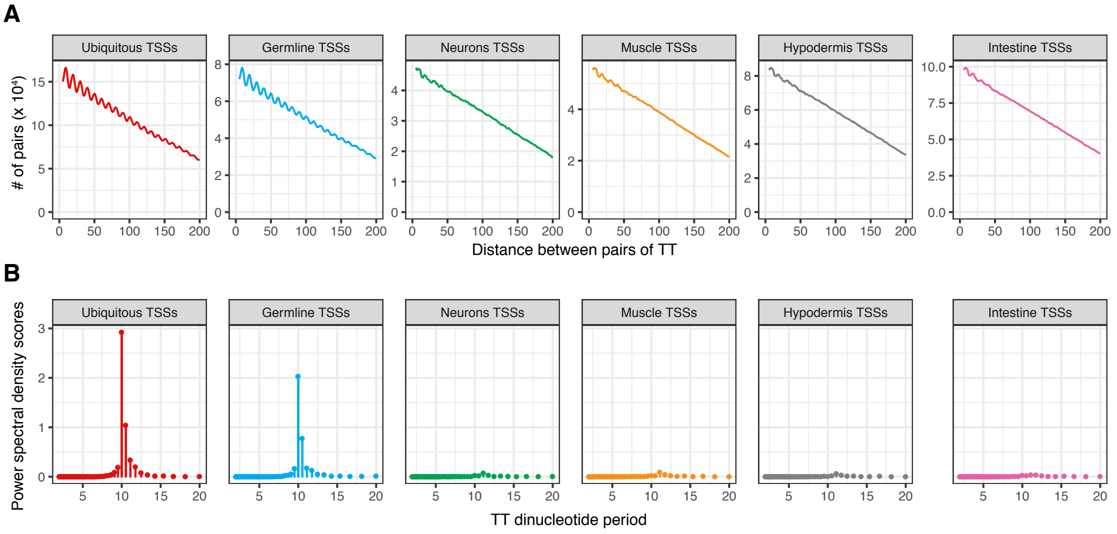


Figure 6.9 – TT periodicity at ubiquitous and tissue-specific TSSs. **A-** Distribution of distances between pairs of TT dinucleotides (TT....TT) found in sequences -50 bp to + 300 bp around TSSs of different types of promoters. **B-** Associated power spectral density values of TT periodicities.

10-bp WW (W = A/T) periodicity in nucleosome positioning (Andersson *et al.*, 2014; Dreos *et al.*, 2016; Field *et al.*, 2008; Haberle *et al.*, 2014; Ioshikhes *et al.*, 1996; Johnson *et al.*, 2006; Mavrich *et al.*, 2008a; Satchwell *et al.*, 1986; Segal *et al.*, 2006; Struhl and Segal, 2013; Wang and Widom, 2005). To investigate whether the T-rich motif found at ubiquitous and germline-specific promoters was part of a larger TT periodic signal, I sought to quantify the TT periodicity in each set of promoters. I computed the distances between all possible pairs of TT dinucleotides in the sequences from -50 bp to +300 bp relative to TSSs of the different classes of promoters. This showed that around ubiquitous and germline-specific TSSs, pairs of TT dinucleotides are generally interspaced by k bases, k being a multiple of 10 (Figure 6.9A). To quantify the overall TT periodicity, I computed the power spectral density for each histogram using a Fourier Transform. This analysis confirmed that TT dinucleotides exhibit strong 10-bp periodicity in ubiquitous and germline-specific promoter sequences but not in somatic-tissue-specific promoters (Figure 6.9B).

To assess whether the 10-bp TT periodicity I measured in the vicinity of ubiquitous and germline TSSs was associated with +1 nucleosomes, I generated genome-wide tracks of 10-bp dinucleotide periodicity (see Chapter 8). I observed that ubiquitous and germline-specific promoter regions harbor a strong 10-bp periodic WW signal immediately downstream of their TSS (Figure 6.10A). At these

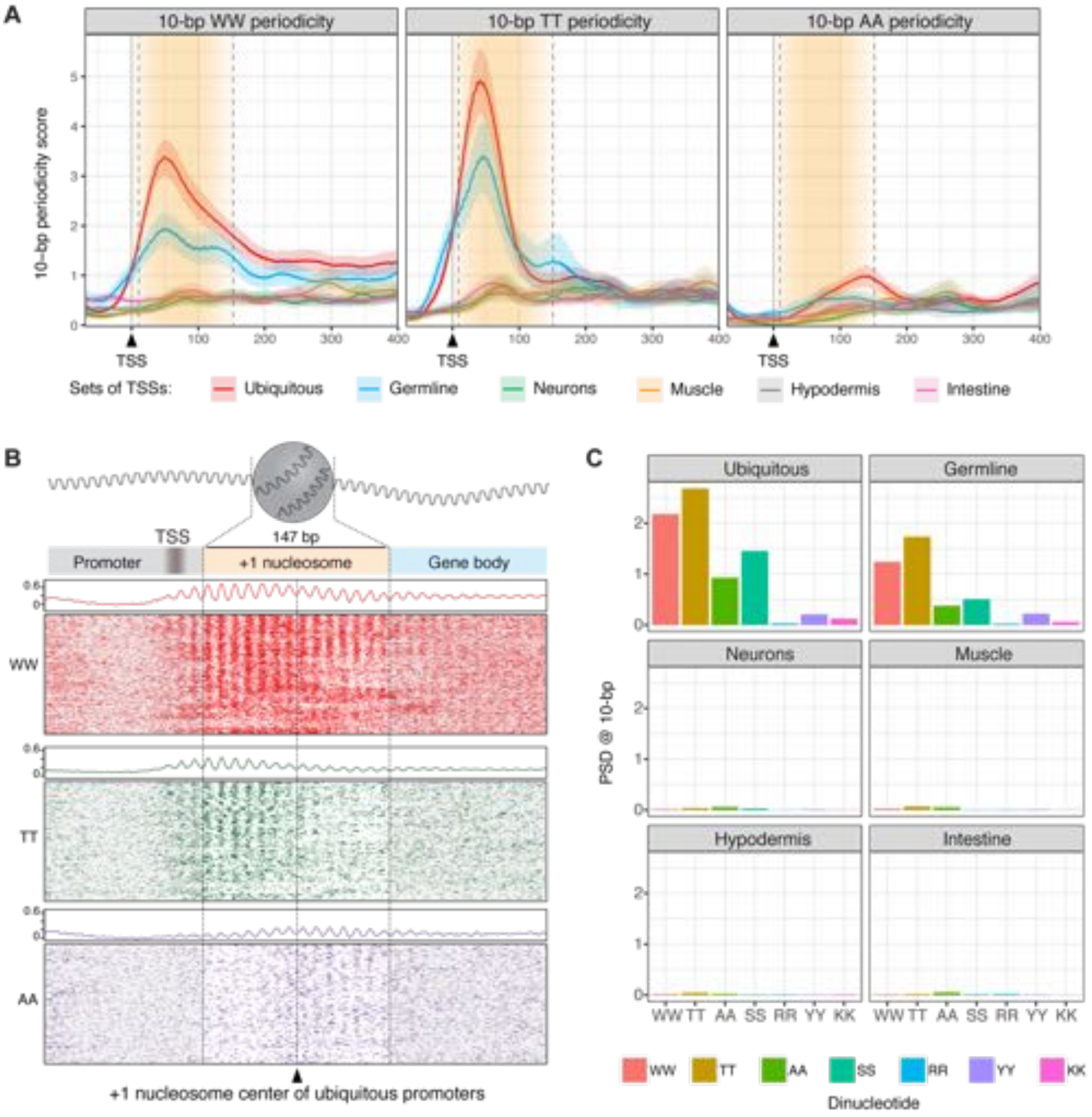


FIGURE 6.10 – 10-bp periodicity of different dinucleotides at ubiquitous and tissue-specific TSSs. **A-** Aggregate plots of 10-bp WW, TT and AA periodicity tracks at ubiquitous and tissue-specific promoters aligned at their TSS. Only the forward promoters are considered here. The reverse promoters show similar results. **B-** WW, TT and AA occurrences in +1 nucleosomal sequences of ubiquitous promoters. The sequences have been aligned at the +1 nucleosome dyad and shifted by a maximum of 5 bp to highlight the periodic occurrence of dinucleotides. The summed occurrences are displayed on top of each heatmap. The average TSS positions of ubiquitous promoters (~20 bp upstream of the +1 nucleosome edge) are displayed by the shaded gray area. See Figure 6.11 for other dinucleotides and other promoter classes. **C-** Power Spectral Density values at a period of 10 bp, for different dinucleotides in each set of promoters.

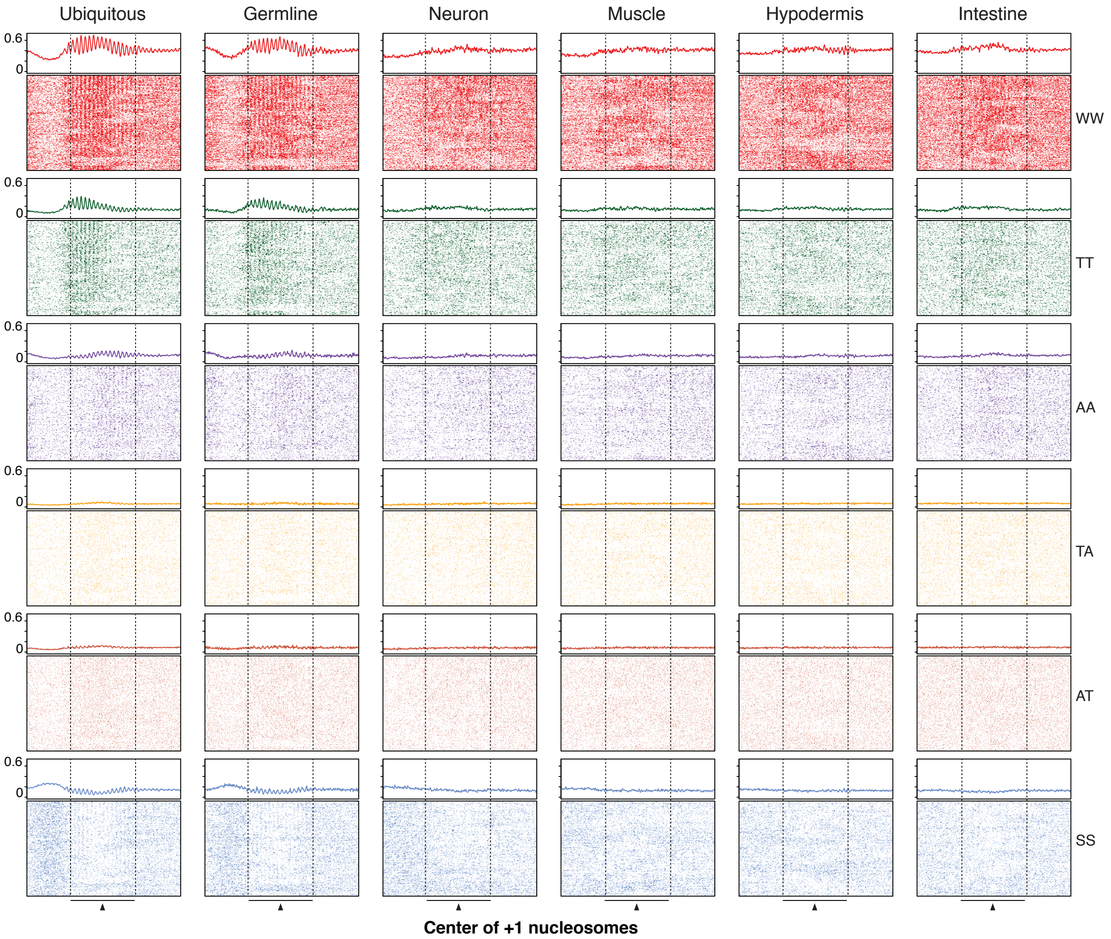


Figure 6.11 – WW, TT, AA, TA, AT and SS dinucleotide occurrences observed at +1 nucleosomes of ubiquitous or tissue-specific promoters (400 bp window centered at nucleosome dyads). The sequences have been aligned at the +1 nucleosome dyad and shifted by a maximum of 5 bp to highlight the periodic occurrence of dinucleotides. The summed occurrences are displayed on top of each heatmap.

promoters, 10-bp WW periodicity strikingly coincides with the position of +1 nucleosomes (Figure 6.10B and Figure 6.11). The 10-bp TT dinucleotide periodicity is the major contributor of the overall 10-bp WW periodicity and is skewed toward the 5'-half of +1 nucleosomes, while a weaker 10-bp AA periodicity peaks over the 3'-half of +1 nucleosomes (Figure 6.10). SS also appeared to occur over +1 nucleosomes of ubiquitous and germline-specific promoters, albeit to a lesser extent (Figure 6.10C and Figure 6.11). Of note, other dinucleotides did not have a strong 10-bp periodicity over these classes of promoters (Figure 6.10, Figure 6.11).

Finally, I sought to investigate whether the 10-bp TT periodicity was correlated with nucleosome occupancy. I grouped +1 nucleosomes of ubiquitous and germline-specific promoters into bins of 20 +1 nucleosomes, based on their occupancy score.

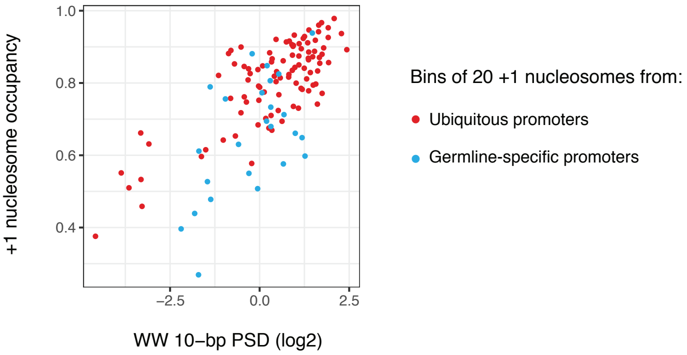


Figure 6.12 – Correlation between +1 nucleosome occupancy and 10-bp WW periodicity in ubiquitous and germline-specific promoters. +1 nucleosomes were binned by their nucleosome occupancy score and the overall 10-bp WW periodicity was assessed in each bin (~ twenty 200-bp long nucleosomal sequences centered at nucleosome dyads). The y axis represents the average nucleosome occupancy in each bin.

I then computed the power spectral density score of 10-bp WW periodicity in the nucleosomal sequences of each bin. I observed a strong correlation (Pearson's r = 0.71) between average nucleosome occupancy of each bin and the strength of 10-bp WW periodicity (Figure 6.12). Therefore, the strength of 10-bp periodic WW signal over +1 nucleosomes at ubiquitous and germline-specific (*i.e.* all the germline-active) promoters is positively correlated with nucleosome occupancy.

Importantly, these results also highlight the absence of any 10-bp periodic dinucleotide signal at somatic-tissue-specific promoters (Figure 6.9, Figure 6.10 and Figure 6.11). This is in line with the absence of positioned +1 nucleosomes at these promoters (Figure 6.3, 6.4 and 6.6).

6.2.2 Positioning of other sequence features at promoters

Other sequence features are found to be enriched at promoters. The Inr initiator sequence, the Sp1 motif and the TATA-box are three well-known core promoter elements that have already been characterized in *C. elegans* promoters (Chen *et al.*, 2013; Saito *et al.*, 2013). I investigated the position of these sequences and their enrichment within sets of ubiquitous or tissue-specific promoters. I found that the Inr motif is enriched in all promoter classes, however, somatic tissue-specific promoters showed higher Inr enrichment than ubiquitous and germline-specific promoters (Figure 6.13, 1.55-fold enrichment, p-value = 6e-14). I further observed that Sp1 and TATA box motifs are both predominantly associated with somatic-

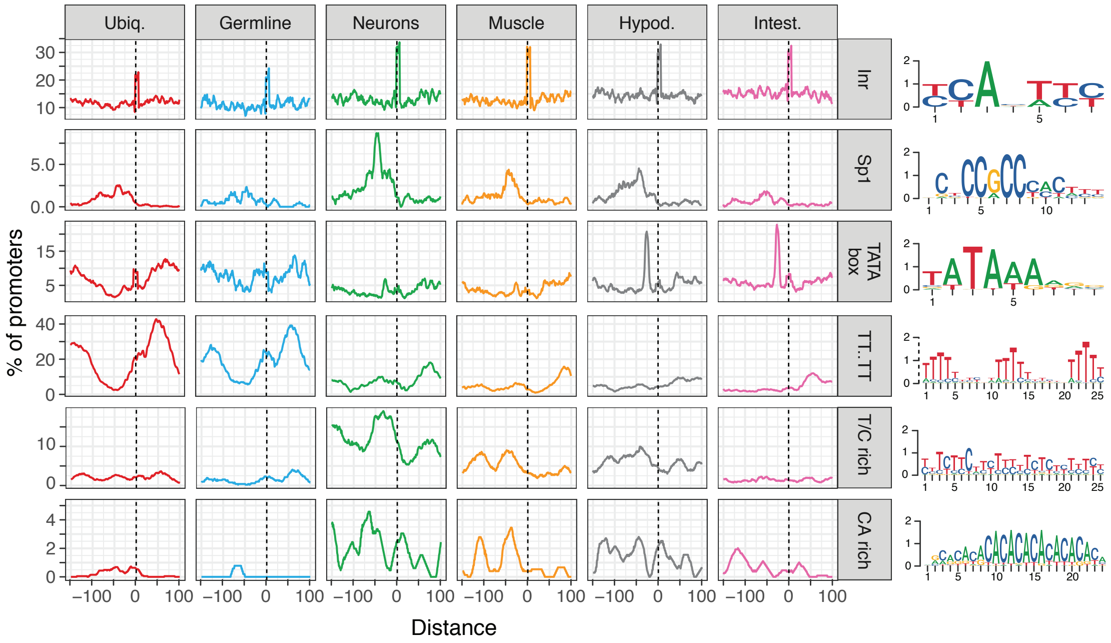


FIGURE 6.13 – Position of major core promoter elements and sequence biases at ubiquitous and tissue-specific promoters aligned to the TSS. Motif Position Weight Matrices (PWM) are displayed on the right. Only promoters with experimentally defined TSSs were considered.

tissue-specific promoters, with striking tissue biases. The Sp1 motif, peaking at -45 bp upstream of the TSS, is enriched at neural, muscle and hypodermal promoters but not at intestinal promoters, whereas the TATA-box motif was predominantly found at hypodermal and intestinal promoters, peaking at -30 bp upstream of the TSS. This enrichment of Sp1 and TATA-box in some but not all of the tissue-specific classes of promoters had not been thoroughly investigated before, likely due to the lack of clear annotation of ubiquitous and tissue-specific sets of promoters in other organisms. Finally, *de novo* motif analysis also revealed that somatic tissue-specific promoters share two dinucleotide composition biases, a T/C-rich stretch and a $(CA)_n$ dinucleotide repeat. Again, these two biases are not found in ubiquitous or germline-specific promoters (Figure 6.13).

The *de novo* motif analysis also uncovered motifs associated with tissue-specific promoters (Figure 6.14 and Figure 6.8 on page 146). For example, as expected, many intestinal promoters harbor the GATA motif bound by ELT-2, while the HLH-1 motif is found specifically at muscle promoters (Figure 6.14; Chen *et al.*, 1994; McGhee *et al.*, 2007). Many of these motifs have peak positions within the NDR, often ~45 bp upstream of the TSS (Figure 6.14).

Altogether, these results highlight the tissue-specific differences in both core

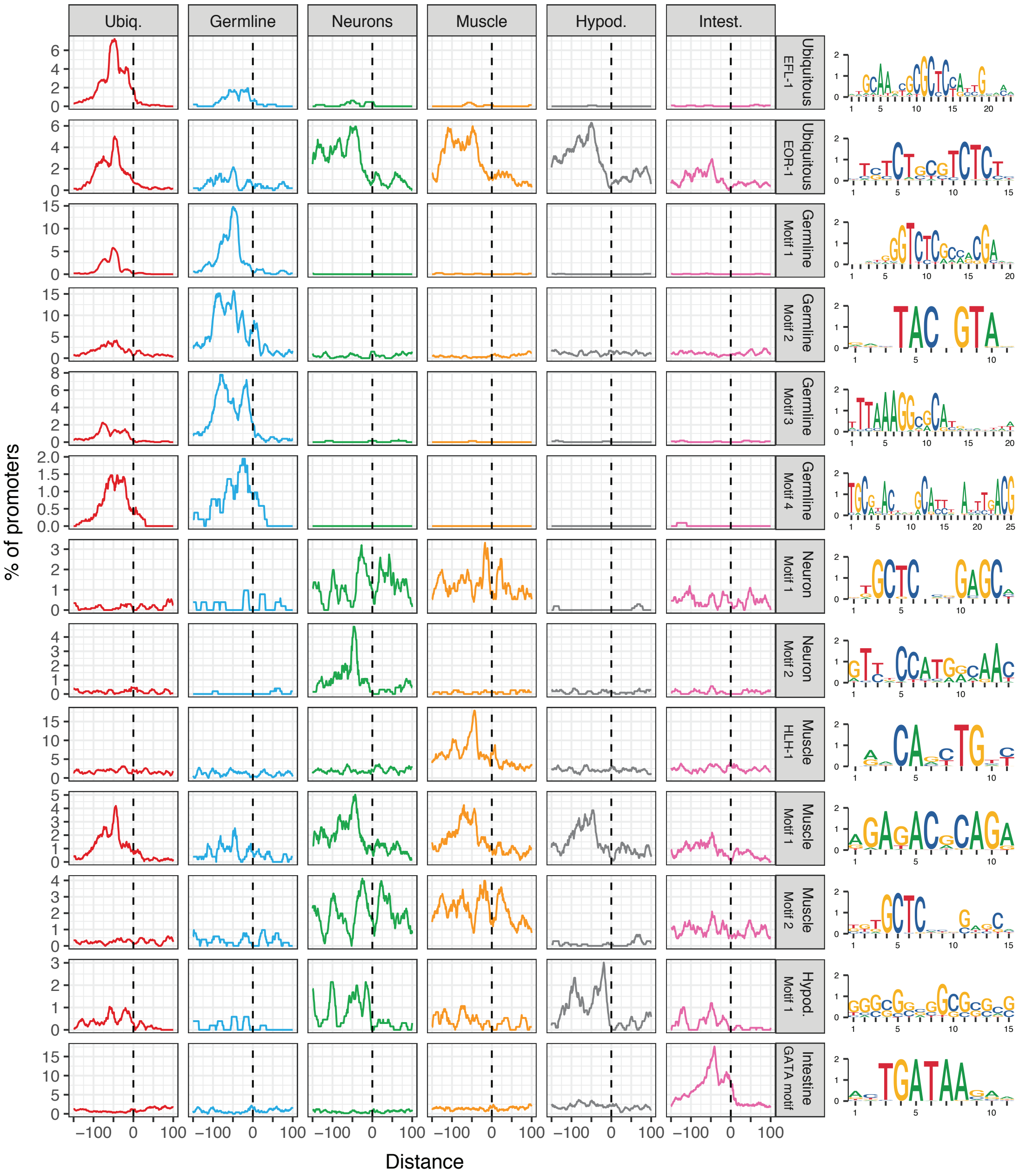


FIGURE 6.14 – Position of DNA motifs enriched at ubiquitous and tissue-specific promoters aligned to the TSS. Motif Position Weight Matrices (PWM) are displayed on the right. Only promoters with experimentally defined TSSs were considered.

promoter elements and TF binding motifs, and bring further insight in the sequence features which may be important for the activity of tissue-specific and/or ubiquitous promoters.

6.3 10-bp WW periodicity at ubiquitous promoters is a feature of non-mammalian genomes

Finally, I wondered whether the 10-bp periodic WW signal is a feature associated with +1 nucleosomes of ubiquitous promoters in other animals. 10-bp periodic WW sequences have been observed or suggested at +1 nucleosomes in yeast, fly, zebrafish and mammals (Albert *et al.*, 2007; Forrest *et al.*, 2014; Haberle *et al.*, 2014; Ioshikhes *et al.*, 2011; Mavrich *et al.*, 2008b; Tolstorukov *et al.*, 2009; Wright and Cui, 2019). However, the specific association of a 10-bp WW signal with different sets of promoters regulating genes with particular patterns of expression has rarely been directly investigated.

I first examined sequences around TSSs of all annotated genes in fly, zebrafish, mouse, and human. I could detect an increase of 10-bp WW periodicity downstream of fly and zebrafish TSSs (Figure 6.15A), as expected from previous reports (Haberle *et al.*, 2014; Mavrich *et al.*, 2008b). As in *C. elegans*, the WW periodicity signals in fly and zebrafish peaked in the 5' half of +1 nucleosomes. However, no comparable increase of 10-bp WW periodicity downstream of TSSs was present around mouse or human promoters (Figure 6.15A).

I then investigated subsets of promoters, to ask (i) whether 10-bp WW periodicity is associated with ubiquitously active promoters and (ii) if it is enriched in ubiquitous promoters compared to promoters with regulated activity. Using the coefficient of variation of gene expression (CV) metric, I considered genes in the bottom 20% of CV values to have broad ubiquitous expression and those in the top 20% to have highly regulated expression (*e.g.* tissue specificity). I then quantified WW 10-bp periodicity in each of the two sets of promoters and for each organism, as described in 6.2.1 on page 146. Though the periodicity strength is generally lower in fly and zebrafish compared to *C. elegans*, I found that in each organism, WW 10-bp periodicity was higher at promoters of broadly expressed genes than at those of highly regulated genes (Figure 6.15B-C). In contrast, neither the broadly

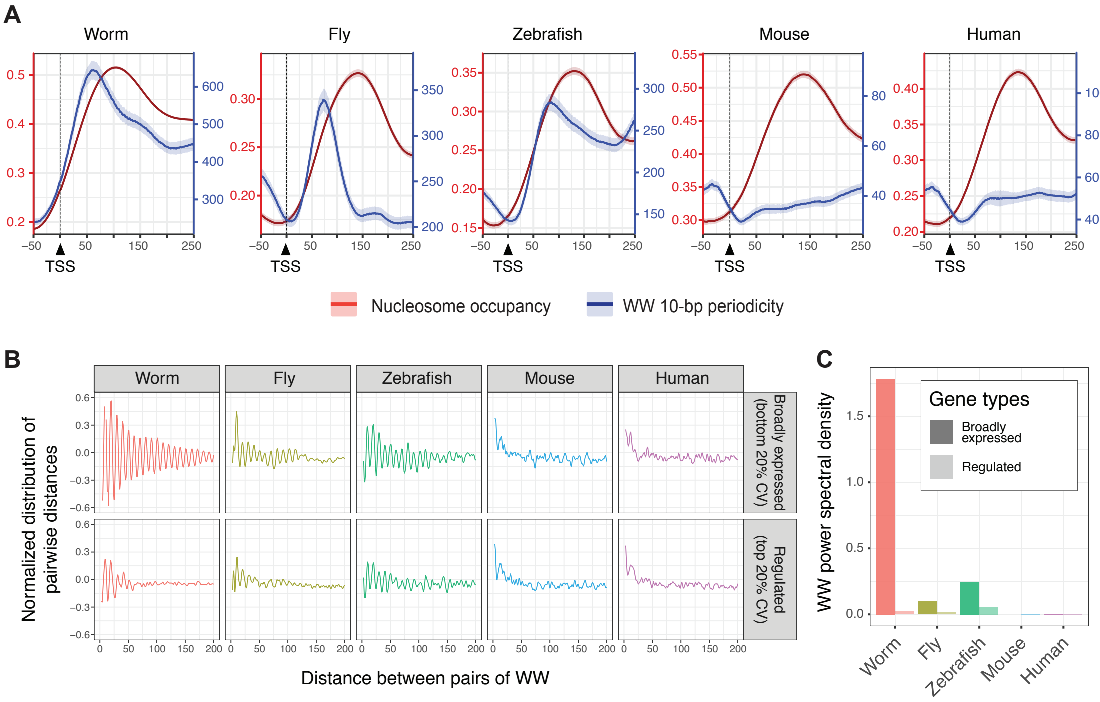


Figure 6.15 – 10-bp WW periodicity at ubiquitous and tissue-specific TSSs. **A-** 10-bp WW periodicity and nucleosome occupancy tracks are plotted over TSS, for all the genes annotated in worm, fly, zebrafish, mouse and human. **B-** Normalized distribution of distances between pairs of WW dinucleotides (WW....WW) found in sequences -50 bp to + 300 bp around TSSs of worm, fly, zebrafish, mouse and human genes. Two sets of TSSs have been analyzed: those associated with genes broadly expressed in the organism (top row: coefficient of variation of expression in bottom 20%) and those associated with genes highly regulated in the organism (bottom row: coefficient of variation of expression in top 20%). The distribution of distances is normalized as explained in 8.2 on page 178. **C-** Power spectral density values of 10-bp WW periodicity.

active nor the regulated groups of mouse and human promoters have comparable WW periodicity signals (Figure 6.15B-C). These results suggest that 10-bp WW periodicity signals are a conserved feature of ubiquitously active promoters in non-mammalian animals, possibly lost in mammals throughout evolution.

6.4 Discussion

Historically, promoters have been grouped based on their sequence or their pattern of transcription initiation (Carninci *et al.*, 2006; Lenhard *et al.*, 2012). This proved to be helpful to redefine the textbook promoter structure. Using such classification, multiple features including intrinsic DNA sequence, chromatin remodelers, DNA binding proteins, and RNA polymerase machinery have been shown to be associated with different types of promoter structures (reviewed in Haberle and Lenhard, 2016). Still, few studies have directly focused on comparing promoter organization based on their spatiotemporal activity. Classifying *C. elegans* promoters into functional sets of ubiquitous, tissue-restricted and tissue-specific promoters, I could directly compare the characteristics of each group of promoters.

Investigating these different classes, I found that strong +1 nucleosome position coinciding with 10-bp periodic WW signal is a key feature of ubiquitous and germline-specific promoters. The association of 10-bp WW periodicity and nucleosome position was first noted by Travers and colleagues in chicken, and is thought to aid nucleosome positioning by conferring sequence-dependent anisotropic bendability to the DNA polymer (Drew and Travers, 1985; Trifonov, 1980; Zhurkin *et al.*, 1979). Since then, this periodicity has been observed in nucleosomal sequences in different eukaryotes including *C. elegans,* but its specific association with +1 nucleosomes of different promoter types was unknown (Albert *et al.*, 2007; Dreos *et al.*, 2016; Field *et al.*, 2008; Forrest *et al.*, 2014; Haberle *et al.*, 2014; Ioshikhes *et al.*, 1996, 2011; Johnson *et al.*, 2006; Mavrich *et al.*, 2008a,b; Peckham *et al.*, 2007; Pich *et al.*, 2018; Satchwell *et al.*, 1986; Segal *et al.*, 2006; Struhl and Segal, 2013; Tolstorukov *et al.*, 2009; Wang and Widom, 2005; Widom, 2001; Wright and Cui, 2019). In contrast to ubiquitous and germline-specific promoters, I found that +1 nucleosomes of *C. elegans* somatic tissue-specific promoters are not associated with a 10-bp WW periodicity signal, have lower occupancy, and inconsistent position relative to the TSS. Instead, I observed intriguing biases

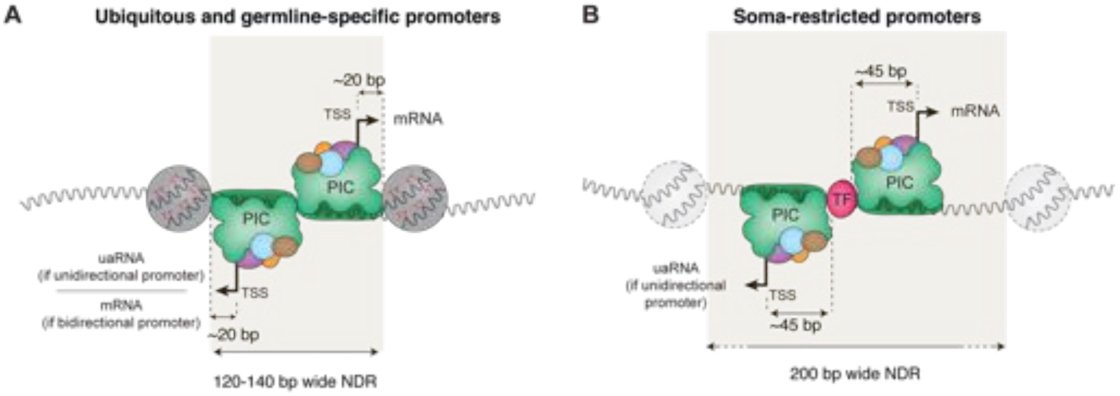

A
Ubiquitous and germline-specific promoters
~20 bp
TSS
mRNA
PIC
PIC
uaRNA
(if unidirectional promoter)
mRNA
(if bidirectional promoter)
TSS
~20 bp
120-140 bp wide NDR
B
Soma-restricted promoters
~45 bp
TSS
mRNA
PIC
TF
PIC
uaRNA
(if unidirectional promoter)
TSS
~45 bp
200 bp wide NDR

FIGURE 6.16 – Two models of PIC positioning at promoters. The nucleosome organization and sequences features found in ubiquitous, germline-specific and somatic-tissue-specific promoters suggest that two models of Pre-Initiation Complex recruitment exist. **A-** In ubiquitous and germline-specific promoters (*i.e.* all germline-active promoters), nucleosomes flank a narrow 120 to 140 bp-wide NDR. Positioning of these nucleosomes is facilitated by the underlying DNA sequence which harbors highly periodic WW (mainly TT) dinucleotides. Thus, the Pre-Initiation Complex (PIC) assembling at the NDR is physically constrained by the +1 nucleosome edge, resulting in transcription initiation ~20 bp upstream of the +1 nucleosome edge. Many of these promoters lead to bidirectional elongative transcription. Otherwise, upstream-antisense RNA (uaRNA) are transcribed. **B-** In soma-restricted promoters, NDRs are wider (> 200 bp) and flanking nucleosomes are weakly positioned and not reproducibly aligned relative to the TSS. Core and transcription factors recruited to the NDR facilitate assembly and positioning of the PIC, resulting in transcription initiation -45 to -50 bp downstream.

in the enrichment of core motifs at these promoters. TATA boxes are primarily found in hypodermal and intestinal promoters whereas Sp1 motifs are most highly enriched in neuronal promoters. In addition, tissue-specific motifs are present, and these often have peak positions around -50bp relative to the mode TSS. Overall, these results are in agreement with the model whereby promoters with high transcriptional plasticity have well-positioned flanking nucleosomes but those with low transcriptional plasticity do not (Tirosh and Barkai, 2008).

Structural studies of the Pre-Initiation Complex (PIC) suggest that it physically interact with DNA from -45 bp to +20 bp relative to the transcription start site (Louder *et al.*, 2016; Robinson *et al.*, 2016; Schilbach *et al.*, 2017). Interestingly, at *C. elegans* ubiquitous and germline promoters, the 5' edges of +1 nucleosomes are located roughly +20 bp downstream of the TSS, which would be at the 3' end of the PIC. This supports a model initially proposed in yeast whereby a positioned +1 nucleosome could facilitate PIC complex assembly by interacting with TFIID (Figure 6.16A) (Jiang and Pugh, 2009). In contrast, at soma-specific promoters which lack strongly positioned nucleosomes, transcription factors might help to locally recruit and position the PIC to initiate transcription ~ 45 bp downstream of their binding site (Figure 6.16B). These models are not mutually exclusive and additional mechanisms also contribute to promoter activity.

Similar to *C. elegans*, I observed that a 10-bp WW periodicity signal is also associated with promoter +1 nucleosomes of broadly expressed genes in zebrafish and fly. This is consistent with a previously described enrichment of 10-bp periodicity in AA and TT dinucleotides downstream of constitutively expressed promoters in zebrafish zygote (Haberle *et al.*, 2014). A weak genome-wide AA/TT periodicity was previously noted in fly but not associated with any class of gene (Mavrich *et al.*, 2008b). In contrast, the periodic WW signal is not detected at promoters of broadly expressed genes in mouse and human, despite their having well positioned +1 nucleosomes. This is consistent with reports showing a low 10-bp WW periodicity in mammal genomes, either around TSSs (Tolstorukov *et al.*, 2009; Wright and Cui, 2019) or genome-wide (Pich *et al.*, 2018).

From these observations, 10-bp WW periodicity seems to contribute to +1 nucleosome positioning at ubiquitously active promoters of non-mammalian eukaryotes, especially those of genes with basal cell functions. 10-bp WW periodicity is also found at promoters in yeast (Mavrich *et al.*, 2008a; Travers *et al.*, 2010), suggesting that this would be an ancient conserved mechanism to regulated housekeeping genes. In contrast, nucleosome positioning at promoters in mammals may rely on other mechanisms, whereas the WW 10-bp periodicity is relatively weaker. This could be linked to the increase of CpG islands in mammalian promoters.

Chapter 7

Gene regulation along *C. elegans* differentiation trajectories

By investigating tissue-specific ATAC-seq and RNA-seq datasets, I was able to identify fundamental differences in gene architecture and promoter organization in individual tissues (Chapter 4, 5 and 6). These datasets were generated from nuclei sorted from bulk tissues in young adult worms. Such approach is poorly fitted to study the mechanisms of gene regulation along cell differentiation trajectories. In Chapter 7, I describe the potential of single-cell-based approaches to study embryonic development and organogenesis. I also present the pilot studies I conducted using single-cell-based techniques to study gene regulation during cell fate determination in *C. elegans* and introduce the challenges and future directions of the project I initiated.

Collaboration note: Yan Dong prepared the collection of live embryos (up to 100-cell stage embryos) used for single-nucleus genomic assays.

7.1 Studying gene regulation in embryogenesis with single-cell approaches

Three main processes take place during embryonic development:

- Transition from maternal to zygotic control of development;
- Cell fate specification;

- Morphogenesis/organogenesis, which leads to the formation of functional tissues.

Precise control of gene expression is crucial for the correct progression of these steps, and relies on many different parameters, from cytoplasmic determinant to pioneer transcription factors. The mechanisms of gene regulation during embryogenesis have been at the center of developmental studies for the past decades, but most of these studies focus on a particular process, such as the genetic control of the development of the vulva in *C. elegans* (Kornfeld, 1997) or the control of the development of the eye in *D. melanogaster* (Bessa *et al.*, 2002).

With the emergence of single-cell approaches, it appears feasible in principle to determine the regulation of chromatin, gene expression, and nuclear organization in every single cell from the zygote to the terminally differentiated cells. *C. elegans* is the only model organism where it would be realistically possible to identify mother and daughter cells throughout the entire embryonic development, as its cell lineage is small, invariable and fully known. Thus, using single-cell methods in *C. elegans* could help determining the mechanisms underlying the regulatory changes occurring during embryonic development. This would reveal the principles by which the genome directs cell fate specification and organogenesis.

7.1.1 Focusing on cell fate specification during early embryogenesis

Single-cell approaches are ideal to study the mechanisms of gene regulation involved in specification during *C. elegans* early embryogenesis, but two points are crucial when designing these assays. Firstly, cell types with unique characteristics are formed at each division, especially during early stages of the embryonic development. For instance, specification in the E lineage is initiated in the first hour after fertilization of the oocyte and involves a cascade of transcription factors differently expressed in the EMS progenitor cell, the E blastomere and the E daughter cells (Figure 7.1). Secondly, because maternal RNA molecules are loaded in the fertilized oocyte and inherited in daughter cells for several cell divisions, profiling cytoplasmic mature mRNA does not directly reflect the transcriptional activity of the zygote. These aspects of embryonic development need to be taken into account when designing the experimental approaches.

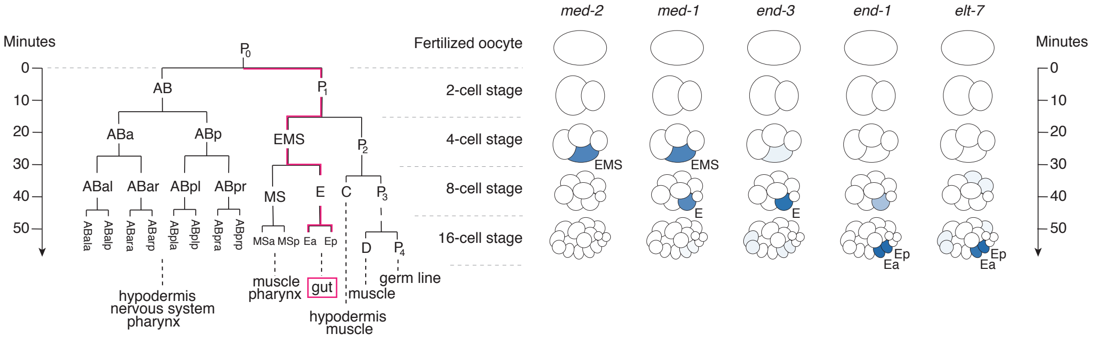


Figure 7.1 – Transient expression of transcription factors during specification. Left: cells are rapidly dividing during the first hours of embryogenesis. The vertical axis indicates time of development (from the first cleavage) at 25 C. The E lineage is highlighted in pink. Right: expression of different transcription factors involved in the E lineage specification during early embryonic development, measured by single-cell RNA-seq of hand-dissected embryos (Tintori *et al.*, 2016).

Two recent studies used single-cell-based approaches to profile gene expression during *C. elegans* embryogenesis. Using the 10X Genomics platform, Packer *et al.* (2019) profiled the transcriptome of ~80,000 individual cells throughout embryogenesis. Using SMART-seq technology, Tintori *et al.* (2016) profiled the transcriptome of 219 cells from 1-cell to 16-cell stage embryos. These two datasets are currently the reference to study genome-wide transcriptomic changes during *C. elegans* embryogenesis at single-cell resolution, but they present several conceptual limitations. First, they focus either on the very first minutes of embryonic development (1-cell to 16-cell embryos) or on the later stages (200-cell to 550-cell embryos), and thus cannot fully characterize gene expression during cell specification (under 100-cell stage embryos). Secondly, these studies mostly quantify cytoplasmic transcript abundance and thus do not measure the transcriptional activity of the zygote. Thirdly, they only focused on transcript quantification and do not provide any insight in the dynamics of chromatin organization during development, *e.g.* in terms of accessibility.

Between 1-cell and 32-cell stages, the embryos undergo drastic changes during Zygotic Genome Activation (ZGA) and quickly after ZGA, lineage specification occurs, restricting populations of cells to a determined fate. I aimed to leverage RNA-seq and ATAC-seq single-cell approaches to investigate the mechanisms required for transcription regulation during ZGA and early cell specification, in individual nuclei of *C. elegans*.

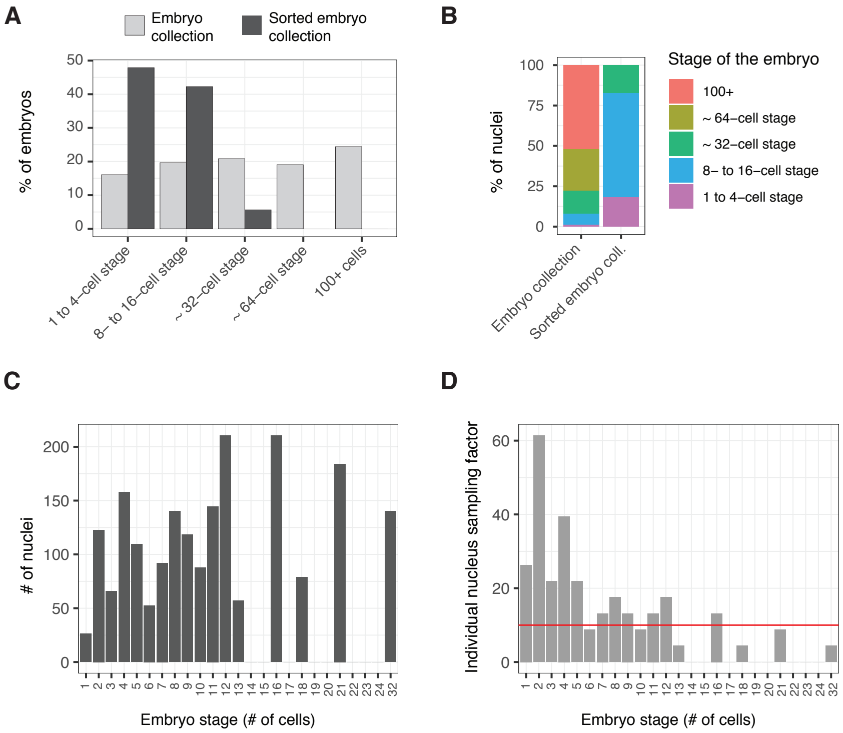


Figure 7.2 – Nuclei obtained from a collection of embryos with or without enrichment of 1- to 32-cell stage embryos. **A-** Distribution of embryo stages in a collection before and after enriching for early embryos (1- to 32-cell stages) by sorting. **B-** Estimated proportion of nuclei from different embryo stages after nuclei isolation, before and after enriching for early embryos by sorting. **C-** Estimated number of nuclei from individual embryo stages, assuming sequencing of 2000 individual nuclei obtained from the sorted embryos. **D-** Estimated sampling of individual nuclei from each embryo stages, assuming sequencing of 2000 individual nuclei obtained from the sorted embryos. The red line represents 10-fold oversampling of each nucleus of an embryo at a given stage.

7.1.2 Profiling gene expression and chromatin accessibility in single nuclei

To study ZGA and early cell specification, I sought to profile both zygotic transcriptional activity and chromatin accessibility in individual nuclei from early embryos, between 1-cell and 32-cell stage or up to 100-cell stage. However, harvesting very early embryos (< 32-cell stage) is challenging and early embryo collections typically contain at least 20% of ~ 100-cell embryos (Figure 7.2A). This is a major issue in single-nucleus experiments, as the nuclei from these late embryos would represent more than 50% of the entire set of extracted nuclei (Figure 7.2B). I first attempted to isolate early frozen embryos (< 32-cell stage embryos) by sorting. To specifically obtain very early (1-cell to 32-cell stages) embryos from a population of mixed embryos, I optimized a cytometry-based sorting method. Staining frozen embryos by DAPI, I could distinguish early embryos (with relatively low DAPI signal) from older embryos (with greater DAPI signal). Thus, I could effectively sort embryos between the 1-cell and the 32-cell stage (Figure 7.2A-C). Overall, nuclei from each cell type and at any embryonic stage (up to the 32-cell stage) are well represented when enriching for very early embryos by sorting(Figure 7.2D).

Using the 10X Genomics workflow, I then performed snATAC-seq and snRNA-seq on nuclei obtained from (i) 1- to 32-cell stage sorted frozen embryos or (ii) up to ~100-cell stage unsorted live embryos. I aimed to sequence ~ 2,000 individual nuclei, which should be sufficient to achieve a ~ 10-fold oversampling of nuclei from each cell type at any embryo stage in the sorted embryos (Figure 7.2D). This should be enough to get preliminary insights on whether we can successfully perform single-cell experiments and identify individual cells from the early cell lineage tree. I recovered 4,945 individual nuclei after snRNA-seq (2,072 from sorted early embryos and 2,873 from unsorted older embryos) and 2,316 individual nuclei after snATAC-seq (1,115 from early embryos and 1,201 from older embryos).

To understand whether the snRNA-seq or snATAC-seq experiments efficiently captured transcript abundance or chromatin accessibility, I plotted histograms of Unique Molecular Identifiers (UMI) counts per nucleus. The number of UMIs per nucleus is a good metric to get preliminary insights about whether a single-cell experiment worked or failed. This revealed that the two snRNA-seq runs had a satisfactory distribution of UMIs per nucleus, suggesting that they both worked (Figure 7.3A). Nuclei from early embryos appeared to have a slightly lower transcript content compared to older embryos (Figure 7.3A). Correspondingly, there are overall slightly fewer detected genes in nuclei from sorted early embryos than in nuclei from older embryos (Figure 7.3B). This could represent the gradual zygotic genome activation (ZGA) during the early division cycles after oocyte fertilization. However, this could also be a technical issue arising from using sorted frozen embryos.

The distribution of UMI counts per nucleus is drastically different between the two snATAC-seq experiments. Nuclei from sorted early embryos have on average ~ 200 UMI / nucleus while those from older embryos have on average ~ 5000 UMI / nucleus (Figure 7.3C). This results in only ~ 100 accessible regions detected per nuclei on average in snATAC-seq from sorted early embryos, compared to almost 3,000 in snATAC-seq from older embryos (Figure 7.3D). This suggests that the snATAC-seq experiment performed on nuclei from sorted early embryos failed,

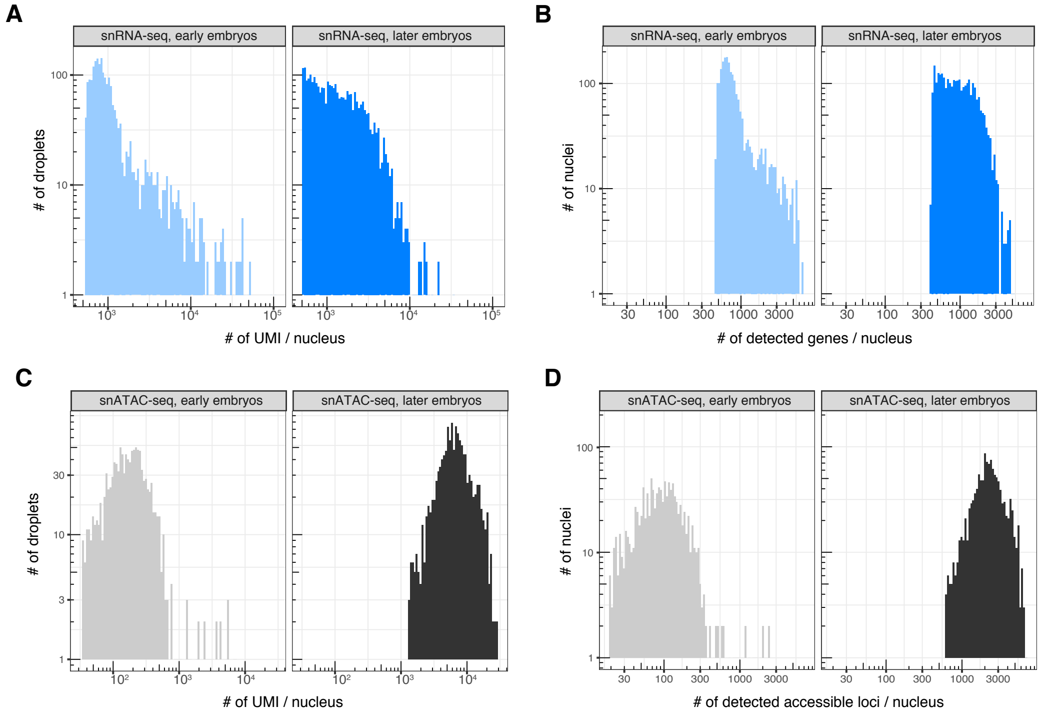


Figure 7.3 – Quality control of snRNA-seq and snATAC-seq pilot experiments. **A-** Distribution of UMI counts per nucleus in snRNA-seq pilot experiments. **B-** Distribution of number of genes detected per nucleus in snRNA-seq pilot experiments. **C-** Distribution of UMI counts per nucleus in snATAC-seq pilot experiments. **D-** Distribution of number of accessible regions detected per nucleus in snATAC-seq pilot experiments. Early embryos are from 1- to 32-cell stage and older embryos are up to ~ 100-cell stage.

potentially because of the embryo sorting step or the fact that these embryos were frozen prior to sorting and nuclei isolation.

Taken together, these metrics suggest that it is possible to enrich for early embryos by sorting frozen embryos and still get good quality snRNA-seq data. However, freezing embryos prior to nuclei isolation and/or enriching for early embryos by sorting may slightly reduce the number of detected genes by snRNA-seq and masking true biological observations. Furthermore, these steps seem to severely impact snATAC-seq. Thus, further controls using synchronous populations of embryos, frozen or not and sorted or not, will need to be performed to understand which of these steps have an impact on snRNA-seq and snATAC-seq (see 7.1.2 on page 162 for further discussion).

7.1.3 Cell markers and differentiation trajectories

To see whether snRNA-seq and snATAC-seq datasets could be leveraged to get useful insights in gene regulation during early embryogenesis in *C. elegans*, further processing of the raw data is required. I processed, filtered and reduced the dimensionality of the snRNA-seq and snATAC-seq datasets using the monocle3 suite (Cao *et al.*, 2019). In the subsequent preliminary analyses of the single-nucleus experiments, I decided to merge the two snRNA-seq datasets together and focus on the snATAC-seq dataset obtained from older embryos.

I embedded the ~ 5,000 individual nuclear transcriptomes into two dimensions using UMAP (Figure 7.4). This revealed that the two snRNA-seq samples have a reasonable overlap, although some populations of nuclei were only found in one of the two samples. This is expected, as each sample is enriched either for early or older embryos. I then clustered the embedded single nuclei using a graph-based approach (Cao *et al.*, 2019). This identified 33 different clusters. To annotate some of these clusters, the transcript abundance of genes known to be specifically expressed in subsets of committed cells can be used. For instance, the well-known cascade of genes inducing cell fate of the E lineage through *med-2* – *end-1/3* – *elt-7* serial expression can be used to characterize cells from the E lineage (Figure 7.5A and Figure 7.1 on page 161). In the merged snRNA-seq dataset, *end-3* transcripts are enriched in clusters 1, 9 and 24, while *end-1* and *elt-7* transcripts are enriched in clusters 19 and 4 (Figure 7.5B). These observations suggest that clusters 1, 9 and

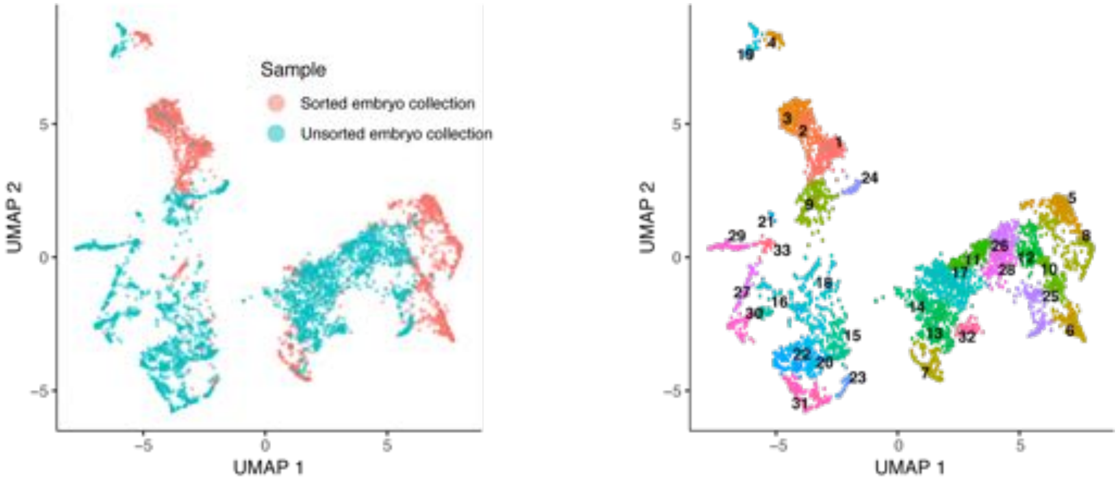


FIGURE 7.4 – UMAP projection of ~5,000 transcriptomes from individual nuclei. Each dot represents a single nucleus, colored according to its method of isolation (left) or its cluster (right).

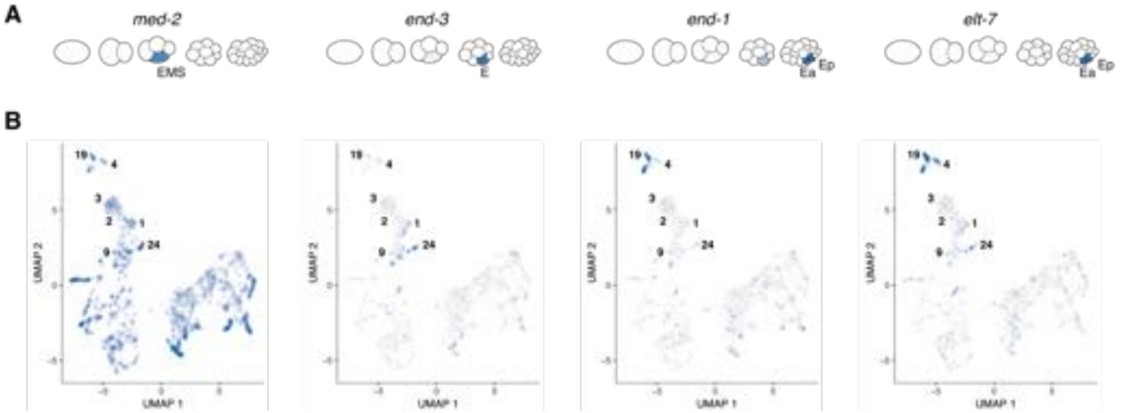


FIGURE 7.5 – Expression of E lineage-related marker genes. **A-** Patterns of expression of different E lineage-related marker genes during early embryonic development, measured by single-cell RNA-seq of hand-dissected embryos (Tintori *et al.*, 2016). **B-** Expression of the corresponding marker genes in ~5,000 individual nuclei. Each dot represents a single nucleus, colored based on the transcript abundance for the indicated gene (*med-2*, *end-3*, *end-1*, *elt-7*).

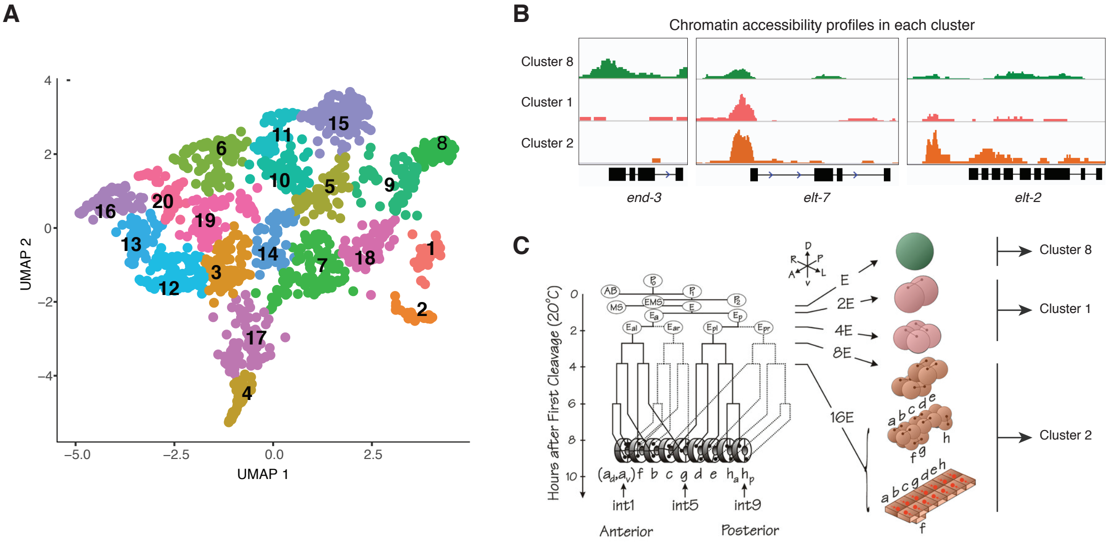


Figure 7.6 – UMAP projection of ~1,200 chromatin accessibility profiles from individual nuclei. Each dot represents a single nucleus, colored according to its cluster. **B-** Cluster-aggregated chromatin accessibility profiles at *end-3*, *elt-7* and *elt-2* loci. The scale is the same across the three tracks and adjusted at each locus. **C-** Annotation of clusters 8, 1 and 2.

24 may consist of E blastomere cells whereas clusters 4 and 19 may represent E daughter cells, Ea and Ep (Figure 7.5B). Importantly, *med-2* transcripts are found across most clusters even though *med-2* is thought to only be zygotically expressed in the EMS precursor. This could originate from abundant maternally inherited cytoplasmic *med-2* mRNA (Maduro *et al.*, 2007). The potential issue of maternal transcripts is further tackled in 7.2.

I also analyzed the ~ 1,200 individual chromatin accessibility profiles obtained by snATAC-seq performed in embryos with a wider range of cell number (1- to ~ 100-cell stage embryos). I embedded the nuclei into two dimensions and obtained 20 clusters. I then generated cluster-specific aggregated genome-wide tracks of chromatin accessibility (Figure 7.6). Here again, markers of individual cell types were found to be specifically accessible in individual clusters. For instance, the *end-3* promoter is accessible in nuclei from cluster 8, the *elt-7* promoter is accessible in nuclei from 3 clusters (cluster 8, 1 and 2) and the *elt-2* promoter is accessible in nuclei from cluster 2 (Figure 7.6B). Based on the known temporal patterns of expression of these genes (Figure 7.5), the clusters 8, 1 and 2 could respectively correspond to the original E blastomere (expressing *end-3,* cluster 8), its early daughter cells (expressing *elt-7*, cluster 1) and the cells appearing later in the E lineage (expressing *elt-7* and *elt-2*, cluster 2) (Figure 7.6C).

Overall, these preliminary observations suggest that single-nucleus RNA and ATAC sequencing can be used to annotate populations of cells across the cell lineage tree during early embryogenesis in *C. elegans*. However, several important technical questions have to be addressed before starting to generate large-scale datasets.

7.2 Challenges arising from single-nucleus approaches

The pilot experiments I conducted highlighted several shortcomings and technical challenges related to both the experimental design and the downstream computational analysis of single-nucleus genomic assays.

Modeling of the cytoplasmic RNA background in snRNA-seq To perform single-nucleus RNA-seq or ATAC-seq from early embryos, I prepare a suspension of single nuclei by breaking open embryos using a ball-bearing Balch homogenizer. This effectively breaks embryo shells and plasmid membranes. However, when cells burst, their cytoplasmic RNA content is released and mixed with the nuclei suspension. During 10X Genomics sample preparation, an emulsion of nuclei in suspension with the ambient cytoplasmic RNA is then obtained. The resulting cytoplasmic RNA "soup" generates a background which can significantly hamper subsequent biocomputational analyses (Kang *et al.*, 2018). Since cytoplasmic RNA is mostly maternally inherited in early embryos, the cytoplasmic RNA content does not reflect active zygotic transcription. This could result in a gene appearing to be transcribed in all the cells when it actually is (partially or totally) maternally inherited (*e.g. med-2*, Figure 7.5A). Tools such as souporcell (Heaton *et al.*, 2019) or SoupX (Young and Behjati, 2018) are being actively developed to measure and correct such cytoplasmic RNA background inherent to emulsion-based applications. These tools could potentially be used to estimate and remove the contribution of maternally inherited RNA in snRNA-seq from early embryos.

Improvement of single-nucleus genomic assays from sorted frozen embryos To study ZGA and the beginning of cell specification, focusing of nuclei from very early embryos is required. To do so, I have isolated for 1-cell to 32-cell stage embryos by sorting them from a frozen collection of embryos (see 7.1.2 on page 162). As stated above, this step may have reduced the quality of snRNA-seq and seems to have severely impacted snATAC-seq (Figure 7.3). Notably, prior to sorting, frozen embryos are stained with DAPI at 0.5 ng/ul, a higher concentration than what is typically used for nuclear sorting (0.025 to 0.05 ng/ul). High concentrations of DAPI, a DNA intercalating agent, could potentially inhibit the efficiency of Tn5-based tagmentation in snATAC-seq. This could be readily assessed by performing snRNA-seq and snATAC-seq on nuclei from DAPI- or DAPI+ embryos. Furthermore, using nuclei from frozen embryos may alter the efficiency of single-cell genomic assays, although successful experiments have been conducted using cryopreserved samples in combination with 10X Genomics workflows. This could also be assessed by performing snRNA-seq and snATAC-seq on nuclei from synchronous populations of live or frozen embryos.

If these control experiments confirm that sorting frozen embryos systematically impacts the efficiency of single-nuclei genomic assays, other approaches should be pursued. For example, the *goa-1* mutant strain could be used. Young adult *goa-1* mutants are characterized by a fast egg laying phenotype and only retain recently fertilized embryos. Thus, this strain could be used to collect very early live embryos, rather than relying on sorting of frozen embryos.

Automatic annotation of cluster identity The comparison of transcriptome (or chromatin accessibility) profiles across clusters allow the *a posteriori* annotation of individual clusters. This process typically relies on ground truth knowledge to infer the nature of a given cluster based on the pattern of expression of few markers (exemplified in Figure 7.5). Manual cluster annotation is a laborious process prone to errors and which can yield suboptimal results. On the other hand, tools such as Garnett (Pliner *et al.*, 2019) or scCATCH (Shao *et al.*, 2020) can use lists of marker genes known to be specifically expressed in a single cell type to automatically annotate individual clusters. Relying on the resources from the worm community (WormBase), compiling lists of marker genes in *C. elegans* should be possible and will help annotating clusters during cell fate specification and organogenesis.

Integration of snATAC-seq and snRNA-seq data In the last five years, emulsion-based single-cell (or single-nucleus) methods have been rapidly expanding (Zhang *et al.*, 2019). Yet, a unified methodology to analyze the different types of single-cell "omics" is still missing (Amezquita *et al.*, 2020; Stuart and Satija, 2019), and integrating snRNA-seq and snATAC-seq together remains a major challenge. New approaches such as canonical correlation analysis (CCA), multi-omics factor analysis (MOFA) or non-negative matrix factorization (NMF) are emerging and could be used to integrate different types of datasets together (*e.g.* single-nucleus RNA-seq and ATAC-seq) (Stuart *et al.*, 2019). Such tools will be helpful to investigate the dynamic relationship between chromatin architecture and gene expression during cell specification or organogenesis in *C. elegans*.

7.3 Toward a developmental single-nucleus atlas

The single-nucleus pilot experiments I conducted in very early (1-cell to 32-cell stage) and embryos up to a later developmental stage (1-cell to ~ 100-cell stage) are promising. I propose strategies to resolve the current technical hurdles and the analytical challenges inherent to single-nucleus RNA-seq and ATAC-seq experiments. Once these issues are addressed, libraries of tens of thousands of nuclei could be obtained and deeply sequenced using these single-nucleus approaches. This will enable the investigation of tissue-specific mechanisms of gene regulation during cell fate specification, ideally from 1-cell to 200-cell embryos.

In *C. elegans,* organogenesis takes place in the second half of the embryogenesis as well as during post-embryonic development. At this point, it becomes challenging to sample cells from each cell type in single-cell-based approaches (Packer *et al.*, 2019). To address this issue, I also plan on sorting nuclei from specific lineages (*e.g.* the intestine lineage) and then perform snRNA-seq and snATAC-seq on these sorted nuclei. Rather than superficially sampling all the cells of the ~ 550-cell embryos undergoing organogenesis, this approach would allow independent in-depth analysis of cell trajectories in each tissue separately, even for those with a relatively small cell lineage (*e.g.* the intestine lineage has only 20 terminally differentiated cells whereas the neurons account for more than 300 cells). Moreover, it would also enable the study of post-embryonic tissue-specific gene regulation in specific developmental contexts at the single-nucleus resolution, such as the activation of the M progenitor cell generating muscle cells during post-embryonic development (Krause and Liu, 2012).

Leveraging single-nucleus experimental approaches in early embryonic development as well as in individual tissues in late embryogenesis and post-embryonic development, a developmental single-nucleus cartography can be generated to investigate the principles by which the genome regulates cell fate specification and organogenesis.

Chapter 8

Computational tools and resources for genomics

During my PhD, I have developed two R packages, VplotR and periodicDNA, to support the analysis of the sequencing datasets I have generated. In Chapter 8, I describe their principles and illustrate how they can be used. I also present JABrowser, a cloud-based genome browser, and RegAtlas, a web application I developed to share the results of my investigation. Together, these utilities provide useful tools and facilitate the dissemination of the datasets I have generated during my PhD as public resources in the larger scientific community.

8.1 VplotR: R package to produce fragment density plots

8.1.1 Fragment lengths bear information

Libraries generated by MNase-seq, DNase-seq or ATAC-seq are usually sequenced in a paired-end manner. This was originally useful to select 147-bp long fragments corresponding to nucleosomal DNA in MNase-seq (*e.g.* Valouev *et al.*, 2008), but elegant approaches have also relied on MNase-seq fragment lengths to study transcription factor binding sites, as binding protects DNA from being cut (Henikoff *et al.*, 2011). The more recent ATAC-seq assay also generates genomic fragments from both nucleosome-spanning DNA and NDRs, resulting in a multi-modal distribution of short or longer fragments (Buenrostro *et al.*, 2013). Because longer

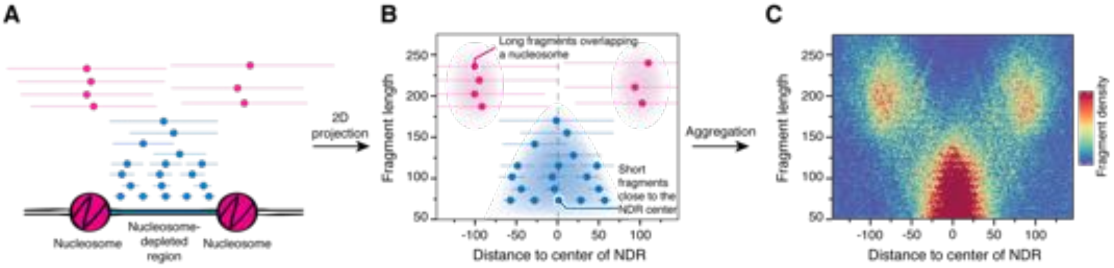

A
B
C
Nucleosome
Nucleosome-depleted region
Nucleosome
2D projection
Long fragments overlapping a nucleosome
Short fragments close to the NDR center
Fragment length
250
200
150
100
50
-100
-50
0
50
100
Distance to center of NDR
Aggregation
Fragment length
Fragment density
Distance to center of NDR

FIGURE 8.1 – Rationale of the VplotR package. **A-** Sequenced fragments (for instance obtained from ATAC-seq) mapping to a locus of interest can originate from either nucleosomal DNA (in pink) or from nucleosome-free DNA (for instance from NDRs, in blue). **B-** The fragments can be embedded in a two-dimension graph. The horizontal coordinate represents the distance from the center of a fragment to the center of a locus of interest (for instance the NDR). The vertical coordinate represents the length of the fragment. **C-** When this projection is done over hundreds of loci, it results in a fragment density plot, *i.e.* a matrix which can be visualized as a heatmap, with the color gradient representing the density of fragments at each set of coordinates.

fragments likely originate from nucleosome-spanning DNA, the distribution of fragments lengths can be used to infer the local arrangement of nucleosomes flanking an NDR (Schep *et al.*, 2015). Thus, integrating ATAC-seq fragment length component into existing analytical frameworks could bring additional insights.

8.1.2 VplotR can illustrate spatial distribution of fragment lengths

I developed VplotR (Serizay, 2020b), an R package which can be used to easily generate fragment density plots, inspired from the visualization approach originally known as “V-plot” (Henikoff *et al.*, 2011). In a fragment density plot, sequenced fragments are projected into a two-dimensional graph: the horizontal axis represents the location of the center of a fragment relative to the center of a locus of interest, while the vertical axis separates the fragments according to their length (Figure 8.1A, B). When multiple loci are aggregated together, the resulting plot represents the density of fragments using a color code (Figure 8.1C). Over nucleosome-depleted regions, such fragment density plot is helpful to highlight the position of flanking nucleosomes, for instance (Figure 8.1C).

The steps performed to generate a fragment density plot using VplotR are the following:

1. Import genomic loci of interest into a *GRanges* object. This is typically done

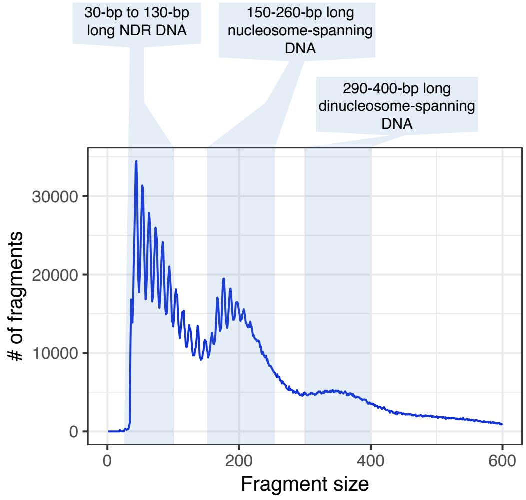


FIGURE 8.2 – Multimodal distribution of ATAC-seq fragment sizes

using methods from rtracklayer package:

```
loci_of_interest <- rtracklayer::import('file.bed')
```

2. Import fragments from a .bam file of paired-end reads aligned and filtered to a genome reference. This can be done using *importPEBamFiles()*, a VplotR function built on top of the Rsamtools package which generates *GRanges* objects from local .bam files.

```
fragments <- VplotR::importPEBamFiles(
    bam_file, where, shift_ATAC_fragments
)
```

- The where argument is used to only import fragments of the .bam file mapping to genomic loci of interest, to reduce computational load.
- The shift_ATAC_fragments boolean argument specifies whether the fragments should be shifted from their location; fragments originating from ATAC-seq experiments are traditionally shifted by -4 / +5 bp to account for Tn5 steric hindrance (Buenrostro *et al.*, 2013).

3. Check whether the fragments lengths show a multimodal distribution (Figure 8.2):

```
Vplot_matrix <- VplotR::getFragmentsDistribution(
        fragments, loci_of_interest
)
```

4. Initiate a *Vplot* object as follows:

```
Vplot_matrix <- VplotR::computeVmat(fragments, loci_of_interest)
```

5. Normalize the *Vplot* object as follows:

```
Vplot_matrix_normalized <- VplotR::normalizeVmat(
    Vplot_matrix, normFun, roll
)
```

- The normFun argument specifies how to normalize the *Vplot* matrix. The matrix can be scaled by dividing each cell by the sum of the entire matrix (normFun = 'pctsum'). This normalization approach is ideal to relatively compare multiple fragment density plots, to identify differences in fragment density patterns. Alternative normalization methods are currently being developed, notably to normalize different fragment density plots by the sequencing depth of the library used to generate each plot. This should allow a direct comparison of absolute density scores rather than relative patterns of fragment density.
- The roll argument specifies the binning window to apply to the matrix. If data is plotted over few genomic loci or if sequencing depth is relatively low, the resulting V-plots may appear grainy; this argument can be used to smooth it.

6. Compute the resulting fragment density plot as follows:

```
Vplot <- VplotR::plotVmat(Vplot_matrix_normalized, ...)
```

- The *plotVmat()* function can take several arguments to customize the appearance of the final fragment density plot. The package documentation provides more information about these arguments (https://js2264.github.io/VplotR).

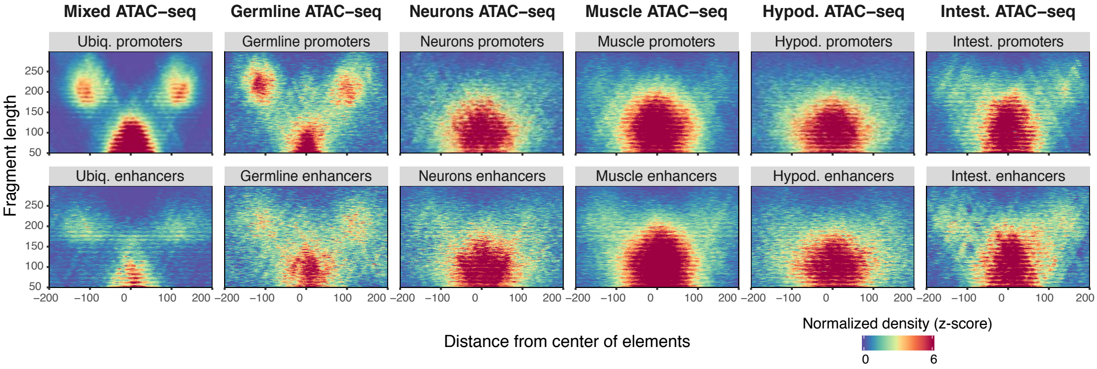


Figure 8.3 – ATAC-seq fragment density plots at ubiquitous or tissue-specific sets of promoters (top row) or enhancers (bottom row). Each plot is independently Z-scored.

- This function generates a *ggplot* object which can be further customized using the ggplot2 package.

The steps 4, 5 and 6 can be streamlined using the different methods defined in the *plotVmat()* function (see the VplotR documentation for more details). Among other methods, *plotVmat()* function can take a nested list of arguments to quickly generate multiple fragment density plots.

A detailed example of a concrete VplotR usage is shown in Appendix Chapter C.

8.1.3 Case study: Investigating nucleosome positioning at enhancers

In Chapter 6, I focused on the organization of promoters whereas that of other types of REs (*e.g.* putative enhancers) has not been discussed. To further illustrate the usefulness of VplotR, I now investigate nucleosome positioning at enhancers (defined in Jänes *et al.*, 2018). It is generally thought that enhancers are also characterized by well-positioning flanking nucleosomes (Andersson and Sandelin, 2019), but how this positioning compares to that observed at promoters remains unclear. I sought to investigate this point in more details using VplotR. I generated ATAC-seq fragment density plots at ubiquitous or tissue-specific sets of promoters or enhancers (Figure 8.3). I observed flanking nucleosome signals at ubiquitous and germline promoters and enhancers, generally absent at somatic-tissue-specific promoters or enhancers. I found that the relative enrichment of nucleosomal versus nucleosome-free fragments seems lower over ubiquitous and germline enhancers than over ubiquitous and germline promoters. This suggests that although -1 and +1 nucleosomes are flanking ubiquitous and germline enhancers, their occupancy is lesser there than at ubiquitous and germline promoters. However, the normalization method used here does not allow direct comparison of absolute density scores across heatmaps. In the future, the additional normalization currently being developed will allow such direct comparisons.

8.1.4 Public availability of VplotR

This case study illustrates how VplotR can be used to analyze chromatin organization at different types of genomic loci. VplotR is primarily designed for exploratory data analysis but also features quantification tools to test hypotheses. It is built in R and relies on the tidyverse environment to generate publication-ready figures in an organized workflow (Wickham *et al.*, 2019). VplotR is already available from Github (https://github.com/js2264/VplotR) and will be submitted for publication in the near future. All the analyses presented in this thesis have been performed using VplotR v0.4.0.

8.2 periodicDNA: R package to analyze k-mers periodicity

8.2.1 DNA sequence influences nucleosome positioning

Soon after solving the structure of nucleosomes, Kornberg raised a fundamental question: whether the positioning of nucleosomes *in vivo* in regard to a DNA locus was “specific” or “statistical” (Kornberg, 1981). Nucleosome “specific” positioning implies that the physicochemical properties of DNA sequences are enough to explain how nucleosomes are arranged along a DNA double-helix (*e.g.* described in Segal *et al.*, 2006). On the contrary, a “statistical” positioning postulates the presence of a “boundary” nucleosome (either a protein or a strong intrinsic positioning sequence, or both) which specifies one end of a nucleosomal array not determined by the physicochemical properties of DNA sequence (*e.g.* described in Mavrich *et al.*, 2008a). Later on, biochemists and computational biologists found out that periodic dinucleotide sequences were associated with positioned nucleosomes, suggesting that the “specific” model contributes – at least to a certain extent – to nucleosome positioning (see Jiang and Pugh, 2009; Struhl and Segal, 2013 for review).

8.2.2 periodicDNA can identify periodic oligonucleotides

To test whether specific periodic sequences were associated with nucleosome positioning in my project, I developed periodicDNA, a package aiming at characterizing periodicities of oligonucleotides (and particularly dinucleotides) (Serizay, 2020a). The package relies on Fourier Transform to identify periodic signals (Bracewell and Bracewell, 1986). It also makes use of the Biostrings package to handle DNA sequences and genome assemblies.

periodicDNA can be used to estimate the power spectral density (PSD) of a given dinucleotide (motif argument) at specific periods (period argument) in a set of sequences of interest (seqs argument), using a simple wrapper function:

```
periodicityScore <- periodicDNA::getPeriodicity(
    motif, seqs, period
)
```

The intermediate steps internally performed when calling this function are the following (Figure 8.4):

1. In each sequence of a set of n sequences (the seqs argument), all the pairs of the dinucleotide of interest (the motif argument, *e.g.* TT) are identified and their pairwise distances are measured.

2. The distribution of the all the resulting pairwise distances (also called “distogram”) is generated.

3. The following normalization steps are then performed:

   (a) The distogram is transformed into a frequency histogram and then normalized by the following steps:

   (b) The frequency histogram follows a marked overall decrease of frequencies with increased pairwise distances. Indeed, for a 200-bp long sequence containing 20 WW dinucleotides exactly distant from each other by 10 base pairs, there are 19 pairs with a pairwise distance of 10 but only 1

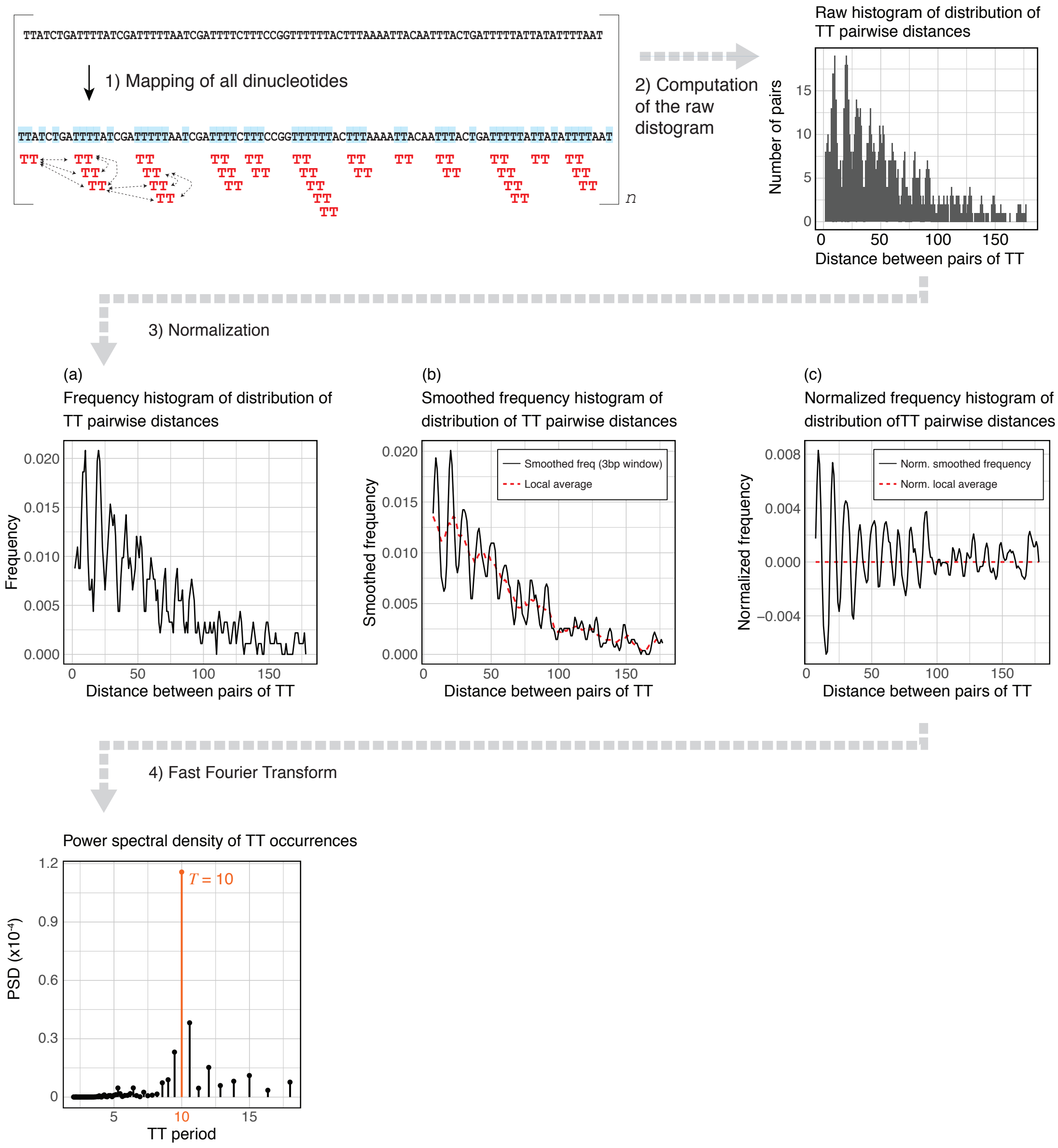

TTATCTGATTTTATCGATTTTTAATCGATTTTCTTTCCGGTTTTTTACTTTAAAATTACAATTTACTGATTTTTATTATATTTTAAT
1) Mapping of all dinucleotides
TTATCTGATTTTATCGATTTTTAATCGATTTTCTTTCCGGTTTTTTACTTTAAAATTACAATTTACTGATTTTTATTATATTTTAAT
TT
n
2) Computation of the raw distogram
Raw histogram of distribution of TT pairwise distances
Number of pairs
0
5
10
15
0
50
100
150
Distance between pairs of TT
3) Normalization
(a)
Frequency histogram of distribution of TT pairwise distances
Frequency
0.000
0.005
0.010
0.015
0.020
Distance between pairs of TT
(b)
Smoothed frequency histogram of distribution of TT pairwise distances
Smoothed frequency
Smoothed freq (3bp window)
Local average
Distance between pairs of TT
(c)
Normalized frequency histogram of distribution ofTT pairwise distances
Normalized frequency
0.008
0.004
0.000
−0.004
Norm. smoothed frequency
Norm. local average
Distance between pairs of TT
4) Fast Fourier Transform
Power spectral density of TT occurrences
PSD (x10⁻⁴)
0
0.3
0.6
0.9
1.2
T = 10
5
10
15
TT period

FIGURE 8.4 – Rationale of the periodicDNA package. Steps are described in the main text. The dotted double-arrows in the first step represent the distances measured by periodicDNA between some of the pairs of TT. For the single sequence shown here, there are 31 individual TT dinucleotides, resulting in $\binom{31}{2} = 465$ different pairs in total.

pair of dinucleotides with a pairwise distance of 190. To overcome this distance decay, the frequency histogram is smoothed using a moving average window of 10 and the resulting smoothed frequency histogram is substracted from the frequency histogram. This effectively transforms the decreasing frequency histogram into a dampened oscillating signal and improves the PSD estimation by Fourier Transform.

(c) The dampened oscillating signal is then scaled (*i.e.* mean-centered and normalized) and smoothed using a moving average window of 3. This last step effectively removes the effect of the latent 3-bp periodicity of most dinucleotides found in eukaryote genomes (Gutiérrez *et al.*, 1994).

4. A Fast Fourier Transform (FFT) is then used to estimate the power spectral density (PSD) of the normalized oscillating distribution at different periods (the period argument).

The PSD can be used in itself to identify which dinucleotide frequencies are enriched in the provided set of sequences. Its amplitude at a given frequency can also be used to compare dinucleotide frequencies across samples.

A manuscript presenting the periodicDNA package and its functionalities is shown in Appendix Chapter D.

8.2.3 Case study: Refining the model of sequence-based nucleosome positioning in *C. elegans*

I previously brought evidence suggesting that highly periodic dinucleotides are associated with -1 and +1 nucleosome positioning at ubiquitous and germline-specific promoters in *C. elegans* (see 6.2.1 on page 146). I revealed that at these sets of promoters, the underlying nucleosomal DNA sequence was characterized by a strong TT 10-bp periodicity, known to facilitate the bending of DNA into a conformation favorable to its wrapping around histones (Travers *et al.*, 2010). The use of my periodicDNA package was instrumental in identifying this feature. Here, I extend the use of periodicDNA to further characterize this periodicity in different genomic loci. I focus on ubiquitous and tissue-specific promoters and enhancers, splitting each element into core (-70 to +70 base pairs around the center of the regulatory element), flanking (-210 to -70 base pairs and +70 to +210 base pairs) and distal sequences (-350 to -210 base pairs and +210 to +350 base pairs) (Figure 8.5A). I then calculated the TT 10-bp periodicity score over core, flanking and distal sequences of ubiquitous or tissue-specific promoters and enhancers.

This showed that ubiquitous and germline-specific promoters have a high TT 10-bp periodicity in the flanking sequences which largely decreases in the immediate neighboring distal sequences (Figure 8.5B). Such TT periodicity is absent in other tissue-specific promoters, as expected from previous results (Figure 8.5B). These observations further support a sequence-specific model of nucleosome positioning at ubiquitous and germline promoters (Figure 8.5C). Interestingly, ubiquitous and germline enhancers show TT 10-bp periodicity in their flanking sequences as well as in their distal sequences (Figure 8.5B). The similarity of TT periodicity signal in the flanking and distal sequences of ubiquitous and germline enhancers could explain why nucleosomes flanking these enhancers may not be as strongly positioned as in the corresponding promoters (Figure 8.5C).

8.2.4 Public availability of periodicDNA

This case study illustrates how periodicDNA could be used to analyze periodicity of oligonucleotides in different types of genomic loci. periodicDNA is built in R and relies on the tidyverse environment to generate publication-ready figures in an organized workflow (Wickham *et al.*, 2019). periodicDNA will soon be submitted for publication and is already available from Github (https://github.com/js2264/periodicDNA). All the analyses presented in this thesis have been performed using periodicDNA v0.2.0.

8.3 JABrowser: a cloud-base genome browser for reproducible investigation

8.3.1 A virtual private server to publicly share data

High-throughput sequencing data is typically processed into a .bigWig file, which can then be loaded locally in genome browsers (*e.g.* using IGV, Robinson *et al.*, 2011). The .bigWig format represents a convenient way for an investigator to dynamically explore genomic tracks, but these files usually exceed hundreds of

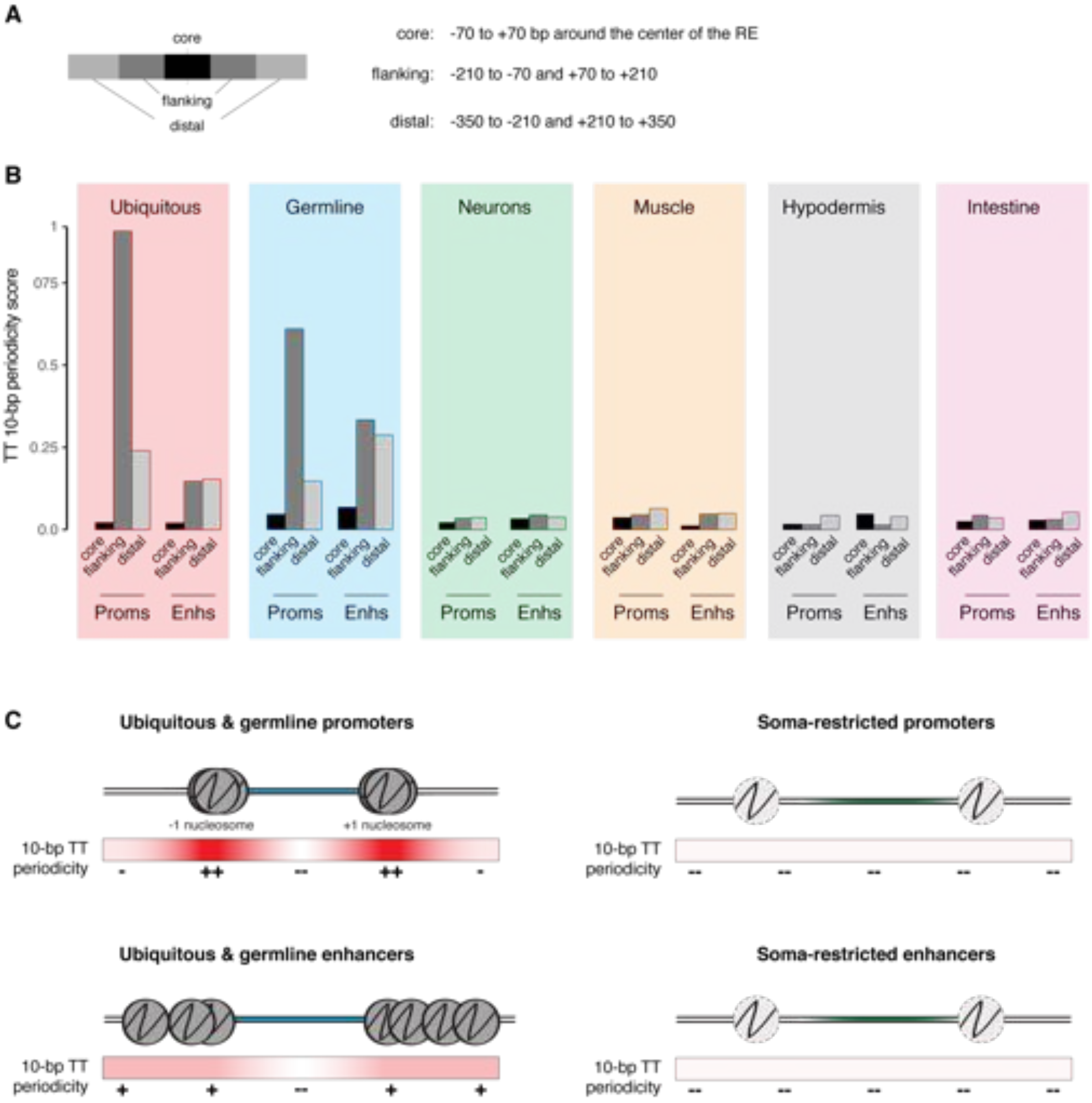


Figure 8.5 – TT periodicity in promoters and enhancers. **A-** Pictogram representing how regulatory elements were divided into core, flanking and distal regions. The core sequence is the 140-bp long sequence at the center of the regulatory element; the flanking sequences range from -210 to -70 and from +70 to +210; the distal sequences range from -350 to -210 and from +210 to +350 (with the center of the regulatory element being the reference). **B-** TT 10-bp periodicity scores obtained from periodicDNA. **C-** Model of sequence-driven nucleosome positioning at different sets of promoters or enhancers. Three different situations are observed: (1) a decrease of TT periodicity on both sides of the flanking nucleosomes favors their precise positioning, (2) a weaker widespread TT periodicity favors nucleosome positioning without local enrichment and (3) absence of TT periodicity does not favor nucleosome positioning. Note that these models do not illustrate the role of other factors such as chromatin remodelers.

megabytes and storing and sending genome-wide .bigWig tracks can still be a hurdle. Thus, sharing findings with others remains a challenging task. Online public genome browsers exist (*e.g.* UCSC genome browser, Tyner *et al.*, 2017) and can be used to remotely browse tracks without file transfers, but quickly become limiting when the files are locally required.

In an effort to make my results and findings available to the broad audience, I decided to publicly share all the datasets I had generated during my PhD on a server. To do so, I set up a web server with DigitalOcean. DigitalOcean is a hosting provider offering virtual private servers (VPS) named “droplets”. Droplets are highly customizable, with the possibility to choose among different distributions and to specify the virtual hardware dedicated to each droplet. I started with the entry-level plan, a Ubuntu 18.04 configuration with a single virtual CPU, 1 GB of memory and 25 GB of SSD storage. Virtual CPUs, memory and additional storage blocks can be dynamically purchased and added to an existing configuration, so my server should not incur any hardware limitation. I then set up a Nginx web server with a standard configuration. Large files such as .bigWig files can be hosted on this server and can be publicly accessed and downloaded, thus facilitating genomic data sharing.

8.3.2 JBrowse: an open-source genome browser

Besides offering a convenient way to share large genomic datasets, a web server can be configured to host a genome browser. A cloud-based genome browser is the ideal way to let everyone dynamically investigate processed datasets. I sought to install a genome browser on the VPS I had set up. I relied on JBrowse, a genome browser actively developed by the GMOD community (Buels *et al.*, 2016). JBrowse is an exceptionally fast genome browser with a fully dynamic HTML5 interface. Its installation is straightforward and once set up on a VPS, it can be served as a static web page. Convenient Perl scripts provided by GMOD can be used to easily integrate any type of data to a JBrowse instance. Importantly, JBrowse processes feature tracks into smaller .json (JavaScript Object Notation) chunks (Figure 8.6) and only the individual .json files required for rendering of each genomic location are transferred from the server to the client. This ensures that most of the computational work (*i.e.* rendering annotation tracks) is performed by

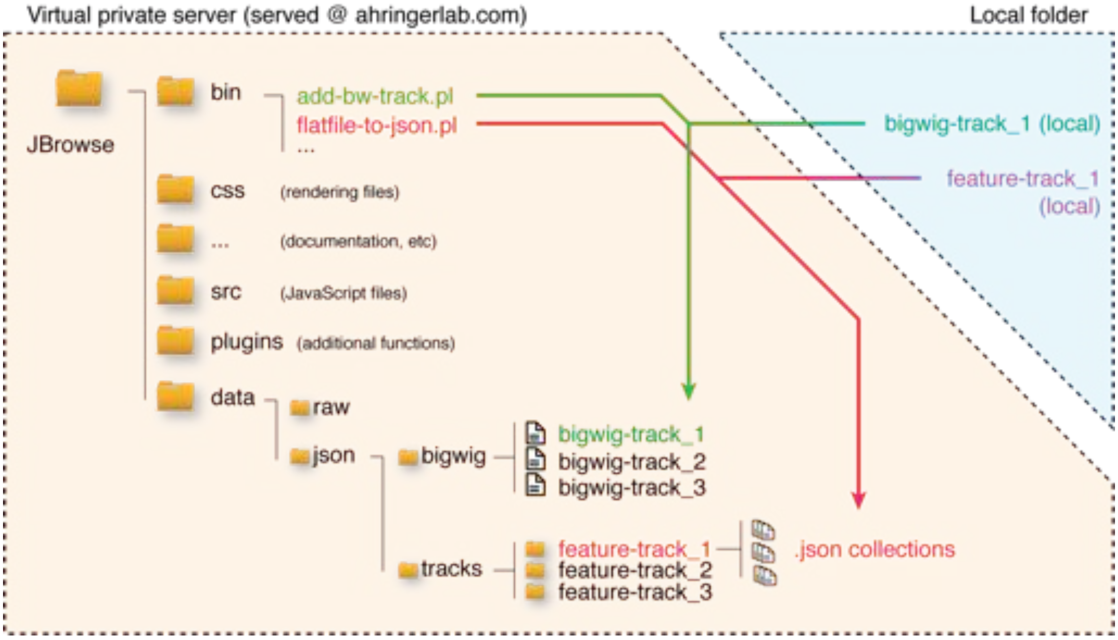
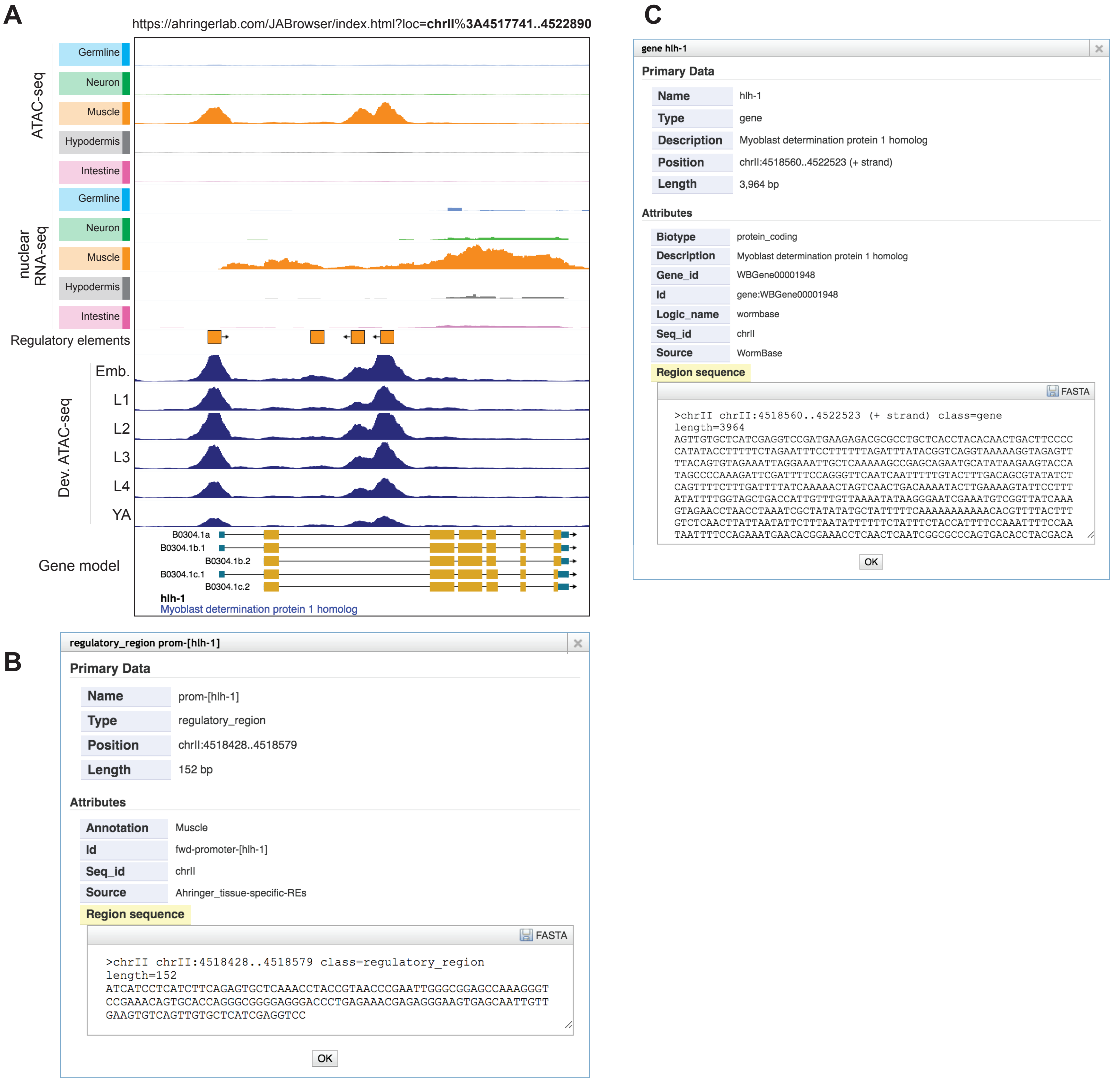


Figure 8.6 – JBrowse directory structure. The main JBrowse folder is located in the `public_html/` folder of the VPS to enable public access. It contains css, src and plugins folders required for the computational steps performed on the client-side. The bin folder contains perl scripts to add data to the genome browser, either from bigwig files or from "flat files" (*i.e.* feature files). Flat files are further processed into small .json collection files.

JavaScript on the client-side. This is a major advantage of JBrowse as it (i) allows for a seamless browsing experience by reducing the amount of data to download, and (ii) prevents server overload. This is particularly useful since the droplet I set up has fairly limited computational resources.

JBrowse is highly configurable through packages already developed and the support of JavaScript callback functions (Buels *et al.*, 2016). Notably, the browser acts as a static page and a unique URL is associated with each possible state of the browser (location, selected tracks, etc). Therefore, it is easy to share an observation with a collaborator by simply copying the URL. A "Share" button is also present in the browser and helps making it even more straightforward. Finally, a "Screenshot" button lets the investigators save high-definition, publication-ready figures of a genomic locus with selected tracks of interest. Thus, JBrowse is a dynamic genome browser integrating powerful extensions which can be used to facilitate genomic exploration and sharing of observations.



regulatory_region prom-[hlh-1]

Primary Data

Name	prom-[hlh-1]
Type	regulatory_region
Position	chrII:4518428..4518579
Length	152 bp

Attributes

Annotation	Muscle
Id	fwd-promoter-[hlh-1]
Seq_id	chrII
Source	Ahringer_tissue-specific-REs

Region sequence

```
>chrII chrII:4518428..4518579 class=regulatory_region
length=152
ATCATCCTCATCTTCAGAGTGCTCAAACCTACCGTAACCCGAATTGGGCGGAGCCAAAGGGT
CCGAAACAGTGCACCAGGGCGGGGAGGGACCCTGAGAAACGAGAGGGAAGTGAGCAATTGTT
GAAGTGTCAGTTGTGCTCATCGAGGTCC
```

gene hlh-1

Primary Data

Name	hlh-1
Type	gene
Description	Myoblast determination protein 1 homolog
Position	chrII:4518560..4522523 (+ strand)
Length	3,964 bp

Attributes

Biotype	protein_coding
Description	Myoblast determination protein 1 homolog
Gene_id	WBGene00001948
Id	gene:WBGene00001948
Logic_name	wormbase
Seq_id	chrII
Source	WormBase

Region sequence

```
>chrII chrII:4518560..4522523 (+ strand) class=gene
length=3964
AGTTGTGCTCATCGAGGTCCGATGAAGAGACGCGCCTGCTCACCTACACAACTGACTTCCCC
CATATACCTTTTTCTAGAATTTCCTTTTTTAGATTTATACGGTCAGGTAAAAAGGTAGAGTT
TTACAGTGTAGAAATTAGGAAATTGCTCAAAAAGCCGAGCAGAATGCATATAAGAAGTACCA
TAGCCCCAAAGATTCGATTTTCCAGGGTTCAATCAATTTTTGTACTTTGACAGCGTATATCT
CAGTTTTCTTTGATTTTATCAAAAACTAGTCAACTGACAAAATACTTGAAAAGTATTCCTTT
ATATTTTGGTAGCTGACCATTGTTTGTTAAAATATAAGGGAATCGAAATGTCGGTTATCAAA
GTAGAACCTAACCTAAATCGCTATATATGCTATTTTTCAAAAAAAAAAACACGTTTTACTTT
GTCTCAACTTATTAATATTCTTTAATATTTTTTCTATTTCTACCATTTTCCAAATTTTCCAA
TAATTTTCCAGAAATGAACACGGAAACCTCAACTCAATCGGCGCCCAGTGACACCTACGACA
```

Figure 8.7 – Overview of a JABrowser instance. **A-** *hlh-1* locus with several loaded tracks. Top 5 tracks: tissue-specific ATAC-seq in YA; 6-10th: tissue-specific nuclear RNA-seq in YA; 11th: annotation of regulatory elements; 12-17th: ATAC-seq across worm development;18th: gene model. **B-** Screenshot of the information panel displayed upon clicking on a regulatory element. **C-** Screenshot of the information panel displayed upon clicking on a gene.

8.3.3 JABrowser: a public JBrowse instance hosted by the Ahringer lab

I configured JABrowser, a JBrowse instance hosting different types of datasets generated in the Ahringer lab (Figure 8.7A). Gene and regulatory annotations have been imported as feature tracks. For each entry, customized information can be displayed by clicking on the feature of interest (Figure 8.7B, C). Chromatin accessibility, gene expression signals and other tracks (*e.g.* ChIP-seq experiments) are stored in bigWig files and are displayed as linear tracks (Figure 8.7A). A convenient track selector is available to filter and select the datasets to visualize. In total, sixty-five tracks are currently hosted in JABrowser:

- Five different feature annotation tracks (for genes, regulatory elements, etc);
- Eleven developmental or aging and five tissue-specific chromatin accessibility tracks;
- Twelve developmental and ten tissue-specific gene expression tracks;
- Twenty-two profiles of histone marks generated by ChIP-seq, covering four different histone modifications across development.

JABrowser is accessible at https://ahringerlab.com/JABrowser.

8.4 RegAtlas: a web app to investigate tissue-specific gene regulation in worm

8.4.1 A responsive infrastructure based on Shiny

Relying on DigitalOcean as a VPS provider has another important benefit: DigitalOcean grants super-user authorizations to droplet managers. This means that once set up, a droplet can really be tailored to the developer's needs and software can be installed and fully configured. I decided to take full advantage of this and I set up an instance of R 3.5.1 (R Core Team, 2019) on my VPS. This allowed me to host R data files and carry out analyses directly in the cloud. Even though this droplet does not provide hefty computation power like other cloud computing systems, it can still remotely run R sessions with reasonable computation requirements.

I sought to develop a web application which could be used to dynamically investigate and download my data. I decided to build such application using Shiny, a versatile R package particularly powerful when used in combination with a VPS (Chang *et al.*, 2019). Shiny aims to combine the computational power of R with the interactivity of the modern web by creating a dynamic app, built on HTML and supporting JavaScript and CSS libraries. Relying on the R language, a developer can write a full-featured web app presented in an intuitive interface. On the client-side, a user without coding experience can communicate with the Shiny app's underlying structure to access, process and visualize data using R powerful packages, all of this directly in a traditional web browser. Shiny apps have been successfully used in all scientific fields and have been particularly effective in genomics projects to give an overview of large datasets (*e.g.* SPACEGERM (Diag *et al.*, 2018), VisCello (Packer *et al.*, 2019) or by Cusanovich *et al.* (2018a)). Using Shiny, I created RegAtlas, a *C. elegans* regulatory atlas accessible at http://ahringerlab.com/RegAtlas.

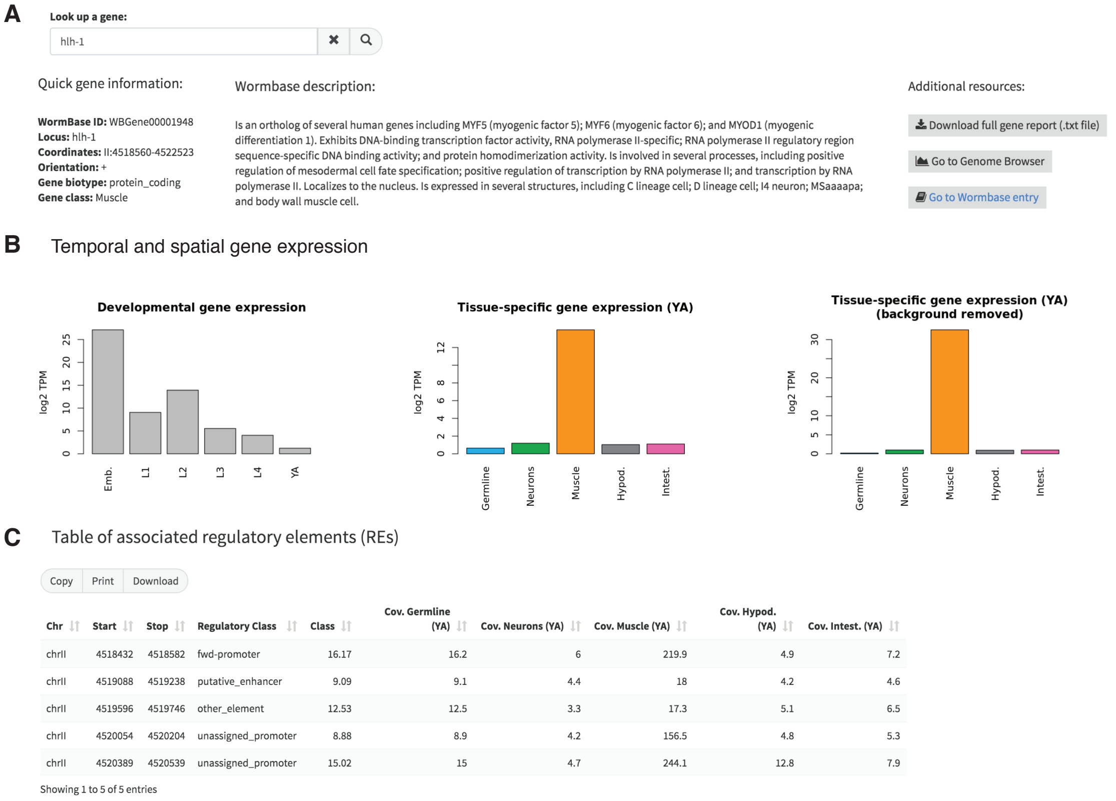
8.4.2 Quick access to gene information

To provide a fully-fledged intuitive user interface, I divided RegAtlas in four segments. The first page of the app primarily focuses on single gene entries, as many users would likely want to query individual genes that they study. This tab can present the information related to specific gene (*e.g. hlh-1*) queried using either its unique WormBase ID or its locus name (Figure 8.8). The first row highlights basic information about the gene of interest (name, genomic location, tissue-specific pattern of expression), a short description based on the Textpresso engine (extracted from WormBase, Müller *et al.*, 2004) and several "quick access" hyperlinks, for instance to download an extended .txt report or to the associated WormBase gene's page (Figure 8.8A). The second row displays gene expression values throughout development (from embryo to young adults, in mixed tissues) and across tissues (in young adult) (Figure 8.8B). Finally, the third row of the page contains a table of annotated regulatory elements associated to the gene on interest (Figure 8.8C).

Thus, this page quickly provides general information for a given gene and can be used by the community as a platform to investigate tissue-specific gene regulation.

8.4.3 Querying lists of genes

The second page of RegAtlas focuses on lists of genes and follows a more analytical approach. A list of genes can be entered by the user by copy-pasting, or can be chosen among the pre-computed gene lists (*e.g.* muscle-specific genes, genes with at least one hypodermis promoter, etc) (Figure 8.9A). Server-side computation in R is then initiated by clicking on the "Perform analysis" button. Once the analysis is finished (typically < 5 seconds), a .gff3 annotation can be downloaded. This .gff3 file contains useful information such as the level of tissue-specific expression of



Quick gene information:

WormBase ID: WBGene00001948
Locus: hlh-1
Coordinates: II:4518560-4522523
Orientation: +
Gene biotype: protein_coding
Gene class: Muscle

Wormbase description:

Is an ortholog of several human genes including MYF5 (myogenic factor 5); MYF6 (myogenic factor 6); and MYOD1 (myogenic differentiation 1). Exhibits DNA-binding transcription factor activity, RNA polymerase II-specific; RNA polymerase II regulatory region sequence-specific DNA binding activity; and protein homodimerization activity. Is involved in several processes, including positive regulation of mesodermal cell fate specification; positive regulation of transcription by RNA polymerase II; and transcription by RNA polymerase II. Localizes to the nucleus. Is expressed in several structures, including C lineage cell; D lineage cell; I4 neuron; MSaaaapa; and body wall muscle cell.

Additional resources:

Download full gene report (.txt file)

Go to Genome Browser

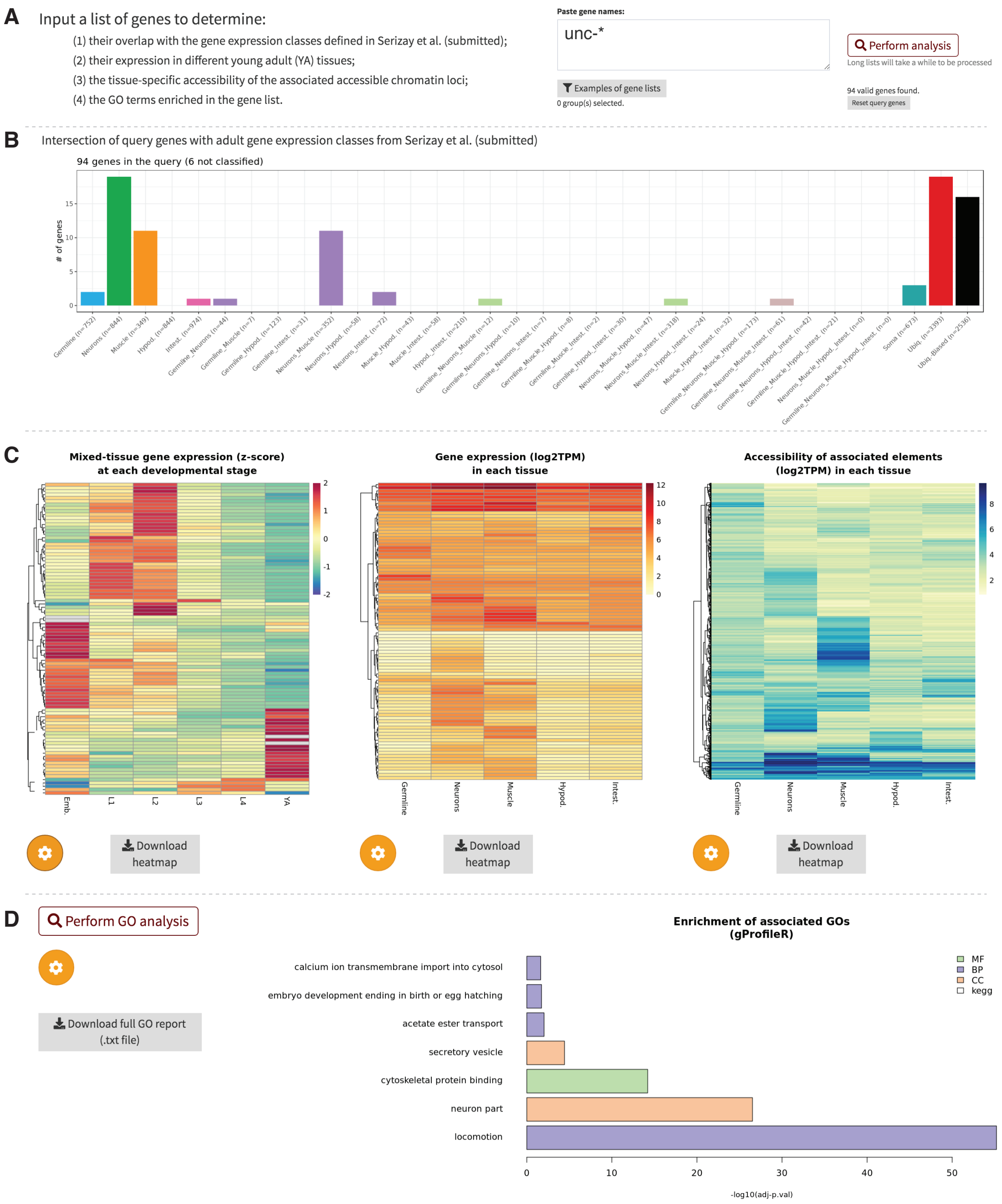
Go to Wormbase entry

Table of associated regulatory elements (REs)

Chr	Start	Stop	Regulatory Class	Class	Cov. Germline (YA)	Cov. Neurons (YA)	Cov. Muscle (YA)	Cov. Hypod. (YA)	Cov. Intest. (YA)
chrII	4518432	4518582	fwd-promoter	16.17	16.2	6	219.9	4.9	7.2
chrII	4519088	4519238	putative_enhancer	9.09	9.1	4.4	18	4.2	4.6
chrII	4519596	4519746	other_element	12.53	12.5	3.3	17.3	5.1	6.5
chrII	4520054	4520204	unassigned_promoter	8.88	8.9	4.2	156.5	4.8	5.3
chrII	4520389	4520539	unassigned_promoter	15.02	15	4.7	244.1	12.8	7.9

Showing 1 to 5 of 5 entries

Figure 8.8 – RegAtlas: overview of the single gene query page. **A-** Genes can be queried by entering their WormBase ID or their locus ID. Basic information are displayed, mostly retrieved from WormBase. **B-** Gene expression values across development (left) and in tissues of young adults (middle and right) are displayed. Note that the right panel displays tissue-specific gene expression values after correction of the background noise, as described in the Information panel of the web app. **C-** Table of associated regulatory elements and their tissue-specific expression in young adult.



A Input a list of genes to determine:

(1) their overlap with the gene expression classes defined in Serizay et al. (submitted);
(2) their expression in different young adult (YA) tissues;
(3) the tissue-specific accessibility of the associated accessible chromatin loci;
(4) the GO terms enriched in the gene list.

Paste gene names: unc-*

0 group(s) selected.

Long lists will take a while to be processed

94 valid genes found.

B Intersection of query genes with adult gene expression classes from Serizay et al. (submitted)

94 genes in the query (6 not classified)

Perform analysis
Examples of gene lists
Reset query genes
# of genes
0
5
10
15
Germline (n=752)
Neurons (n=844)
Muscle (n=349)
Hypod. (n=844)
Intest. (n=974)
Germline_Neurons (n=44)
Germline_Muscle (n=7)
Germline_Hypod. (n=123)
Germline_Intest. (n=31)
Neurons_Muscle (n=352)
Neurons_Hypod. (n=58)
Neurons_Intest. (n=72)
Muscle_Hypod. (n=43)
Muscle_Intest. (n=58)
Hypod._Intest. (n=210)
Germline_Neurons_Muscle (n=12)
Germline_Neurons_Hypod. (n=10)
Germline_Neurons_Intest. (n=7)
Germline_Muscle_Hypod. (n=8)
Germline_Muscle_Intest. (n=2)
Germline_Hypod._Intest. (n=30)
Neurons_Muscle_Hypod. (n=47)
Neurons_Muscle_Intest. (n=318)
Neurons_Hypod._Intest. (n=24)
Muscle_Hypod._Intest. (n=32)
Germline_Neurons_Muscle_Hypod. (n=173)
Germline_Neurons_Muscle_Intest. (n=61)
Germline_Neurons_Hypod._Intest. (n=42)
Germline_Muscle_Hypod._Intest. (n=21)
Neurons_Muscle_Hypod._Intest. (n=0)
Germline_Neurons_Muscle_Hypod._Intest. (n=0)
Soma (n=673)
Ubiq. (n=3393)
Ubiq.-Biased (n=2536)
C
Mixed-tissue gene expression (z-score) at each developmental stage
2
1
0
-1
-2
Emb.
L1
L2
L3
L4
YA
Gene expression (log2TPM) in each tissue
12
10
8
6
4
2
0
Germline
Neurons
Muscle
Hypod.
Intest.
Accessibility of associated elements (log2TPM) in each tissue
8
6
4
2
Germline
Neurons
Muscle
Hypod.
Intest.
Download heatmap
Download heatmap
Download heatmap
D
Perform GO analysis
Download full GO report (.txt file)
Enrichment of associated GOs (gProfileR)
calcium ion transmembrane import into cytosol
embryo development ending in birth or egg hatching
acetate ester transport
secretory vesicle
cytoskeletal protein binding
neuron part
locomotion
0
10
20
30
40
50
-log10(adj-p.val)
MF
BP
CC
kegg

Figure 8.9 – RegAtlas: overview of the multiple genes query page. **A-** List of genes can be queried by typing or copying their WormBase ID or their locus ID. Lists of genes are available as example. Regular expressions can be used to match several names (here all the *unc* genes are queried) **B-** Barplot representing the intersection of the query with the tissue-specific and ubiquitous sets of genes. **C-** Heatmaps of temporal and spatial expression of the query and their associated regulatory elements. **D-** Results of Gene Ontology terms enrichment analysis performed on the query using gProfiler (Reimand *et al.*, 2007).

the queried genes, or the tissue-specific accessibility of their associated regulatory elements. Furthermore, the intersection of the query list with tissue-specific and ubiquitous gene sets computed during my PhD is shown as a barplot (Figure 8.9B). Heatmaps of gene expression throughout development and across tissues in young adult worms are also displayed and many plotting parameters can be customized using dropdown menus (Figure 8.9C). Tissue-specific chromatin accessibility scores of the regulatory elements associated to the query are also shown as a third heatmap. Finally, GO term enrichment analyses can be performed on the list query. The gProfiler R package (Reimand *et al.*, 2007) is used to identify GO terms enriched in genes from the query (Figure 8.9D).

This page provides an analytical framework to examine lists of genes. Importantly, all the results can be individually exported in files readable by standard computing tools, therefore facilitating the use of this app as a tool for reproducible investigation.

8.4.4 Integration of JABrowser

Shiny apps essentially behave like dynamic web pages, modified and reloaded upon client input. Moreover, because the genome browser I developed is based on JBrowse and behaves as a web page with a unique URL (see Section 8.3), it is easily embeddable into other web pages using standard HTML code. I leveraged these features of Shiny and JBrowse to integrate my genome browser into RegAtlas. The integrated genome browser is accessible in the third page of the app, and presents the same features as in its standalone version. Because the URL of the embedded browser is obtained from the Shiny code, it can be programmatically updated upon client's input. This ensures that when a user queries a gene in the first tab, the browser URL is updated to reflect this change. This allowed me to integrate a hyperlink in the first tab of the app to redirect the user to the browser already focused on the queried gene, so that navigating across the app is a seamless experience.

8.4.5 Quick and anonymous access to entire datasets

Finally, all the processed genome-wide tissue-specific ATAC-seq or RNA-seq datasets can be anonymously downloaded with a single click, as (1) browsable tracks in

.bigWig format, (2) annotation features in .gff3 format or (3) as text tables in .txt format. The latter tables have been formatted so that their import in R is facilitated. Moreover, it is possible to dynamically explore and filter these tables within the web application. Thus, this page favors both small-scale data investigation and large-scale data sharing.

Conclusion

At the beginning of my PhD, I focused on investigating the dynamics of chromatin accessibility during *C. elegans* life. By characterizing sets of promoters whose activity is temporally coordinated, I shed light on functional networks of transcription factors during development and aging. This work has been integrated in a larger project annotating, describing and characterizing promoters and enhancers across the *C. elegans* genome and throughout its life.

At the same time, I developed and optimized a method to sort nuclei from individual tissues in *C. elegans*. This method allowed me to investigate gene regulation in individual tissues of *C. elegans* at the young adult stage. I profiled chromatin accessibility and gene expression across the five main tissues of the nematode. By classifying the chromatin accessible site in tissue-specific, tissue-restricted and ubiquitous sets, I showed that most of the regulatory elements in *C. elegans* genome are active in a single tissue or in a subset of tissues, rather than ubiquitously. I also uncovered large differences in gene structures: ubiquitously expressed genes and germline-specific genes have a particularly simple structure while genes where expression is restricted to somatic tissues have a more complex organization, with more regulatory elements associated with each gene. Finally, I provide evidence that the spatial folding of the chromatin exhibits tissue-specific characteristics. In the germline, chromatin folds into small communities while in somatic tissues, it has a more complex 3D network that relies on soma regulatory elements. The function of these soma regulatory elements, and whether mechanisms of liquid-liquid phase-separation are required for the segregation of the chromatin into large communities, are possible paths to explore in the future.

I then focused on the characteristics of ubiquitous and tissue-specific promoters. I showed that ubiquitous promoters are characterized by well-positioned -1 and +1 nucleosomes associated with a 10-bp WW periodic signal, as previously described for broadly active promoters. Strikingly, promoters active in adult germline also share this organization, while those whose activity is restrained to somatic tissues have a much less defined promoter structure. Notably, positioning of the +1 nucleosome at ubiquitous and germline-specific promoters is well-aligned with that of the TSS. For the first time, these results identify a different organization between promoters active in germline (*i.e.* germline-specific and ubiquitous promoters) and those restricted to somatic tissues. These findings could suggest that different mechanisms of positioning of Pre-Initiation Complex exist in the two different groups of promoters. In the future, characterizing molecular arrangement of the RNA Polymerase II and other general transcription factors at these loci may shed light on the contribution of promoter sequences and positioning of the PIC. My results also suggest that 10-bp WW periodicity could be an ancient conserved signal contributing to +1 nucleosome positioning at ubiquitously active promoters of non-mammalian eukaryotes, whereas nucleosome positioning in mammals may rely on other mechanisms.

Overall, combinations of sequence features and DNA binding motifs are strikingly different at ubiquitous, germline and somatic-tissue-specific promoters in *C. elegans*. In the future, these clear differences could be leveraged to annotate tissue-specific regulatory element activity based on their sequence. Relying on powerful sequence classification methods built on Long Short Term Memory neural networks, characteristics of tissue-specific and ubiquitous promoters in *C. elegans* could be used to identify and classify regulatory elements in other nematode species, for example.

Toward the end of my PhD, I initiated a new project aiming at characterizing the genomic regulation and gene expression changes that drive cell specification and differentiation during development. I started optimizing single-nuclei approaches to profile chromatin accessibility and gene expression in individual nuclei using the 10X Genomics platform. I performed preliminary experiments in very early embryos (1-cell to 32-cell stage) and later embryos (up to 100-cell stage). This showed promising results suggesting that these methods could be used to characterize gene regulation during early embryonic development in individual cells. In the future, organogenesis during embryogenesis and in post-embryonic development could also be investigated by using single-cell approaches in combination with the nuclear sorting method I developed.

Finally, I strove to make all my findings and my methodology easily and anonymously available to the broad community. I created two R packages to support my research, which can now be used to answer other biological questions. I also designed RegAtlas, a web interface to dynamically browse and easily download the key datasets I generated during my PhD. In the future, these efforts should facilitate the dissemination of this data and promote the investigation of the mechanisms of tissue-specific gene regulation in *C. elegans* as well as in other organisms.

Bibliography

Albert, I., Mavrich, T.N., Tomsho, L.P., Qi, J., Zanton, S.J., Schuster, S.C. and Pugh, B.F. Translational and rotational settings of H2A.Z nucleosomes across the saccharomyces cerevisiae genome. *Nature*, *446*(7135):572, 2007.

Ali, T., Renkawitz, R. and Bartkuhn, M. Insulators and domains of gene expression. *Curr. Opin. Genet. Dev.*, *37*:17, 2016.

Allen, M.A., Hillier, L.W., Waterston, R.H. and Blumenthal, T. A global analysis of C. elegans trans-splicing. *Genome research*, *21*(2):255, 2011.

Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M.L., Huber, W., Morgan, M., Gottardo, R. and Hicks, S.C. Orchestrating single-cell analysis with bioconductor. *Nat. Methods*, *17*(2):137, 2020.

Andersson, R. and Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, 2019.

Andersson, R., Sandelin, A. and Danko, C.G. A unified architecture of transcriptional regulatory elements. *Trends Genet.*, *31*(8):426, 2015.

Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493):455, 2014.

Andrioli, L.P.M., Vasisht, V., Theodosopoulou, E., Oberstein, A. and Small, S. Anterior repression of a drosophila stripe enhancer requires three position-specific mechanisms. *Development*, *129*(21):4931, 2002.

Araya, C.L., Kawli, T., Kundaje, A., Jiang, L., Wu, B., Vafeados, D., Terrell, R., Weissdepp, P., Gevirtzman, L., Mace, D., Niu, W., Boyle, A.P., Xie, D., Ma, L., Murray, J.I., Reinke, V., Waterston, R.H. and Snyder, M. Regulatory analysis of the c. elegans genome with spatiotemporal resolution. *Nature*, *512*(7515):400, 2014.

Bagijn, M.P., Goldstein, L.D., Sapetschnig, A., Weick, E.M., Bouasker, S., Lehrbach, N.J., Simard, M.J. and Miska, E.A. Function, Targets, and Evolution of Caenorhabditis elegans piRNAs. *Science*, *337*(6094):574, 2012.

Bahr, C., von Paleske, L., Uslu, V.V., Remeseiro, S., Takayama, N., Ng, S.W., Murison, A., Langenfeld, K., Petretich, M., Scognamiglio, R., Zeisberger, P., Benk, A.S., Amit, I., Zandstra, P.W., Lupien, M., Dick, J.E., Trumpp, A. and Spitz, F. A myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature*, *553*(7689):515, 2018.

Bannister, A.J. and Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.*, *21*(3):381, 2011.

Beagan, J.A. and Phillips-Cremins, J.E. On the existence and functionality of topologically associating domains. *Nat. Genet.*, *52*(1):8, 2020.

Benchetrit, H., Jaber, M., Zayat, V., Sebban, S., Pushett, A., Makedonski, K., Zakheim, Z., Radwan, A., Maoz, N., Lasry, R., Renous, N., Inbar, M., Ram, O., Kaplan, T. and Buganim, Y. Direct induction of the three pre-implantation blastocyst cell types from fibroblasts. *Cell Stem Cell*, *24*(6):983, 2019.

Bessa, J., Gebelein, B., Pichaud, F., Casares, F. and Mann, R.S. Combinatorial control of drosophila eye development by eyeless, homothorax, and teashirt. *Genes Dev.*, *16*(18):2415, 2002.

Blazie, S.M., Babb, C., Wilky, H., Rawls, A., Park, J.G. and Mangone, M. Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in caenorhabditis elegans intestine and muscles. *BMC Biol.*, *13*:4, 2015.

Bonn, S., Zinzen, R.P., Perez-Gonzalez, A., Riddell, A., Gavin, A.C. and M Furlong, E.E. Cell type–specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. *Nat. Protoc.*, *7*, 2012.

Bowman, G.D. and Poirier, M.G. Post-translational modifications of histones that influence nucleosome dynamics. *Chem. Rev.*, *115*(6):2274, 2015.

Bozek, M., Cortini, R., Storti, A.E., Unnerstall, U., Gaul, U. and Gompel, N. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the drosophila blastoderm. *Genome Res.*, *29*(5):771, 2019.

Bracewell, R.N. and Bracewell, R.N. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M.M. and Lorincz, M.C. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nat. Commun.*, *6*:6033, 2015.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L. and Holmes, I.H. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, *17*:66, 2016.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 2013.

Bulut-Karslioglu, A., Macrae, T.A., Oses-Prieto, J.A., Covarrubias, S., Percharde, M., Ku, G., Diaz, A., McManus, M.T., Burlingame, A.L. and Ramalho-Santos, M. The transcriptionally permissive chromatin state of embryonic stem cells is acutely tuned to translational output. *Cell Stem Cell*, *22*(3):369, 2018.

Buratowski, S., Hahn, S., Guarente, L. and Sharp, P.A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, *56*(4):549, 1989.

C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, *282*(5396):2012, 1998.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., Adey, A., Waterston, R.H., Trapnell, C. and Shendure, J. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352):661, 2017.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C. and Shendure, J. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745):496, 2019.

Carelli, F.N., Sharma, G. and Ahringer, J. Broad chromatin domains : An important facet of genome regulation. *Bioessays*, *1700124*:1, 2017.

Carlberg, C. and Campbell, M.J. Vitamin D receptor signaling mechanisms: integrated actions of a well-defined transcription factor. *Steroids*, *78*(2):127, 2013.

Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, *38*(6):626, 2006.

Catarino, R.R. and Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.*, *32*(3-4):202, 2018.

Chang, W., Cheng, J., Allaire, J., Xie, Y. and McPherson, J. *Shiny: Web Application Framework for R*, 2019. R package version 1.3.2.

Chatterjee, S. and Ahituv, N. Gene regulatory elements, major drivers of human disease. *Annu. Rev. Genomics Hum. Genet.*, *18*:45, 2017.

Chen, L., Krause, M., Sepanski, M. and Fire, A. The caenorhabditis elegans MYOD homologue HLH-1 is essential for proper muscle function and complete morphogenesis. *Development*, *120*(6):1631, 1994.

Chen, R.A., Down, T.A., Stempor, P., Chen, Q.B., Egelhofer, T.A., Hillier, L.W., Jeffers, T.E. and Ahringer, J. The landscape of RNA polymerase II transcription initiation in C . elegans reveals promoter and enhancer architectures. *Genome Res.*, pages 1339–1347, 2013.

Chen, R.A.J., Stempor, P., Down, T.A., Zeiser, E., Feuer, S.K. and Ahringer, J. Extreme HOT regions are CpG-dense promoters in c. elegans and humans. *Genome Res.*, *24*(7):1138, 2014.

Chen, T. and Dent, S.Y.R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.*, *15*(2):93, 2014.

Chen, Z., Eastburn, D.J. and Han, M. The caenorhabditis elegans nuclear receptor gene nhr-25 regulates epidermal cell development. *Mol. Cell. Biol.*, *24*(17):7345, 2004.

Clapier, C.R. and Cairns, B.R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.*, *78*:273, 2009.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., Kathiria, A., Cho, S.W., Mumbach, M.R., Carter, A.C., Kasowski, M., Orloff, L.A., Risca, V.I., Kundaje, A., Khavari, P.A., Montine, T.J., Greenleaf, W.J. and Chang, H.Y. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*, *14*(10):959, 2017.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A. and Lis, J.T. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, *46*(12):1311, 2014.

Corsi, A.K., Wightman, B. and Chalfie, M. A transparent window into biology: A primer on caenorhabditis elegans. *Genetics*, *200*(2):387, 2015.

Cox, G.N. and Hirsh, D. Stage-specific patterns of collagen gene expression during development of Caenorhabditis elegans. *Molecular and cellular biology*, *5*(2):363, 1985.

Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J. and Meyer, B.J. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, *523*(7559):240, 2015.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, *Complex Systems*:1695, 2006.

Cusanovich, D.A., Reddington, J.P., Garfield, D.A., Daza, R.M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H.A., Christiansen, L., Qiu, X., Steemers, F.J., Trapnell, C., Shendure, J. and Furlong, E.E.M. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, *555*(7697):538, 2018a.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., Lee, C., Regalado, S.G., Read, D.F., Steemers, F.J., Disteche, C.M., Trapnell, C. and Shendure, J. A Single-Cell atlas of in vivo mammalian chromatin accessibility. *Cell*, *174*(5):1309, 2018b.

Cutter, A.R. and Hayes, J.J. A brief review of nucleosome structure. *FEBS Lett.*, *589*(20 Pt A):2914, 2015.

Danzer, J.R. and Wallrath, L.L. Mechanisms of HP1-mediated gene silencing in drosophila. *Development*, *131*(15):3571, 2004.

Daugherty, A.C., Yeo, R., Buenrostro, J.D., Greenleaf, W.J., Kundaje, A. and Brunet, A. Chromatin accessibility dynamics reveal novel functional enhancers in c. elegans. *Genome Res.*, page 088732, 2017.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C. and Huang, T.H.M. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.*, *24*(4):167, 2008.

Diag, A., Schilling, M., Klironomos, F., Ayoub, S. and Rajewsky, N. Spatiotemporal m(i)RNA architecture and 3' UTR regulation in the c. elegans germline. *Dev. Cell*, *47*(6):785, 2018.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398):376, 2012.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K. and Young, R.A. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 2014.

Dreos, R., Ambrosini, G. and Bucher, P. Influence of rotational nucleosome positioning on transcription start site selection in animal promoters. *PLoS Comput. Biol.*, *12*(10):e1005144, 2016.

Drew, H.R. and Travers, A.A. DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.*, *186*(4):773, 1985.

Dukler, N., Gulko, B., Huang, Y.F. and Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nat. Genet.*, *49*(1):2, 2016.

Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414):57, 2012.

Dupuy, D. *et al.* Genome-scale analysis of in vivo spatiotemporal promoter activity in caenorhabditis elegans. *Nat. Biotechnol.*, *25*(6):663, 2007.

Ernst, J. and Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, *28*(8):817, 2010.

Evans, K.J., Huang, N., Stempor, P., Chesney, M.A., Down, T.A. and Ahringer, J. Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *Proc. Natl. Acad. Sci. U. S. A.*, *113*(45):E7020, 2016.

Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A. and Schier, A.F. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392), 2018.

Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.*, *7*(9):1728, 2012.

Feric, M., Vaidya, N., Harmon, T.S., Kriwacki, R.W., Pappu, R.V., Brangwynne, C.P., Mitrea, D.M., Zhu, L. and Richardson, T.M. Coexisting liquid phases underlie nucleolar subcompartments. *Cell*, *165*:1686, 2016.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I.K., Sharon, E., Lubling, Y., Widom, J. and Segal, E. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, *4*(11):e1000216, 2008.

Fischle, W., Tseng, B.S., Dormann, H.L., Ueberheide, B.M., Garcia, B.A., Shabanowitz, J., Hunt, D.F., Funabiki, H. and Allis, C.D. Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature*, *438*(7071):1116, 2005.

Foley, J.W. and Sidow, A. Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics*, *14*(1):1, 2013.

Folick, A., Oakley, H.D., Yu, Y., Armstrong, E.H., Kumari, M., Sanor, L., Moore, D.D., Ortlund, E.A., Zechner, R. and Wang, M.C. Aging. lysosomal signaling molecules regulate longevity in caenorhabditis elegans. *Science*, *347*(6217):83, 2015.

Forrest, A.R.R. *et al.* A promoter-level mammalian expression atlas. *Nature*, *507*(7493):462, 2014.

Fox, R.M., Von Stetina, S.E., Barlow, S.J., Shaffer, C., Olszewski, K.L., Moore, J.H., Dupuy, D., Vidal, M. and Miller, 3rd, D.M. A gene expression fingerprint of c. elegans embryonic motor neurons. *BMC Genomics*, *6*:42, 2005.

Fox, R.M., Watson, J.D., Stetina, S.E., Mcdermott, J., Brodigan, T.M., Fukushige, T., Krause, M. and Miller, Iii, D.M. The embryonic muscle transcriptome of caenorhabditis elegans. *Genome Biol.*, *8*(9), 2007.

Frøkjaer-Jensen, C., Davis, M.W., Hopkins, C.E., Newman, B.J., Thummel, J.M., Olesen, S.P., Grunnet, M. and Jorgensen, E.M. Single-copy insertion of transgenes in caenorhabditis elegans. *Nat. Genet.*, *40*(11):1375, 2008.

Fuda, N.J., Ardehali, M.B. and Lis, J.T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, *461*(7261):186, 2009.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L. Formation of chromosomal domains by loop extrusion. *Cell Rep.*, *15*(9):2038, 2016.

Fukushige, T., Hawkins, M.G. and McGhee, J.D. The GATA-factor elt-2 is essential for formation of the caenorhabditis elegans intestine. *Dev. Biol.*, *198*(2):286, 1998.

Furlong, E.E.M. and Levine, M. Developmental enhancers and chromosome topology. *Science*, *361*(6409):1341, 2018.

Gao, L., Wu, K., Liu, Z., Yao, X., Yuan, S., Tao, W., Yi, L., Yu, G., Hou, Z., Fan, D., Tian, Y., Liu, J., Chen, Z.J. and Liu, J. Chromatin accessibility landscape in human early embryos and its association with evolution. *Cell*, *173*(1):248, 2018.

Gaydos, L.J., Rechtsteiner, A., Egelhofer, T.A., Carroll, C.R. and Strome, S. Antagonism between MES-4 and polycomb repressive complex 2 promotes appropriate gene expression in c. elegans germ cells. *Cell Rep.*, *2*(5):1169, 2012.

Gelino, S., Chang, J.T., Kumsta, C., She, X., Davis, A., Nguyen, C., Panowski, S. and Hansen, M. Intestinal autophagy improves healthspan and longevity in c. elegans during dietary restriction. *PLoS Genet.*, *12*(7):e1006135, 2016.

Gerstein, M.B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, *330*(6012):1775, 2010.

Gerstein, M.B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature*, *512*(7515):445, 2014.

Gilbert, S.F. *The Developmental Mechanics of Cell Specification.* Sinauer Associates, 2000.

Gilleard, J.S. and McGhee, J.D. Activation of hypodermal differentiation in the caenorhabditis elegans embryo by GATA transcription factors ELT-1 and ELT-3. *Mol. Cell. Biol.*, *21*(7):2533, 2001.

Gilleard, J.S., Shafi, Y., Barry, J.D. and McGhee, J.D. ELT-3: A caenorhabditis elegans GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev. Biol.*, *208*(2):265, 1999.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. FAIRE (Formaldehyde-Assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, *17*(6):877, 2007.

Girvan, M. and Newman, M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12):7821, 2002.

Gissendanner, C.R. and Sluder, A.E. nhr-25, the caenorhabditis elegans ortholog of ftz-f1, is required for epidermal and somatic gonad development. *Dev. Biol.*, *221*(1):259, 2000.

Goudeau, J., Bellemin, S., Toselli-Mollereau, E., Shamalnasab, M., Chen, Y. and Aguilaniu, H. Fatty acid desaturation links germ cell loss to longevity through NHR-80/HNF4 in c. elegans. *PLoS Biol.*, *9*(3):e1000599, 2011.

Gracida, X., Norris, A.D. and Calarco, J.A. *Regulation of Tissue-Specific Alternative Splicing: C. elegans as a Model System*, pages 229–261. Springer, Cham, 2016.

Guerrero, L., Marco-Ferreres, R., Serrano, A.L., Arredondo, J.J. and Cervera, M. Secondary enhancers synergise with primary enhancers to guarantee fine-tuned muscle gene expression. *Dev. Biol.*, *337*(1):16, 2010.

Gutiérrez, G., Oliver, J.L. and Marín, A. On the origin of the periodicity of three in protein coding DNA sequences. *J. Theor. Biol.*, *167*(4):413, 1994.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., Brummelkamp, T.R., de Wit, E. and Rowland, B.D. The cohesin release factor WAPL restricts chromatin loop extension. *Cell*, *169*(4):693, 2017.

Haberle, V. and Lenhard, B. Promoter architectures and developmental gene regulation. *Semin. Cell Dev. Biol.*, *57*:11, 2016.

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., Gehrig, J., Dong, X., Akalin, A., Suzuki, A.M., van IJcken, W.F.J., Armant, O., Ferg, M., Strähle, U., Carninci, P., Müller, F. and Lenhard, B. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, *507*(7492):381, 2014.

Hada, K., Asahina, M., Hasegawa, H., Kanaho, Y., Slack, F.J. and Niwa, R. The nuclear receptor gene nhr-25 plays multiple roles in the caenorhabditis elegans heterochronic gene network to control the larva-to-adult transition. *Dev. Biol.*, *344*(2):1100, 2010.

Haenni, S., Ji, Z., Hoque, M., Rust, N., Sharpe, H., Eberhard, R., Browne, C., Hengartner, M.O., Mellor, J., Tian, B. and Furger, A. Analysis of c. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic Acids Res.*, *40*(13):6304, 2012.

Haines, J.E. and Eisen, M.B. Patterns of chromatin accessibility along the anterior-posterior axis in the early drosophila embryo. *PLoS Genet.*, *14*(5):e1007367, 2018.

Hamm, D.C. and Harrison, M.M. Regulatory principles governing the maternal-to-zygotic transition: insights from drosophila melanogaster. *Open Biol.*, *8*(12):180183, 2018.

Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, *2*(3):666, 2012.

Heaton, H., Talman, A.M., Knights, A., Imaz, M., Gaffney, D., Durbin, R., Hemberg, M. and Lawniczak, M. souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes. *bioRxiv*, page 699637, 2019.

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. and Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences*, *109*(43):17507, 2012.

Heger, P., Marin, B. and Schierenberg, E. Loss of the insulator protein CTCF during nematode evolution. *BMC Molecular Biology*, *10*:84, 2009.

Henikoff, J.G., Belsky, J.A., Krassovsky, K., MacAlpine, D.M. and Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(45):18318, 2011.

Henry, G.L., Davis, F.P., Picard, S. and Eddy, S.R. Cell type–specific genomics of Drosophila neurons. *Nucleic Acids Research*, *40*(19):9691, 2012.

Herndon, L.A., Schmeissner, P.J., Dudaronek, J.M., Brown, P.A., Listner, K.M., Sakano, Y., Paupard, M.C., Hall, D.H. and Driscoll, M. Stochastic and genetic factors influence tissue-specific decline in ageing c. elegans. *Nature*, *419*(6909):808, 2002.

Herr, W. The SV40 enhancer: Transcriptional regulation through a hierarchy of combinatorial interactions. *Semin. Virol.*, *4*(1):3, 1993.

Hnisz, D., Day, D.S. and Young, R.A. Insulated neighborhoods: Structural and functional units of mammalian gene control. *Cell*, 2016.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. and Sharp, P.A. A phase separation model for transcriptional control. *Cell*, *169*:13, 2017.

Ho, J.W.K. *et al.* Comparative analysis of metazoan chromatin organization. *Nature*, *512*(7515):449, 2014.

Horn, M., Geisen, C., Cermak, L., Becker, B., Nakamura, S., Klein, C., Pagano, M. and Antebi, A. DRE-1/FBXO11-dependent degradation of BLMP-1/BLIMP-1 governs c. elegans developmental timing and maturation. *Dev. Cell*, *28*(6):697, 2014.

Hou, C., Li, L., Qin, Z.S. and Corces, V.G. Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. *Mol. Cell*, *48*(3):471, 2012.

Hsu, H.T., Chen, H.M., Yang, Z., Wang, J., Lee, N.K., Burger, A., Zaret, K., Liu, T., Levine, E. and Mango, S.E. Recruitment of RNA polymerase II by the pioneer transcription factor PHA-4. *Science*, *348*(6241):1372, 2015.

Huang, N., Seow, W.Q. and Ahringer, J. High-resolution mapping of regulatory element interactions and genome architecture using ARC-C. *bioRxiv*, page 467506, 2018.

Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A., McKay, S., Okada, H.M., Pan, J., Schulz, A.K., Tu, D., Wong, K., Zhao, Z., Alexeyenko, A., Burglin, T., Sonnhammer, E., Schnabel, R., Jones, S.J., Marra, M.A., Baillie, D.L. and Moerman, D.G. High-throughput in vivo analysis of gene expression in caenorhabditis elegans. *PLoS Biol.*, *5*(9):e237, 2007.

Hyun, K., Jeon, J., Park, K. and Kim, J. Writing, erasing and reading histone lysine methylations. *Exp. Mol. Med.*, *49*(4):e324, 2017.

Ibrahim, M.M., Karabacak, A., Glahs, A., Kolundzic, E., Hirsekorn, A., Carda, A., Tursun, B., Zinzen, R.P., Lacadie, S.A. and Ohler, U. Determinants of promoter and enhancer transcription directionality in metazoans. *Nat. Commun.*, *9*(1):4472, 2018.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. and Trifonov, E.N. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, *262*(2):129, 1996.

Ioshikhes, I., Hosid, S. and Pugh, B.F. Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.*, *21*(11):1863, 2011.

Iwafuchi-Doi, M. and Zaret, K.S. Pioneer transcription factors in cell reprogramming. *Genes Dev.*, *28*(24):2679, 2014.

Jacob, T.C. and Kaplan, J.M. The EGL-21 carboxypeptidase E facilitates acetylcholine release at caenorhabditis elegans neuromuscular junctions. *J. Neurosci.*, *23*(6):2122, 2003.

Jänes, J., Dong, Y., Schoof, M., Serizay, J., Appert, A., Cerrato, C., Woodbury, C., Chen, R., Gemma, C., Huang, N., Kissiov, D., Stempor, P., Steward, A., Zeiser, E., Sauer, S. and Ahringer, J. Chromatin accessibility dynamics across c. elegans development and ageing. *Elife*, *7*, 2018.

Jiang, C. and Pugh, B.F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.*, *10*(3):161, 2009.

Jin, V.X., Singer, G.A.C., Agosto-Pérez, F.J., Liyanarachchi, S. and Davuluri, R.V. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics*, *7*:114, 2006.

Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P. and Fire, A.Z. Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin. *Genome Res.*, *16*(12):1505, 2006.

Joshi, O., Wang, S.Y., Kuznetsova, T., Spivakov, M., Burgess, D. and Stunnenberg Correspondence, H.G. Dynamic reorganization of extremely Long-Range Promoter-Promoter interactions between two states of pluripotency. *Cell Stem Cell*, 2015.

Kadonaga, J.T., Jones, K.A. and Tjian, R. Promoter-specific activation of RNA polymerase II transcription by Sp1. *Trends in Biochemical Sciences*, *11*(1):20, 1986.

Kaletsky, R., Yao, V., Williams, A., Runnels, A.M., Tadych, A., Zhou, S., Troyanskaya, O.G. and Murphy, C.T. Transcriptome analysis of adult caenorhabditis elegans cells reveals tissue-specific gene and isoform expression. *PLoS Genet.*, *14*(8):e1007559, 2018.

Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., Gate, R.E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L.A. and Ye, C.J. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, *36*(1):89, 2018.

Kaplan, R.E.W. and Baugh, L.R. L1 arrest, daf-16/FoxO and nonautonomous control of post-embryonic development. *Worm*, *5*(2):e1175196, 2016.

Kasper, D.M., Wang, G., Gardner, K.E., Johnstone, T.G. and Reinke, V. The c. elegans SNAPc component SNPC-4 coats piRNA domains and is globally required for piRNA abundance. *Dev. Cell*, *31*(2):145, 2014.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. and Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17):2204, 2010.

Kimble, J. and Crittenden, S.L. Germline proliferation and its control. *WormBook*, *1-14.*, 2005.

Kolundzic, E., Ofenbauer, A., Bulut, S.I., Uyar, B., Baytek, G., Sommermeier, A., Seelk, S., He, M., Hirsekorn, A., Vucicevic, D., Akalin, A., Diecke, S., Lacadie, S.A. and Tursun, B. FACT sets a barrier for cell fate reprogramming in caenorhabditis elegans and human cells. *Dev. Cell*, *46*(5):611, 2018.

Kornberg, R. The location of nucleosomes in chromatin: specific or statistical. *Nature*, *292*(5824):579, 1981.

Kornberg, R.D. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(32):12955, 2007.

Kornfeld, K. Vulval development in caenorhabditis elegans. *Trends Genet.*, *13*(2):55, 1997.

Kouzarides, T. Chromatin modifications and their function. *Cell*, 2007.

Krause, M., Harrison, S.W., Xu, S.Q., Chen, L. and Fire, A. Elements regulating cell- and stage-specific expression of the c. elegans MyoD family homolog hlh-1. *Dev. Biol.*, *166*(1):133, 1994.

Krause, M. and Liu, J. Somatic muscle specification during embryonic and post-embryonic development in the nematode c. elegans. *Wiley Interdiscip. Rev. Dev. Biol.*, *1*(2):203, 2012.

Kroetz, M.B. and Zarkower, D. Cell-Specific mRNA profiling of the caenorhabditis elegans somatic gonadal precursor cells identifies suites of Sex-Biased and Gonad-Enriched transcripts. *G3*, *5*(12):2831, 2015.

Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T. and Meyer, B.J. Condensin controls recruitment of RNA polymerase II to achieve nematode x-chromosome dosage compensation. *Elife*, *2*:e00808, 2013.

Kruse, K., Díaz, N., Enriquez-Gasca, R., Gaume, X., Torres-Padilla, M.E. and Vaquerizas, J.M. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis, 2019.

Kudron, M.M. *et al.* The ModERN resource: Genome-Wide binding profiles for hundreds of drosophila and caenorhabditis elegans transcription factors. *Genetics*, *208*(3):937, 2018.

Kulakovskiy, I.V. and Makeev, V.J. DNA sequence motif: a jack of all trades for ChIP-Seq data. *Adv. Protein Chem. Struct. Biol.*, *91*:135, 2013.

Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539):317, 2015.

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S. and Narlikar, G.J. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, *547*, 2017.

Latorre, I., Chesney, M.A., Garrigues, J.M., Stempor, P., Appert, A., Francesconi, M., Strome, S. and Ahringer, J. The DREAM complex promotes gene body H2A.Z for target repression. *Genes and Development*, *29*(5):495, 2015.

Lee, C.Y.S., Lu, T. and Seydoux, G. Nanos promotes epigenetic reprograming of the germline by down-regulation of the THAP transcription factor LIN-15B. *Elife*, *6*, 2017.

Lee, J.S., Shukla, A., Schneider, J., Swanson, S.K., Washburn, M.P., Florens, L., Bhaumik, S.R. and Shilatifard, A. Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell*, *131*(6):1084, 2007.

Lee, J.S., Smith, E. and Shilatifard, A. The language of histone crosstalk. *Cell*, *142*(5):682, 2010.

Lee, J.Y. and Goldstein, B. Mechanisms of cell positioning during c. elegans gastrulation. *Development*, *130*(2):307, 2003.

Lenhard, B., Sandelin, A. and Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, *13*(4):233, 2012.

Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, *20*(17):R754, 2010.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14):1754, 2009.

Li, X. and Manley, J.L. Cotranscriptional processes and their influence on genome stability. *Genes Dev.*, *20*(14):1838, 2006.

Lin, K., Hsin, H., Libina, N. and Kenyon, C. Regulation of the caenorhabditis elegans longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nat. Genet.*, *28*(2):139, 2001.

Liu, C., Wang, M., Wei, X., Wu, L., Xu, J., Dai, X., Xia, J., Cheng, M., Yuan, Y., Zhang, P., Li, J., Feng, T., Chen, A., Zhang, W., Chen, F., Shang, Z., Zhang, X., Peters, B.A. and Liu, L. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci Data*, *6*(1):65, 2019a.

Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., Gong, F., Zhang, S., Wei, X., Wang, M., Zhao, L., Hu, L., Wang, J., Yang, H., Zhu, S., Chen, F., Lu, G., Shang, Z. and Lin, G. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nat. Commun.*, *10*(1):364, 2019b.

Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, *16*:22, 2015.

Louder, R.K., He, Y., López-Blanco, J.R., Fang, J., Chacón, P. and Nogales, E. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature*, *531*(7596):604, 2016.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12):550, 2014.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A. and Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*(5):1012, 2015.

Ma, X., Zhan, G., Sleumer, M.C., Chen, S., Liu, W., Zhang, M.Q. and Liu, X. Analysis of c. elegans muscle transcriptome using trans-splicing-based RNA tagging (SRT). *Nucleic Acids Res.*, *44*(21), 2016.

Ma'ayan, A. Introduction to Network Analysis in Systems Biology. *Science signaling*, *4*(190):tr5, 2011.

Maduro, M.F., Broitman-Maduro, G., Mengarelli, I. and Rothman, J.H. Maternal deployment of the embryonic SKN-1–>MED-1,2 cell specification pathway in c. elegans. *Dev. Biol.*, *301*(2):590, 2007.

Maniatis, T. and Reed, R. An extensive network of coupling among gene expression machines. *Nature*, *416*(6880):499, 2002.

Mann, F.G., Van Nostrand, E.L., Friedland, A.E., Liu, X. and Kim, S.K. Deactivation of the GATA transcription factor ELT-2 is a major driver of normal aging in c. elegans. *PLoS Genet.*, *12*(4):e1005956, 2016.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I. and Pugh, B.F. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, *18*(7):1073, 2008a.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., Gilmour, D.S., Albert, I. and Pugh, B.F. Nucleosome organization in the drosophila genome. *Nature*, *453*(7193):358, 2008b.

McGee, M.D., Rillo, R., Anderson, A.S. and Starr, D.A. UNC-83 IS a KASH protein required for nuclear migration and is recruited to the outer nuclear membrane by a physical interaction with the SUN protein UNC-84. *Mol. Biol. Cell*, *17*(4):1790, 2006.

McGee, M.D., Weber, D., Day, N., Vitelli, C., Crippen, D., Herndon, L.A., Hall, D.H. and Melov, S. Loss of intestinal nuclei and intestinal integrity in aging c. elegans. *Aging Cell*, *10*(4):699, 2011.

McGhee, J. The c. elegans intestine. *WormBook*, pages 1–36, 2007.

McGhee, J.D., Sleumer, M.C., Bilenky, M., Wong, K., McKay, S.J., Goszczynski, B., Tian, H., Krich, N.D., Khattra, J., Holt, R.A., Baillie, D.L., Kohara, Y., Marra, M.A., Jones, S.J.M., Moerman, D.G. and Robertson, A.G. The ELT-2 GATA-factor and the global regulation of transcription in the c. elegans intestine. *Dev. Biol.*, *302*(2):627, 2007.

McGhee, J.D., Fukushige, T., Krause, M.W., Minnema, S.E., Goszczynski, B., Gaudet, J., Kohara, Y., Bossinger, O., Zhao, Y., Khattra, J., Hirst, M., Jones, S.J.M., Marra, M.A., Ruzanov, P., Warner, A., Zapf, R., Moerman, D.G. and Kalb, J.M. ELT-2 is the predominant transcription factor controlling differentiation and function of the c. elegans intestine, from embryo to adult. *Dev. Biol.*, *327*(2):551, 2009.

McKay, S.J. *et al.* Gene expression profiling of cells, tissues, and developmental stages of the nematode c. elegans. *Cold Spring Harb. Symp. Quant. Biol.*, *68*(October):159, 2003.

McMurchy, A.N., Stempor, P., Gaarenstroom, T., Wysolmerski, B., Dong, Y., Aussianikava, D., Appert, A., Huang, N., Kolasinska-Zwierz, P., Sapetschnig, A., Miska, E. and Ahringer, J. A team of heterochromatin factors collaborates with small RNA pathways to combat repetitive elements and germline stress. pages 560–564, 2015.

Meissner, B., Warner, A., Wong, K., Dube, N., Lorch, A., McKay, S.J., Khattra, J., Rogalski, T., Somasiri, A., Chaudhry, I., Fox, R.M., Miller, 3rd, D.M., Baillie, D.L., Holt, R.A., Jones, S.J.M., Marra, M.A. and Moerman, D.G. An integrated strategy to study muscle development and myofilament structure in caenorhabditis elegans. *PLoS Genet.*, *5*(6):e1000537, 2009.

Mello, C.C., Kramer, J.M., Stinchcomb, D. and Ambros, V. Efficient gene transfer in c.elegans: extrachromosomal maintenance and integration of transforming sequences. *EMBO J.*, *10*(12):3959, 1991.

Merritt, C., Rasoloson, D., Ko, D. and Seydoux, G. 3' utrs are the primary regulators of gene expression in the c. elegans germline. *Curr. Biol.*, *18*(19):1476, 2008.

ModENCODE. Integrative analysis of the c. elegans genome by the modENCODE project. *Science*, *330*(June), 2011.

Müller, H.M., Kenny, E.E. and Sternberg, P.W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, *2*(11):e309, 2004.

Murray, J.I., Bao, Z., Boyle, T.J., Boeck, M.E., Mericle, B.L., Nicholas, T.J., Zhao, Z., Sandel, M.J. and Waterston, R.H. Automated analysis of embryonic gene expression with cellular resolution in c. elegans. *Nat. Methods*, *5*(8):703, 2008.

Murray, J.I., Boyle, T.J., Preston, E., Murray, J.I., Boyle, T.J., Preston, E., Vafeados, D., Mericle, B., Weisdepp, P., Zhao, Z., Bao, Z., Boeck, M. and Waterston, R.H. Multidimensional regulation of gene expression in the C . elegans embryo. *Genome Res.*, pages 1282–1294, 2012.

Nègre, N. *et al.* A cis-regulatory map of the drosophila genome. *Nature*, *471*(7339):527, 2011.

Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, *489*(7414):83, 2012.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J. and Heard, E. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, *485*:381, 2012.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. and Mirny, L. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *BioRxiv*, (1211), 2017.

Okkema, P.G., Harrison, S.W., Plunger, V., Aryana, A. and Fire, A. Sequence Requirements for Myosin Gene Expression and Regulation in Caenorhabditis Elegans. *Genetics*, *135*(2):385, 1993.

Ortiz, M.A., Noble, D., Sorokin, E.P. and Kimble, J. A new dataset of spermatogenic vs. oogenic transcriptomes in the nematode caenorhabditis elegans. *G3*, *4*(9):1765, 2014.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Akiyama, J.A., Afzal, V., Lopez-Rios, J., Dickel, D.E., Visel, A. and Pennacchio, L.A. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, *554*(7691):239, 2018.

Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., Waterston, R.H. and Murray, J.I. A lineage-resolved molecular atlas of c. elegans embryogenesis at single-cell resolution. *Science*, *365*(6459), 2019.

Pálfy, M., Schulze, G., Valen, E. and Vastenhouw, N.L. Chromatin accessibility established by pou5f3, sox19b and nanog primes genes for activity during zebrafish genome activation, 2019.

Pandya-Jones, A., Markaki, Y., Serizay, J., Chitiashvilli, T., Mancia, W., Damianov, A., Chronis, C., Papp, B., Chen, C.K., McKee, R., Wang, X.J., Chau, A., Leonhardt, H., Zheng, S., Guttman, M., Black, D.L. and Plath, K. An <em>Xist</em>-dependent protein assembly mediates <em>Xist</em> localization and gene silencing. *bioRxiv*, page 2020.03.09.979369, 2020.

Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, *10*(10):669, 2009.

Park, S., Hannenhalli, S. and Choi, S. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. *BMC Genomics*, *15*(1):526, 2014.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., NISC Comparative Sequencing Program, Black, B.L., Visel, A., Pennacchio, L.A., Collins, F.S., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors and NISC Comparative Sequencing Program Authors. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.*, *110*(44):17921, 2013.

Pauli, F., Liu, Y., Kim, Y.A., Chen, P.J. and Kim, S.K. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in c. elegans. *Development*, *133*(2):287, 2006.

Pecci, A., Viegas, L.R., Barañao, J.L. and Beato, M. Promoter choice influences alternative splicing and determines the balance of isoforms expressed from the Mousebcl-X gene. *J. Biol. Chem.*, *276*(24):21062, 2001.

Peckham, H.E., Thurman, R.E., Fu, Y., Stamatoyannopoulos, J.A., Noble, W.S., Struhl, K. and Weng, Z. Nucleosome positioning signals in genomic DNA. *Genome Res.*, 2007.

Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M. and Guigó, R. Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat. Genet.*, *47*(10):1158, 2015.

Pervouchine, D.D., Djebali, S., Breschi, A., Davis, C.A., Barja, P.P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L.H., Fastuca, M., Drenkow, J., Wang, H., Bussotti, G., Pei, B., Balasubramanian, S., Monlong, J., Harmanci, A., Gerstein, M., Beer, M.A., Notredame, C., Guigó, R. and Gingeras, T.R. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat. Commun.*, *6*:5903, 2015.

Petrella, L.N., Wang, W., Spike, C.A., Rechtsteiner, A., Reinke, V. and Strome, S. synmuv B proteins antagonize germline fate in the intestine and ensure c. elegans survival. *Development*, *138*(6):1069, 2011.

Pich, O., Muiños, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A. and Lopez-Bigas, N. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell*, *175*(4):1074, 2018.

Pike, J.W. and Meyer, M.B. The vitamin D receptor: new paradigms for the regulation of gene expression by 1,25-dihydroxyvitamin d(3). *Endocrinol. Metab. Clin. North Am.*, *39*(2):255, 2010.

Pliner, H.A., Shendure, J. and Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods*, *16*(10):983, 2019.

Porter, M.A., Onnela, J.P. and Mucha, P.J. Communities in networks, 2009.

Pozner, A., Lotem, J., Xiao, C., Goldenberg, D., Brenner, O., Negreanu, V., Levanon, D. and Groner, Y. Developmentally regulated promoter-switch transcriptionally controls runx1 function during embryonic hematopoiesis. *BMC Dev. Biol.*, *7*:84, 2007.

Quillien, A., Abdalla, M., Yu, J., Ou, J., Zhu, L.J. and Lawson, N.D. Robust identification of developmentally active endothelial enhancers in zebrafish using FANS-Assisted ATAC-Seq. *Cell Rep.*, *20*(3):709, 2017.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6):841, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019.

Reimand, J., Kull, M., Peterson, H., Hansen, J. and Vilo, J. g:profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, *35*(Web Server issue):W193, 2007.

Reinke, V., Krause, M. and Okkema, P. *Transcriptional regulation of gene expression in C. elegans.* WormBook, 2018.

Riddle, D.L., Blumenthal, T., Meyer, B.J. and Priess, J.R. *The Biological Model.* Cold Spring Harbor Laboratory Press, 1997.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. Integrative genomics viewer. *Nat. Biotechnol.*, *29*(1):24, 2011.

Robinson, P.J., Trnka, M.J., Bushnell, D.A., Davis, R.E., Mattei, P.J., Burlingame, A.L. and Kornberg, R.D. Structure of a complete Mediator-RNA polymerase II Pre-Initiation complex. *Cell*, *166*(6):1411, 2016.

Robson, M.I., Ringel, A.R. and Mundlos, S. Regulatory landscaping: How Enhancer-Promoter communication is sculpted in 3D. *Mol. Cell*, *74*(6):1110, 2019.

Rougvie, A.E. and Ambros, V. The heterochronic gene lin-29 encodes a zinc finger protein that controls a terminal differentiation event in Caenorhabditis elegans. *Development (Cambridge, England)*, *121*(8):2491, 1995.

Rowley, M.J., Nichols, M.H., Lyu, X., Ando-Kuri, M., Rivera, I.S.M., Hermetz, K., Wang, P., Ruan, Y. and Corces, V.G. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell*, *67*(5):837, 2017.

Roy, P.J., Stuart, J.M., Lund, J. and Kim, S.K. Chromosomal clustering of muscle-expressed genes in caenorhabditis elegans. *Nature*, *418*(6901):975, 2002.

Roy, S. *et al.* Identification of Functional Elements and Regulatory Circuits by <em>Drosophila</em> modENCODE. *Science*, *330*(6012):1787, 2010.

Saito, T.L., Hashimoto, S.I., Gu, S.G., Morton, J.J., Stadler, M., Blumenthal, T., Fire, A. and Morishita, S. The transcription start site landscape of c. elegans. *Genome Res.*, *23*(8):1348, 2013.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D.A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, *8*(6):424, 2007.

Sandhir, R. and Berman, N.E.J. Age-dependent response of CCAAT/enhancer binding proteins following traumatic brain injury in mice. *Neurochem. Int.*, *56*(1):188, 2010.

Sangaletti, R. and Bianchi, L. A method for culturing embryonic c. elegans cells. *J. Vis. Exp.*, (79):e50649, 2013.

Sankaran, V.G. and Orkin, S.H. The switch from fetal to adult hemoglobin. *Cold Spring Harb. Perspect. Med.*, *3*(1):a011643, 2013.

Satchwell, S.C., Drew, H.R. and Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, *191*(4):659, 1986.

Saunders, A., Core, L.J. and Lis, J.T. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.*, *7*(8):557, 2006.

Schaner, C.E. and Kelly, W.G. Germline chromatin. *WormBook : the online review of C. elegans biology*, *1-14.*, 2006.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. and Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.*, *22*(9):1748, 2012.

Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G. and Greenleaf, W.J. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.*, 2015.

Schilbach, S., Hantsche, M., Tegunov, D., Dienemann, C., Wigge, C., Urlaub, H. and Cramer, P. Structures of transcription pre-initiation complex with TFIIH and mediator. *Nature*, *551*(7679):204, 2017.

Schoenfelder, S. and Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.*, *20*(8):437, 2019.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G. and Zhao, K. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, *132*(5):887, 2008.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.P.Z. and Widom, J. A genomic code for nucleosome positioning. *Nature*, 2006.

Segref, A., Cabello, J., Clucas, C., Schnabel, R. and Johnstone, I.L. Fate specification and tissue-specific cell cycle control of the caenorhabditis elegans intestine. *Mol. Biol. Cell*, *21*(5):725, 2010.

Sen, P., Shah, P.P., Nativio, R. and Berger, S.L. Epigenetic mechanisms of longevity and aging. *Cell*, *166*(4):822, 2016.

Serizay, J. periodicDNA. *GitHub*, 2020a.

Serizay, J. VplotR. *GitHub*, 2020b.

Serizay, J. and Ahringer, J. Genome organization at different scales: nature, formation and function. *Curr. Opin. Cell Biol.*, *52*:145, 2018.

Serizay, J., Dong, Y., Janes, J., Chesney, M.A., Cerrato, C. and Ahringer, J. Tissue-specific profiling reveals distinctive regulatory architectures for ubiquitous, germline and somatic genes. *bioRxiv*, page 2020.02.20.958579, 2020.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, *148*(3):458, 2012.

Shao, X., Liao, J., Lu, X., Xue, R., Ai, N. and Fan, X. scCATCH: Automatic annotation on cell types of clusters from Single-Cell RNA sequencing data. *iScience*, *23*(3):100882, 2020.

Shatkin, A.J. and Manley, J.L. The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.*, *7*(10):838, 2000.

Shlyueva, D., Stampfel, G. and Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, *15*(4):272, 2014.

Skene, P.J. and Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, *6*, 2017.

Spencer, W.C., McWhirter, R., Miller, T., Strasbourger, P., Thompson, O., Hillier, L.W., Waterston, R.H. and Miller, D.M. Isolation of specific neurons from c. elegans larvae for gene expression profiling. *PLoS One*, *9*(11):1, 2014.

Spencer, W.C., Zeller, G., Watson, J.D., Henz, S.R., Watkins, K.L., Mcwhirter, R.D., Petersen, S., Sreedharan, V.T., Widmer, C., Jo, J., Reinke, V., Petrella, L., Strome, S., Stetina, S.E.V., Katz, M., Shaham, S., Ra, G. and Iii, D.M.M. A spatial and temporal map of c. elegans gene expression. *Genome Res.*, pages 325–341, 2011.

Spieth, J., Lawson, D., Davis, P., Williams, G. and Howe, K. *Overview of gene structure in C. elegans*. WormBook, 2018.

Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin. Cell Dev. Biol.*, *57*:57, 2016.

Spitz, F. and Furlong, E.E.M. Transcription Factors: From Enhancer Binding to Developmental Control. *Nature Reviews. Genetics*, *13*(9):613, 2012.

Spitz, F. and M Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, *13*, 2012.

Steiner, F.A., Talbert, P.B., Kasinathan, S., Deal, R.B. and Henikoff, S. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Res.*, *22*(4):766, 2012.

Stetina, S.E., Watson, J.D., Fox, R.M., Olszewski, K.L., Spencer, W.C., Roy, P.J., Miller, Iii, D.M. and Stetina, V. Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the c. elegans nervous system. *Genome Biol.*, *8*(7), 2007.

Stiernagle, T. *Maintenance of C. elegans.* WormBook, 2006.

Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X. and Karpen, G.H. Phase separation drives heterochromatin domain formation. *Nature*, 2017.

Struhl, K. and Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, *20*, 2013.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, 3rd, W.M., Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. Comprehensive integration of Single-Cell data. *Cell*, *177*(7):1888, 2019.

Stuart, T. and Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.*, *20*(5):257, 2019.

Sulston, J.E. and Horvitz, H.R. Post-embryonic cell lineages of the nematode, caenorhabditis elegans. *Dev. Biol.*, *56*(1):110, 1977.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J. and Mering, C.v. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, *47*(D1):D607, 2019.

Takayama, J., Faumont, S., Kunitomo, H., Lockery, S.R. and Iino, Y. Single-cell transcriptional analysis of taste sensory neuron pair in caenorhabditis elegans. *Nucleic Acids Res.*, *38*(1):131, 2010.

Tan, N.Y. and Khachigian, L.M. Sp1 Phosphorylation and Its Regulation of Gene Transcription. *Molecular and Cellular Biology*, *29*(10):2483, 2009.

Tepper, R.G., Ashraf, J., Kaletsky, R., Kleemann, G., Murphy, C.T. and Bussemaker, H.J. PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. *Cell*, *154*(3):676, 2013.

Tessarz, P. and Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. *Nat. Rev. Mol. Cell Biol.*, *15*(11):703, 2014.

Thomas, S., Li, X.Y., Sabo, P.J., Sandstrom, R., Thurman, R.E., Canfield, T.K., Giste, E., Fisher, W., Hammonds, A., Celniker, S.E., Biggin, M.D. and Stamatoyannopoulos, J.A. Dynamic reprogramming of chromatin accessibility during drosophila embryo development. *Genome Biol.*, 2011.

Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature*, *489*(7414):75, 2012.

Tian, Y., Garcia, G., Bian, Q., Steffen, K.K., Joe, L., Wolff, S., Meyer, B.J. and Dillin, A. Mitochondrial stress induces chromatin reorganization to promote longevity and UPR(mt). *Cell*, *165*(5):1197, 2016.

Timmons, L., Tabara, H., Mello, C.C. and Fire, A.Z. Inducible Systemic RNA Silencing in Caenorhabditis elegans. *Molecular Biology of the Cell*, *14*(7):2972, 2003.

Tintori, S.C., Nishimura, E.O., Golden, P., Lieb, J.D. and Correspondence, B.G. A transcriptional lineage of the early c. elegans embryo. *Dev. Cell*, *38*:430, 2016.

Tirosh, I. and Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.*, 2008.

Tolstorukov, M.Y., Kharchenko, P.V., Goldman, J.A., Kingston, R.E. and Park, P.J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes, 2009.

Travers, A., Hiriart, E., Churcher, M., Caserta, M. and Di Mauro, E. The DNA sequence-dependence of nucleosome positioning in vivo and in vitro. *J. Biomol. Struct. Dyn.*, *27*(6):713, 2010.

Trifonov, E.N. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.*, *8*(17):4041, 1980.

Turek, M. and Bringmann, H. Gene expression changes of caenorhabditis elegans larvae during molting and sleep-like lethargus. *PLoS One*, *9*(11):e113269, 2014.

Tyner, C. *et al.* The UCSC genome browser database: 2017 update. *Nucleic Acids Res.*, *45*(D1):D626, 2017.

Uesaka, M., Kuratani, S., Takeda, H. and Irie, N. Recapitulation-like developmental transitions of chromatin accessibility in vertebrates. *Zoological Lett*, *5*:33, 2019.

Uno, M., Honjoh, S., Matsuda, M., Hoshikawa, H., Kishimoto, S., Yamamoto, T., Ebisuya, M., Yamamoto, T., Matsumoto, K. and Nishida, E. A fasting-responsive signaling pathway that extends life span in c. elegans. *Cell Rep.*, *3*(1):79, 2013.

Valensisi, C., Liao, J.L., Andrus, C., Battle, S.L. and Hawkins, R.D. cChIP-seq: a robust small-scale method for investigation of histone modifications. *BMC Genomics*, *16*(1):1, 2015.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A.Z. and Johnson, S.M. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.*, *18*(7):1051, 2008.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W.M.P., van Schaik, F.M.A., Varier, R.A., Baltissen, M.P.A., Stunnenberg, H.G., Mann, M. and Timmers, H.T.M. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, *131*(1):58, 2007.

Wagner, C.R., Kuervers, L., Baillie, D.L. and Yanowitz, J.L. xnd-1 regulates the global recombination landscape in caenorhabditis elegans. *Nature*, *467*(7317):839, 2010.

Wang, J.P.Z. and Widom, J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Res.*, *33*(21):6743, 2005.

Warner, A.D., Gevirtzman, L., Hillier, L.W., Ewing, B. and Waterston, R.H. The c. elegans embryonic transcriptome with tissue, time, and alternative splicing resolution. *Genome Res.*, *29*(6):1036, 2019.

Werber, M., Wittler, L., Timmermann, B., Grote, P. and Herrmann, B.G. The tissue-specific transcriptomic landscape of the mid-gestational mouse embryo. *Development*, *141*(11):2325, 2014.

West, S.M., Mecenas, D., Gutwein, M., Aristizábal-Corrales, D., Piano, F. and Gunsalus, K.C. Developmental dynamics of gene expression and alternative polyadenylation in the caenorhabditis elegans germline. *Genome Biol.*, *19*(1):8, 2018.

White, J.G., Southgate, E., Thomson, J.N. and Brenner, S. The structure of the nervous system of the nematode caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *314*(1165):1, 1986.

White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T., Füllgrabe, A., Davis, M.P., Enright, A.J. and Busch-Nentwich, E.M. A high-resolution mRNA expression time course of embryonic development in zebrafish. *Elife*, *6*, 2017.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. and Young, R.A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2):307, 2013.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H. Welcome to the tidyverse. *JOSS*, *4*(43):1686, 2019.

Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, *34*(3):269, 2001.

Wolkow, C.A., Herndon, L.A. and Hall, D.H. The aging cuticle. *Wormatlas*, 2017.

WormBase. WormBase : Nematode Information Resource, 2020. [Online; accessed 18. Feb. 2020].

Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., Bulut, S., Tursun, B. and Akalin, A. HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Research*, *47*(11):5735, 2019.

Wright, G.M. and Cui, F. The nucleosome position-encoding WW/SS sequence pattern is depleted in mammalian genes relative to other eukaryotes. *Nucleic Acids Res.*, *47*(15):7942, 2019.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., Xu, F., Chang, Z., Na, J., Yang, X. and Xie, W. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, *534*(7609):652, 2016.

Wu, X., Shi, Z., Cui, M., Han, M. and Ruvkun, G. Repression of germline RNAi pathways in somatic cells by retinoblastoma pathway chromatin complexes. *PLoS Genet.*, *8*(3):e1002542, 2012.

Yadav, T., Quivy, J.P. and Almouzni, G. Chromatin plasticity: A versatile landscape that underlies cell fate and identity. *Science*, *361*(6409):1332, 2018.

Yáñez-Cuna, J.O., Kvon, E.Z. and Stark, A. Deciphering the transcriptional cis-regulatory code. *Trends Genet.*, *29*(1):11, 2013.

Young, M.D. and Behjati, S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data, 2018.

Yu, G., Wang, L.G., Han, Y. and He, Q.Y. clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS*, *16*(5):284, 2012.

Yuan, G.C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., Quackenbush, J., Saadatpour, A., Schroeder, T., Shivdasani, R. and Tirosh, I. Challenges and emerging directions in single-cell analysis. *Genome Biol.*, *18*(1):84, 2017.

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J. and Rando, O.J. Genome-Scale identification of nucleosome positions in s. cerevisiae. *Science*, 2005.

Zaret, K.S. and Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, 2011.

Zeiser, E., Frøkjær-Jensen, C., Jorgensen, E. and Ahringer, J. MosSCI and gateway compatible plasmid toolkit for constitutive and inducible expression of transgenes in the c. elegans germline. *PLoS One*, *6*(5), 2011.

Zentner, G.E. and Henikoff, S. Surveying the epigenomic landscape, one base at a time. *Genome Biol.*, *13*(10):250, 2012.

Zhang, S., Banerjee, D. and Kuhn, J.R. Isolation and Culture of Larval Cells from C. elegans. *PLoS ONE*, *6*(4), 2011.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y. and Wang, J. Comparative analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq systems. *Mol. Cell*, *73*(1):130, 2019.

Zhang, Y., Ma, C., Delohery, T., Nasipak, B., Foat, B.C., Bounoutas, A., Bussemaker, H.J., Kim, S.K. and Chalfie, M. Identification of genes expressed in c. elegans touch receptor neurons. *Nature*, *418*(6895):331, 2002.

Zhao, Y. and Garcia, B.A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harbor Perspectives in Biology*, *7*(9):a025064, 2015.

Zhou, V.W., Goren, A. and Bernstein, B.E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, *12*(1):7, 2011.

Zhurkin, V.B., Lysov, Y.P. and Ivanov, V.I. Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, *6*(3):1081, 1979.

Appendix A

Genome organization at different scales: nature, formation and function

This appendix introduces a review entitled "Genome organization at different scales: nature, formation and function", which I wrote toward the end of my first year of PhD. My PhD project was at first focusing on chromatin physical interactions, to continue the work initiated by Wei Qiang Seow and Ni Huang (Huang *et al.*, 2018). When Julie received an invitation to write a review about genome 3D organization and forwarded it to the lab, I considered this as a chance to capitalize the knowledge I acquired in this field as well as to improve my writing skills early on during my PhD. The publication is presented in its final edited form hereafter and have been reprinted with permission from the publisher. It has been published in the "Cell nucleus" special issue of *Current Opinion in Cell Biology* volume 52.

DOI: https://doi.org/10.1016/j.ceb.2018.03.009

Note: Since this publication, new key results have been obtained, notably clarifying the biological importance of these compartments and how they are formed (Haarhuis *et al.*, 2017; Kruse *et al.*, 2019; Nuebler *et al.*, 2017; Rowley *et al.*, 2017). These recent results have been comprehensively documented in a review which also clarifies the nomenclature in use (Beagan and Phillips-Cremins, 2020).

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Cell Biology

Genome organization at different scales: nature, formation and function

Jacques Serizay and Julie Ahringer

Since the discovery of chromosome territories, it has been clear that DNA within the nucleus is spatially organized. During the last decade, a tremendous body of work has described architectural features of chromatin at different spatial scales, such as A/B compartments, topologically associating domains (TADs), and chromatin loops. These features correlate with domains of chromatin marking and gene expression, supporting their relevance for gene regulation. Recent work has highlighted the dynamic nature of spatial folding and investigated mechanisms of their formation. Here we discuss current understanding and highlight key open questions in chromosome organization in animals.

Address
The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge CB2 1QN, United Kingdom

Corresponding author: Ahringer, Julie (ja219@cam.ac.uk)

Current Opinion in Cell Biology 2018, **52**:145–153

This review comes from a themed issue on **Cell nucleus**

Edited by **Mary C Dasso** and **Beatriz MA Fontoura**

https://doi.org/10.1016/j.ceb.2018.03.009

Introduction

The current view of nuclear organization has come predominantly from applying variations of two major types of method, (i) microscopic observations or (ii) assessment of chromatin interactions using Chromosome Conformation Capture (3C) techniques (see [1] for a concise review of these methods). Using microscopy to visualise fluorescent probes targeted to specific loci can reveal the spatial location of whole chromosomes and the relative positions of loci with respect to each other or to landmarks such as the nuclear envelope or nucleoli. On the other hand, 3C methods detect interactions between two regions of chromatin [2]. In these methods, the physical proximity of two regions of DNA within the nucleus is inferred from the frequency of ligation events generated between them following nuclear fixation and digestion. Although captured ligation events are referred to as 'chromatin interactions,' in reality they represent regions of DNA that were close enough to be ligated together, which could be because of a direct interaction between these regions or because the regions occupied the same general vicinity. Applying the 3C technique genome-wide (Hi-C), chromatin interactions can be mapped across the genome [3], with resolution related to the depth of sequencing [3,4]. Importantly, 3C methods and microscopy are highly complementary. 3C methods identify putative chromatin interactions usually from cell populations and cannot assess the frequency of occurrence of the identified interactions across the population. On the other hand, microscopy can be used to validate interactions and their frequency, by visualising large numbers of individual nuclei. Live imaging is also powerful to investigate the stability of interactions and the dynamics of the association of proteins with chromatin.

Applying these methods has led to the definition of different types of chromatin organization, such as chromosome territories, compartments, TADs, insulated domains, contact domains, and loops. Here we discuss their properties and potential relationships.

Large-scale organization: chromosome territories and compartments

The initial visualization of the spatial positioning of chromosomes by microscopy demonstrated that their organization is actively regulated within the nucleus. Individual chromosomes are spatially organized in interphase nuclei, occupying distinct chromosome territories (CTs), and adopting relatively reproducible positions in different cells with limited intermingling (Figure 1a). Additionally, inactive regions of chromatin are often found in proximity to the nuclear envelope whereas active chromatin generally has a more internal position within the nucleus [5].

More recently, 3C-based procedures have been instrumental in assessing 3D structure of individual chromosomes at increasingly higher resolution. Using Hi-C to derive average chromosome conformations from capturing pair-wise interactions in populations of cells revealed that chromosomes have two major types of structural domains, termed A and B compartments [3]. The A compartment contains active chromatin (denoted by transcriptional activity, higher chromatin accessibility and H3K36me3 deposition) while the B compartment, more compacted, is associated with inactive chromatin (denoted by low transcriptional activity, association with the nuclear lamina and H3K27me3 deposition) [3,4].

Importantly, the plaid pattern obtained by plotting pairwise correlation scores of interaction landscapes, when observed across entire chromosomes (Figure 1b), reveals that chromatin interactions are more frequent between regions of the same compartment type (A with A, and B with B) [3]. A recent Hi-C study conducted on single mammalian cells provided striking views of the spatial arrangements of A and B compartments [6••]. In modelling the arrangement of all chromosomes within the nucleus, it was shown that DNA from the A compartment is organized in an inner ring-shaped structure, while DNA from the B compartment preferentially associates with the lamina and the edges of nucleoli (Figure 1a). These results are consistent with previous studies that used microscopy to map the locations of active and inactive chromatin within nuclei [5,7].

A single-cell Hi-C study also highlighted the stochastic positions of A and B compartments in interphase cells [6••]. Although a locus on a given chromosome occupies the same compartment in different nuclei, the spatial folding of the chromosome varies between nuclei (Figure 1a). This is in agreement with the finding that positions of lamina-associated chromatin (largely corresponding to the B compartment) are not heritable. Instead, these regions are randomly redirected to the nuclear lamina or near nucleoli after mitosis, with some of them switching from a nuclear lamina position to a nucleolar associated location [8]. These studies show that chromosomes have different conformations in different cells and that A compartment active chromatin and B compartment inactive chromatin are spatially segregated both within chromosomes and globally within nuclei.

Importantly, A/B compartment organization is only observed in interphase. During mitosis, chromatin structure is radically rearranged (Figure 1c) [5,9,10]. Hi-C studies performed on synchronised cells showed that minutes after entering prophase, chromosomes lose A/B compartment organization and progressively generate and compact arrays of loops arranged around helical scaffolds of condensin I and II complexes. This raises the question of how compartment structure is reformed.

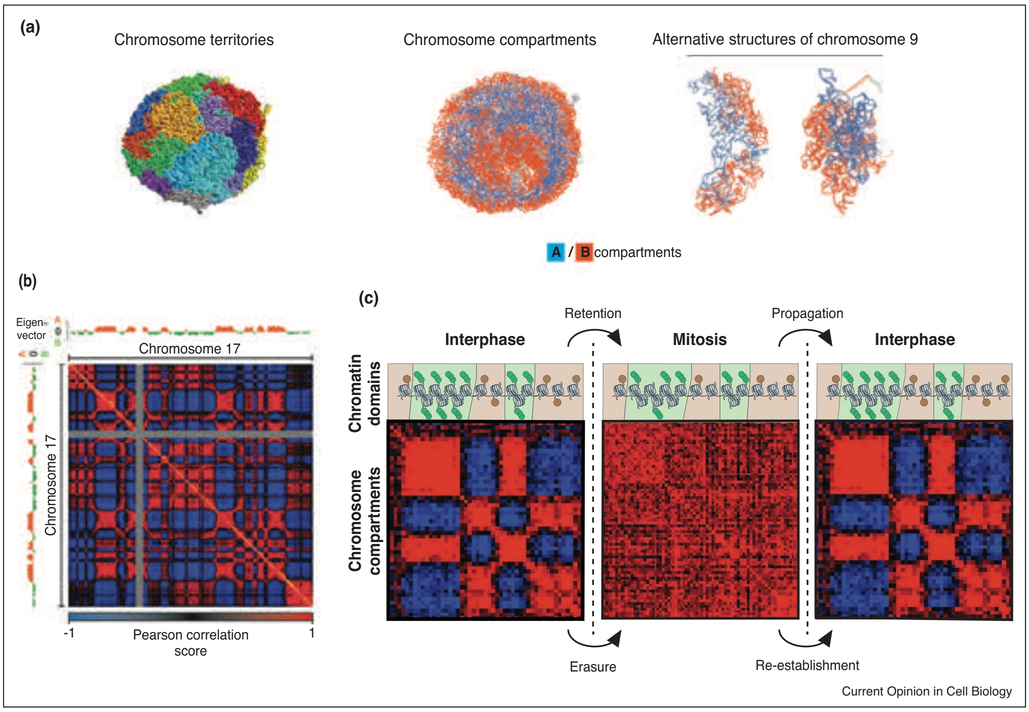
Although a relationship between transcriptional activity and compartments is clear, the mechanism of compartment formation and function are not yet understood. A striking feature of A and B compartments is their different chromatin composition, including histone modifications associated with gene activity or inactivity, respectively. Chromatin state domains, which are defined by differently marked chromatin, have been noted to subdivide the genomes of animals, and their position in the genome is relatively constant during development [11]. Interestingly, super-resolution imaging has shown that different chromatin state domains (e.g., active, inactive, Polycomb marked) have distinct types of 3D organization, with Polycomb-marked chromatin having the densest packing [12,13•]. Furthermore, altering local chromatin composition through targeting histone modifiers can drive repositioning to different compartments [14]. Whereas histone modifications can be inherited through cell division, most compartment interactions are lost during mitosis but regained after division [9,15] (Figure 1c). These data suggest a model where the formation and structure of chromosome compartments relies on chromatin domains [16••,17,18••]. In such a model, chromatin reorganization that occurs during mitosis would prevent A/B compartment interactions, while retention of chromatin domain marking would provide a framework for regenerating compartments in daughter cells (Figure 1c).

What might cause the segregation of chromatin into two types of spatial compartment? A growing body of work has shown that liquid–liquid phase separation can drive the formation of non-membrane bound compartments in the nucleus and cytoplasm [19]. For instance, the nucleolus is a phase separated compartment containing several different immiscible liquid-like sub-compartments, and HP1 containing heterochromatin has liquid-like properties and appears to form by phase separation [20–23]. The formation of these membrane-less compartments is thought to be driven by the local condensation of proteins containing unstructured regions. It is plausible that domains of particular chromatin modifications and/or proteins could drive phase-separated compartments that organize chromosome structure.

Intermediate scale organization: topologically associating domains

At a more local scale, chromatin interaction studies mostly in Drosophila and mammalian cells have described the segmentation of the genome into small physical domains of tens of kilobases up to a few megabases, and generally containing a small number (e.g., 1–10) of genes [4,24–28]. These self-interacting domains are variously termed ‘Topologically Associating Domains’ (TADs) [24–26], sub-TADs [27], ‘contact domains’ [4] and ‘insulated neighbourhoods’ [28]. They are defined based on observing frequent chromatin interactions within a region and relatively fewer interactions with neighbouring chromatin. Because these differently named domains are defined in a similar way, and it is unknown whether they are functionally different, we will refer to this class of chromosome segmentation domain as ‘topologically associating domains’ (TADs) without distinction. The properties of TADs support the view that they represent functional domains. For example, histone modification and replication timing are often similar across individual TADs [4,29]. Additionally, TADs appear to constrain the regulatory activity of enhancers [30].

Figure 1



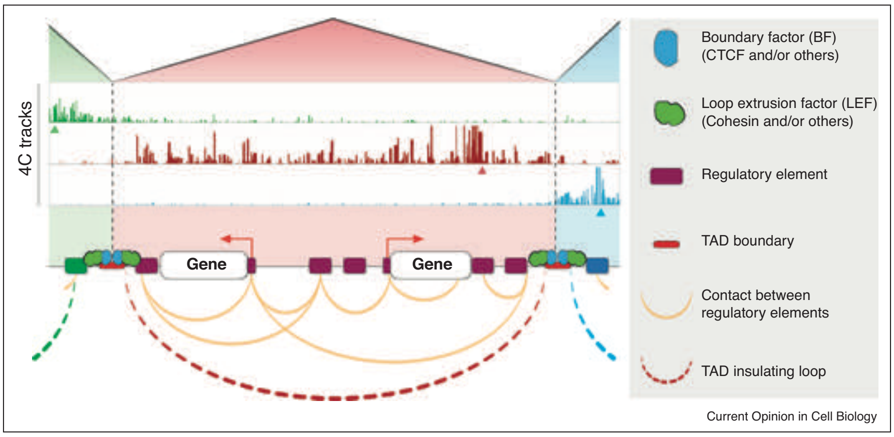
Large-scale chromosome organization. **(a)** Computational model of the 3D structure of a haploid mouse ES genome using data from a single-cell Hi-C experiment. Left: Modelled arrangement of the chromosomes within a single nucleus. Each chromosome is coloured differently. Center: Cross-section of the modelled nucleus, with A compartment in blue and B compartment in red. The B compartment is enriched at the nuclear lamina and in a central ring that surrounds the nucleolus. Right: Different structural organization of chromosome 9 modelled from two different single-cell Hi-C datasets. Figures extracted from [6••]. **(b)** Pearson correlation map of chromatin interactions on Chromosome 17 at a resolution of 500 kb. The eigenvector obtained by principal component analysis (PCA) reveals segregation of the chromosome in two compartments, A (positive values) and B (negative values). Data visualised using Juicebox and obtained from [4]. **(c)** A/B compartments are present in interphase, lost in mitosis and re-established after cell division. A/B compartment re-establishment could potentially rely on retained chromatin domains defined by histone modifications. The Pearson correlation maps of interactions are coloured as in (b). Data obtained from [9] and visualised using Juicebox [4].

TAD boundaries

The positions of TAD boundaries defined from studies on populations of cells appear relatively conserved in different cell types and across evolution [27,28,31–34]. In mammals, TAD boundaries interact more frequently with each other than with any other locus within the TAD and usually show binding of the CCCTC binding factor CTCF and the cohesin complex [4,24,25]. CTCF was initially identified as a protein with insulator activity, and its binding motifs at interacting boundaries are almost always oppositely oriented [4,34,35]. These observations have led to the notion that a chromosome domain is constrained within an insulating loop anchored by oppositely oriented CTCF proteins at the two boundaries of the domain (Figure 2). This model is supported by the analyses of mutants with deletions or inversions of CTCF sites at TAD boundaries, which led to predicted fusions or alterations of TADs [36,37].

The importance of TAD domain organization is also supported by gene expression and phenotypic alterations that are associated with TAD perturbations. In late embryonic development in the mouse, deleting a boundary between TADs that separate Hox genes alters gene expression and leads to skeletal defects [38]. In human and mouse, the inversion, deletion or duplication of TADs or TAD boundaries was shown to alter expression of genes located in the affected TADs, resulting in heart or limb pathologies [39,40], Cook syndrome [41•] or cancer susceptibility [42].

Figure 2


4C tracks
Gene
Gene
Boundary factor (BF) (CTCF and/or others)
Loop extrusion factor (LEF) (Cohesin and/or others)
Regulatory element
TAD boundary
Contact between regulatory elements
TAD insulating loop
Current Opinion in Cell Biology

Topologically associating domain (TAD) organization in mammals. Three theoretical TADs (green, red and blue) are depicted. 4C tracks from [39] are used to illustrate the 'insulating' properties of TAD boundaries (4C experiments assess the interactions between one specific locus and the rest of the genome; the assessed locus in each 4C experiment is indicated by an arrowhead). Insulating loops between TAD boundaries are represented by dashed lines while contacts between regulatory elements are represented by solid yellow lines.

Mechanism and dynamics of domain formation

The cohesin complex forms a ring structure that entraps DNA for sister chromatid cohesion in meiosis and mitosis [43]. The enrichment of cohesin at TAD boundaries in interphase cells, together with its ability to entrap DNA, has led to a 'loop extrusion' model to describe the formation of insulating loops [44,45•] (Figure 3). In this model, a loop of DNA is dynamically extruded by a loop extrusion factor (LEF) that contains cohesin (Figure 3b–f). Encountering a 'boundary factor' (BF) such as CTCF would stabilize the complex (Figure 3e–i). This model would explain the enrichment of cohesin and CTCF at TAD boundaries and the strong interaction signal observed between these regions. Of note, consistent with these roles, cohesin binding is located on the inner edge of the TAD relative to CTCF (Figure 3e) [46].

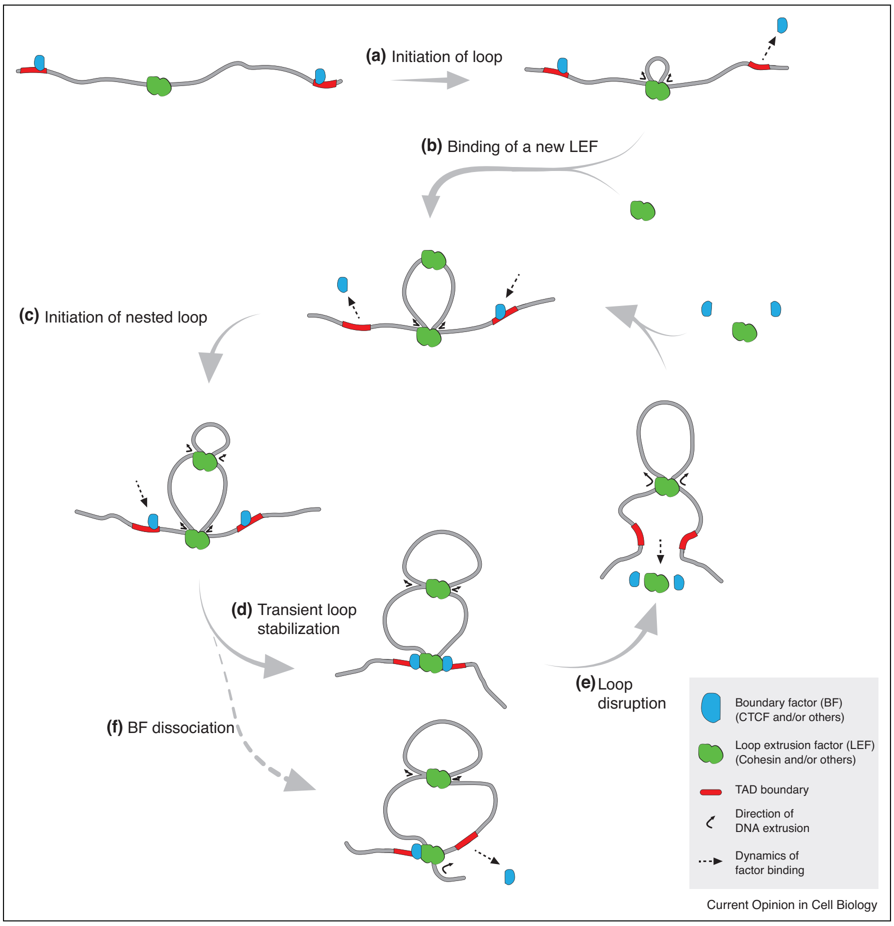
Increasing experimental and modelling studies have given strong support to the involvement of cohesin and loop extrusion in regulating chromosome organization (see [47] for a recent review). However, their mechanisms are still unclear. For example, the factors or processes providing the force for loop extrusion are not yet known. Transcriptional activity is correlated with TADs, and a recent computational model suggests that the negative supercoiling generated by transcription could provide energy for loop extrusion by 'pushing' cohesin handcuffs [22,45•,48]. However, TADs may not rely on transcription, as they start forming in Drosophila embryogenesis before the onset of the majority of zygotic transcription, and still form even after chemical inhibition of RNA polymerase [49,50].

The dynamics of cohesin and CTCF binding to chromatin argue that loops are not static structures but instead are constantly forming and collapsing (Figure 3d–f). Cohesin has a residence time of ~22 min, and CTCF, potentially playing the role of an insulating loop anchor, has a residence time of ~1 min [51••]. This implies that cohesin/CTCF loops are present only transiently even when ends are at TAD boundaries (Figure 3). The binding dynamics also explains how an extruding loop could bypass a TAD boundary to form a larger loop. Finally, dynamic binding suggests that nested extrusion would be expected to form within existing loops. A dynamic nature of chromosome domains is also supported by single-cell Hi-C studies [6••,52,53•]. Although averaged TAD boundary positions converge to those defined using a large number of cells, individual cells differ in TAD positions, and TADs can transgress conserved TAD boundaries. These studies support the view of dynamic loop formation and collapse and indicate that TADs are not stable structures (Figure 3).

Factors involved in the formation of domains and boundaries

A series of recent studies directly investigated the roles of cohesin and CTCF in interphase chromosome organization by removing them in mammalian cells [16••,54••]. Loss of CTCF, the Rad21 component of cohesin, or the cohesin loading factor Nipbl, led to the loss of TADs and

Figure 3


(a) Initiation of loop
(b) Binding of a new LEF
(c) Initiation of nested loop
(d) Transient loop stabilization
(e) Loop disruption
(f) BF dissociation
Boundary factor (BF) (CTCF and/or others)
Loop extrusion factor (LEF) (Cohesin and/or others)
TAD boundary
Direction of DNA extrusion
Dynamics of factor binding
Current Opinion in Cell Biology

Model of dynamic loop extrusion. A loop extrusion factor (LEF) binds to a segment of chromatin between two boundary factors (BF) located on TAD boundaries and initiates loop extrusion **(a)**. Although this loop is growing, a new LEF could bind within the loop **(b)**, leading to the extrusion of a secondary nested loop **(c)**. If BFs are present when the loop ends reach a TAD boundary, the loop is temporarily stabilized **(d)** then disrupts when a LEF or LEF/BF complex dissociates **(e)**. Alternatively, if a BF is not present, the loop could bypass the TAD boundary **(f)**. Loops could potentially also dissociate during any phase of extrusion. Model based on Refs. [44,45•].

loops [16••,54••,55], underlining the important structural role of both CTCF and cohesin in forming loops and insulated domains. In line with these results, the cohesin release factor WAPL was shown to restrict loop extension, as evidenced by the increase in loop size upon its depletion [56••]. However, although loops and TAD structure were lost upon CTCF or cohesin removal, A/B compartment structure remained intact, indicating that TADs and compartments are two independent types of structure [16••,54••,55]. CTCF or cohesin loss did not cause widespread transcriptional changes but only affected the expression of a limited set of genes, suggesting that much of normal gene expression is not dependent on TAD structure. It may be that compartments, which are retained, are important in this context.

The regulation of nucleosome dynamics at TAD boundaries also has the potential to control boundary 'strength' (i.e. the level of segregation of interactions on each side of the boundary). TAD boundaries are sensitive to DNAse I digestion which indicates a lower nucleosome density [57,58]. Moreover, loss of the nucleosome remodelling protein BRG1 increases nucleosome occupancy at TAD boundaries and reduces boundary strength and CTCF binding [59]. In addition to affecting the binding of boundary factors, nucleosome dynamics has the potential to affect boundary function through changing local chromatin flexibility (see [60] for further discussion).

Importantly, factors involved in domain formation appear to differ in different animals. Mammals show strong CTCF/cohesin loop anchors at TAD boundaries [4,27] whereas in Drosophila, CTCF sites are at a small proportion of TAD boundaries and are not usually in inverted orientation [18••]. Instead, Drosophila TAD boundaries are enriched for a number of other architectural proteins, such as CP190 and BEAF [57,58]. Furthermore, recent studies indicate that the prevalent strong loop anchors observed in mammals do not exist in Drosophila and that many TAD 'boundaries' are instead domains of active genes [18••,57,61•].

Domains in other organisms

The widespread TAD structure described in mammals and Drosophila has not been observed in other organisms such as *Caenorhabditis elegans* [62] and *Arabidopsis thaliana* [63]. However, this difference may be due to technical and/or biological limitations, such as Hi-C map resolution and gene spacing. Notably, TAD-like structures are visible in gene-depleted regions of these otherwise compact genomes [18••]. Although TADs are not apparent in *C. elegans*, a larger domain structure required for dosage compensation has been observed on the X chromosome [62]. Additionally, *C. elegans* autosomes are demarcated by alternating chromatin domains of H3K27me3 and H3K36me3 which contain genes with different modes of regulation [11,15]. Although the relationship between this chromatin domain pattern and spatial organization is not yet known, a similar chromatin domain organization of high versus low levels of H3K27me3 occurs in Drosophila sperm, and this pattern aligns well with TADs and TAD boundaries, respectively [17,61•]. The alignment of histone modification domains with TADs together with the finding that compartments and histone modification patterns are not generally affected by loss of cohesin or CTCF in mammals suggests that chromatin domains may provide a primary level of 1D chromatin organization and regulation upon which higher-level organizational mechanisms act.

Small-scale chromatin interactions

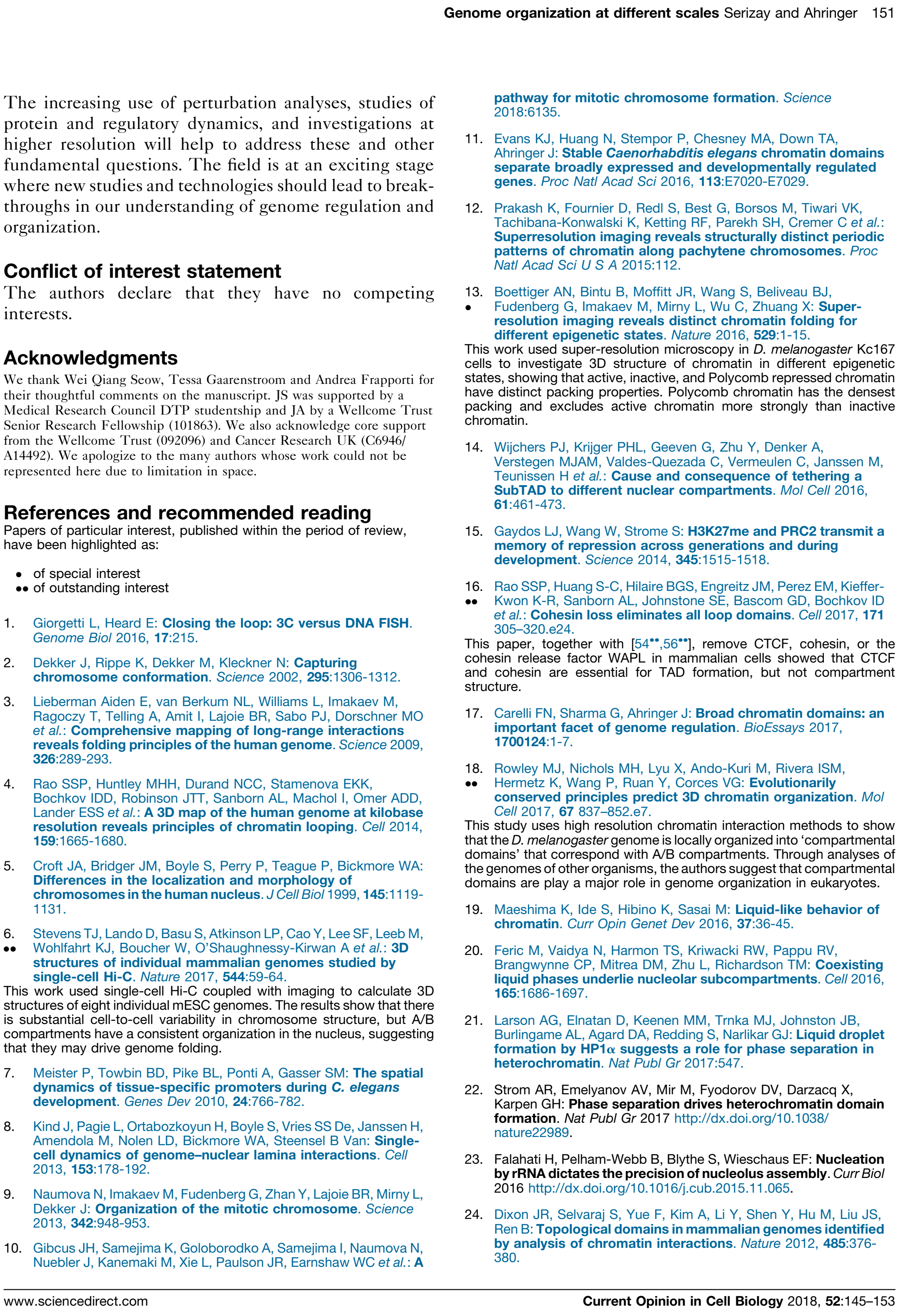
Variant 3C methods such as 4C, 5C, ChIA-PET or promoter capture, focusing on selected regions of the genomes, have uncovered extensive contacts between regulatory elements (i.e. promoters and enhancers), especially within TADs, which are not generally visible using genome wide methods such as Hi-C [64–67]. Enhancers usually contact multiple promoters and vice versa (Figure 2), and interacting regions show correlated activity, suggesting that contacts have functions in transcriptional control. Some genomic regions, such as Frequently Interacting REgions (FIREs) show particularly dense local interactions [31,68] and are associated with networks of co-expressed tissue-specific genes clustered within the same domain [68]. Their function is not yet known, but they might serve as a platform for transcription regulation in a domain. The anchors of enhancer/promoter interactions are less enriched for the combination of CTCF and cohesin compared to loop anchors at insulating TAD boundaries suggesting alternative mechanisms for their formation [4,27,31,40]. This observation could explain the relatively weak effect of CTCF and cohesin depletion on gene regulation [16••,54••].

There is evidence that both pre-established loops and *de novo* loop formation play roles in regulating transcriptional output. In Drosophila and mammals, interactions between enhancers and promoters are detected before gene activation and are associated with paused RNA polymerase, suggesting that such contacts prime later expression [31,64,69]. Similarly, during early neural lineage commitment, enrichment of transcription factor YY1 at a set of pre-established regulatory loops is associated with transcription activation [33]. During macrophage development, transcription activation is associated with both the formation of new regulatory loops and increased acetylation of H3K27 at pre-existing loop anchors [31]. Finally, directly inducing contact between an enhancer and a promoter can drive transcription, supporting the functionality of interactions [70,71].

In summary, the current data support roles for chromatin interactions in regulating gene expression and controlling chromosome organization. Yet the mechanisms that govern patterns of regulatory element interactions are still poorly understood.

Conclusion

In this review, we have highlighted the diverse and versatile mechanisms implemented within the nucleus to build spatially organized and regulated chromatin. Although recent work has provided a remarkable improvement in our understanding of genome organization, many outstanding questions remain, such as (1) How are higher-order structures such as A/B compartments formed? Do liquid–liquid phase transitions play a role? (2) How are TADs formed? What provides the force for loop extrusion? (3) How are contacts between regulatory elements made and what are their functions? What are the roles of transcription factors? (4) How many different types of loop exist, and what are their functions?



The increasing use of perturbation analyses, studies of protein and regulatory dynamics, and investigations at higher resolution will help to address these and other fundamental questions. The field is at an exciting stage where new studies and technologies should lead to breakthroughs in our understanding of genome regulation and organization.

Conflict of interest statement

The authors declare that they have no competing interests.

Acknowledgments

We thank Wei Qiang Seow, Tessa Gaarenstroom and Andrea Frapporti for their thoughtful comments on the manuscript. JS was supported by a Medical Research Council DTP studentship and JA by a Wellcome Trust Senior Research Fellowship (101863). We also acknowledge core support from the Wellcome Trust (092096) and Cancer Research UK (C6946/A14492). We apologize to the many authors whose work could not be represented here due to limitation in space.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Giorgetti L, Heard E: **Closing the loop: 3C versus DNA FISH**. *Genome Biol* 2016, **17**:215.

2. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *Science* 2002, **295**:1306-1312.

3. Lieberman Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al.*: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**:289-293.

4. Rao SSP, Huntley MHH, Durand NCC, Stamenova EKK, Bochkov IDD, Robinson JTT, Sanborn AL, Machol I, Omer ADD, Lander ESS *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell* 2014, **159**:1665-1680.

5. Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA: **Differences in the localization and morphology of chromosomes in the human nucleus**. *J Cell Biol* 1999, **145**:1119-1131.

6. •• Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, Leeb M, Wohlfahrt KJ, Boucher W, O'Shaughnessy-Kirwan A *et al.*: **3D structures of individual mammalian genomes studied by single-cell Hi-C**. *Nature* 2017, **544**:59-64.

This work used single-cell Hi-C coupled with imaging to calculate 3D structures of eight individual mESC genomes. The results show that there is substantial cell-to-cell variability in chromosome structure, but A/B compartments have a consistent organization in the nucleus, suggesting that they may drive genome folding.

7. Meister P, Towbin BD, Pike BL, Ponti A, Gasser SM: **The spatial dynamics of tissue-specific promoters during *C. elegans* development**. *Genes Dev* 2010, **24**:766-782.

8. Kind J, Pagie L, Ortabozkoyun H, Boyle S, Vries SS De, Janssen H, Amendola M, Nolen LD, Bickmore WA, Steensel B Van: **Single-cell dynamics of genome–nuclear lamina interactions**. *Cell* 2013, **153**:178-192.

9. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny L, Dekker J: **Organization of the mitotic chromosome**. *Science* 2013, **342**:948-953.

10. Gibcus JH, Samejima K, Goloborodko A, Samejima I, Naumova N, Nuebler J, Kanemaki M, Xie L, Paulson JR, Earnshaw WC *et al.*: **A pathway for mitotic chromosome formation**. *Science* 2018:6135.

11. Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J: **Stable *Caenorhabditis elegans* chromatin domains separate broadly expressed and developmentally regulated genes**. *Proc Natl Acad Sci* 2016, **113**:E7020-E7029.

12. Prakash K, Fournier D, Redl S, Best G, Borsos M, Tiwari VK, Tachibana-Konwalski K, Ketting RF, Parekh SH, Cremer C *et al.*: **Superresolution imaging reveals structurally distinct periodic patterns of chromatin along pachytene chromosomes**. *Proc Natl Acad Sci U S A* 2015:112.

13. • Boettiger AN, Bintu B, Moffitt JR, Wang S, Beliveau BJ, Fudenberg G, Imakaev M, Mirny L, Wu C, Zhuang X: **Super-resolution imaging reveals distinct chromatin folding for different epigenetic states**. *Nature* 2016, **529**:1-15.

This work used super-resolution microscopy in *D. melanogaster* Kc167 cells to investigate 3D structure of chromatin in different epigenetic states, showing that active, inactive, and Polycomb repressed chromatin have distinct packing properties. Polycomb chromatin has the densest packing and excludes active chromatin more strongly than inactive chromatin.

14. Wijchers PJ, Krijger PHL, Geeven G, Zhu Y, Denker A, Verstegen MJAM, Valdes-Quezada C, Vermeulen C, Janssen M, Teunissen H *et al.*: **Cause and consequence of tethering a SubTAD to different nuclear compartments**. *Mol Cell* 2016, **61**:461-473.

15. Gaydos LJ, Wang W, Strome S: **H3K27me and PRC2 transmit a memory of repression across generations and during development**. *Science* 2014, **345**:1515-1518.

16. •• Rao SSP, Huang S-C, Hilaire BGS, Engreitz JM, Perez EM, Kieffer-Kwon K-R, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID *et al.*: **Cohesin loss eliminates all loop domains**. *Cell* 2017, **171** 305–320.e24.

This paper, together with [54••,56••], remove CTCF, cohesin, or the cohesin release factor WAPL in mammalian cells showed that CTCF and cohesin are essential for TAD formation, but not compartment structure.

17. Carelli FN, Sharma G, Ahringer J: **Broad chromatin domains: an important facet of genome regulation**. *BioEssays* 2017, **1700124**:1-7.

18. •• Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG: **Evolutionarily conserved principles predict 3D chromatin organization**. *Mol Cell* 2017, **67** 837–852.e7.

This study uses high resolution chromatin interaction methods to show that the *D. melanogaster* genome is locally organized into 'compartmental domains' that correspond with A/B compartments. Through analyses of the genomes of other organisms, the authors suggest that compartmental domains are play a major role in genome organization in eukaryotes.

19. Maeshima K, Ide S, Hibino K, Sasai M: **Liquid-like behavior of chromatin**. *Curr Opin Genet Dev* 2016, **37**:36-45.

20. Feric M, Vaidya N, Harmon TS, Kriwacki RW, Pappu RV, Brangwynne CP, Mitrea DM, Zhu L, Richardson TM: **Coexisting liquid phases underlie nucleolar subcompartments**. *Cell* 2016, **165**:1686-1697.

21. Larson AG, Elnatan D, Keenen MM, Trnka MJ, Johnston JB, Burlingame AL, Agard DA, Redding S, Narlikar GJ: **Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin**. *Nat Publ Gr* 2017:547.

22. Strom AR, Emelyanov AV, Mir M, Fyodorov DV, Darzacq X, Karpen GH: **Phase separation drives heterochromatin domain formation**. *Nat Publ Gr* 2017 http://dx.doi.org/10.1038/nature22989.

23. Falahati H, Pelham-Webb B, Blythe S, Wieschaus EF: **Nucleation by rRNA dictates the precision of nucleolus assembly**. *Curr Biol* 2016 http://dx.doi.org/10.1016/j.cub.2015.11.065.

24. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012, **485**:376-380.

25. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J *et al.*: **Spatial partitioning of the regulatory landscape of the X-inactivation centre**. *Nature* 2012, **485**:381-385.

26. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome**. *Cell* 2012, **148**:458-472.

27. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y *et al.*: **Architectural protein subclasses shape 3D organization of genomes during lineage commitment**. *Cell* 2013 http://dx.doi.org/10.1016/j.cell.2013.04.053.

28. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K *et al.*: **Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes**. *Cell* 2014 http://dx.doi.org/10.1016/j.cell.2014.09.030.

29. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK *et al.*: **Topologically associating domains are stable units of replication-timing regulation**. *Nature* 2014, **515**:402-405.

30. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Xois Spitz F, Spitz F: **Functional and topological characteristics of mammalian regulatory domains**. *Genome Res* 2014, **24**:390-400.

31. Phanstiel DH, Van Bortle K, Spacek DV, Hess GT, Saad Shamim M, Machol I, Love MI, Lieberman Aiden E, Bassik MC, Snyder MP: **Static and dynamic DNA loops form AP-1 bound activation hubs during macrophage development**. *Mol Cell* 2017 http://dx.doi.org/10.1101/142026.

32. Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, Kim DS, Boxer LD, Cairns J, Spivakov M *et al.*: **Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation**. *Nat Genet* 2017 http://dx.doi.org/10.1038/ng.3935.

33. Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski CV, Gillis DR, Phillips-Cremins JE: **YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment**. *Genome Res* 2017, **27**:1139-1152.

34. Rudan MV, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S: **Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture**. *Cell Rep* 2015, **10**:1297-1309.

35. Bell O, Tiwari VK, Thomä NH, Schübeler D: **Determinants and dynamics of genome accessibility**. *Nat Rev Genet* 2011, **12**:554-564.

36. Guo Y, Xu Q, Canzio D, Krainer AR, Maniatis T, Wu Q: **CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function**. *Cell* 2015, **162**:900-910.

37. de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, Splinter E, Wijchers PJ, Krijger PHL, de Laat W: **CTCF binding polarity determines chromatin looping**. *Mol Cell* 2015, **60**:676-684.

38. Narendra V, Bulajic M, Dekker J, Mazzoni EO, Reinberg D: **CTCF-mediated topological boundaries during development foster appropriate gene regulation**. *Genes Dev* 2016, **30**:2657-2662.

39. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R *et al.*: **Disruptions of topological chromatin domains cause pathogenic rewiring of gene–enhancer interactions**. *Cell* 2015, **161**:1012-1025.

40. Lee DP, Lek Wen Tan W, George Anene-Nzelu C, Yiqing Li P, Anh Luu Danh T, Tiang Z, Ling Ng S, Autio MI, Jiang J, Fullwood M *et al.*: **Gene neighbourhood integrity disrupted by CTCF loss in vivo**. *Biorxiv* 2017 http://dx.doi.org/10.1101/187393.

41. • Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L *et al.*: **Formation of new chromatin domains determines pathogenicity of genomic duplications**. *Nature* 2016, **538**:265-269.

This work investigating spatial organization of the Sox9 neighbourhood in mice and humans demonstrates that genomic duplications introducing new TAD boundaries can have pathological effects.

42. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA *et al.*: **Activation of proto-oncogenes by disruption of chromosome neighborhoods**. *Science* 2016, **351**:1454-1458.

43. Merkenschlager M, Nora EP: **CTCF and cohesin in genome folding and transcriptional gene regulation**. *Annu Rev Genomics Hum Genet* 2016, **17**:17-43.

44. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J *et al.*: **Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes**. *Proc Natl Acad Sci* 2015, **112**:201518552.

45. • Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny L: **Formation of chromosomal domains by loop extrusion**. *Cell Rep* 2016, **15**:2038-2049.

These papers [44,45•] propose a loop extrusion model based on computational modelling to explain the mechanism of formation of TADs. The models explain many biological observations (e.g., loop peaks at TAD boundaries, nested TADs, CTCF motif orientation) and highlights the dynamics of loop formation.

46. Uusküla-reimand L, Hou H, Samavarchi-tehrani P, Rudan MV, Liang M, Medina-rivera A, Mohammed H, Schmidt D, Schwalie P, Young EJ *et al.*: **Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders**. *Genome Biol* 2016 http://dx.doi.org/10.1186/s13059-016-1043-8.

47. Barrington C, Finn R, Hadjur S: **Cohesin biology meets the loop extrusion model**. *Chromosom Res* 2017 http://dx.doi.org/10.1007/s10577-017-9550-3.

48. Racko D, Benedetti F, Dorier J, Stasiak A: **Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes**. *Nucleic Acids Res* 2018, **46**:1648-1660.

49. Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM: **Chromatin architecture emerges during zygotic genome activation independent of transcription**. *Cell* 2017, **169** 216–228.e19.

50. Ke Y, Xu Y, Chen X, Feng S, Liu Z, Sun Y, Yao X: **3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis**. *Cell* 2017, **170** 367–381.e20.

51. •• Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X: **CTCF and cohesin regulate chromatin loop stability with distinct dynamics**. *Elife* 2017, **6**:1-33.

Using single-molecule imaging to measure chromatin binding dynamics in mESCs, this paper shows that CTCF and cohesin do not form a stable complex, but instead rapidly exchange on chromatin, with CTCF binding having a much shorter residence time than cohesin. The binding dynamics suggest that chromatin loops are continually forming and breaking.

52. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P: **Single-cell Hi-C reveals cell-to-cell variability in chromosome structure**. *Nature* 2013:502.

53. • Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV, Abdennur N, Razin SV, Mirny L, Tachibana-Konwalski K: **Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition**. *Nature* 2017, **544**:110-114.

Single nucleus Hi-C in mouse oocytes and early zygotes shows that the positions of TADs vary substantially between individual cells, supporting their dynamic nature. The authors also observe TADs and loops, but not compartments, in maternal chromatin, suggesting the different structures are formed by different mechanisms.

54. •• Nora EP, Goloborodko A, Valton A-L, Dekker J, Mirny L, Gibcus JH, Uebersohn A, Abdennur N, Bruneau BG: **Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization**. *Cell* 2017, **169**:930-944.

This paper, together with [16••,56••], remove CTCF, cohesin, or the cohesin release factor WAPL in mammalian cells showed that CTCF and cohesin are essential for TAD formation, but not compartment structure.

55. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, Fonseca NA, Huber W, Haering C, Mirny L *et al.*: **Two independent modes of chromatin organization revealed by cohesin removal**. *Nature* 2017 http://dx.doi.org/10.1038/nature22989.

56. •• Haarhuis JHI, van der Weide RH, Blomen VA, Yáñez-Cuna JO, Amendola M, van Ruiten MS, Krijger PHL, Teunissen H, Medema RH, van Steensel B *et al.*: **The cohesin release factor WAPL restricts chromatin loop extension**. *Cell* 2017, **169** 693–707.e14.

This paper, together with [16••,54••], remove CTCF, cohesin, or the cohesin release factor WAPL in mammalian cells showed that CTCF and cohesin are essential for TAD formation, but not compartment structure.

57. Stadler M, Haines JE, Eisen MB: **Convergence of topological domain boundaries, insulators, and polytene interbands revealed by high-resolution mapping of chromatin contacts in the early *Drosophila melanogaster* embryo**. *Elife* 2017 http://dx.doi.org/10.1101/149344.

58. Bortle K Van, Nichols MH, Li L, Ong C, Takenaka N, Qin ZS, Corces VG: **Insulator function and topological domain border strength scale with architectural protein occupancy**. *Genome Biol* 2014 http://dx.doi.org/10.1186/gb-2014-15-5-r82.

59. Barutcu AR, Lajoie BR, Fritz AJ, Mccord RP, Nickerson JA, Wijnen AJ Van, Lian JB, Stein JL, Dekker J, Stein GS *et al.*: **SMARCA4 regulates gene expression and higher-order chromatin structure in proliferating mammary epithelial cells**. *Genome Res* 2016 http://dx.doi.org/10.1101/gr.201624.115.

60. Dixon JR, Gorkin DU, Ren B: **Chromatin domains: the unit of chromosome organization**. *Mol Cell* 2016, **62**:668-680.

61. • El-sharnouby S, Fischer B, Magbanua JP, Umans B, Flower R, Choo SW, Russell S, White R: **Regions of very low H3K27me3 partition the Drosophila genome into topological domains**. *PLoS One* 2017 http://dx.doi.org/10.1371/journal.pone.0172725.

The study shows a correspondence between genome organization and domains of high and low H3K27me3 levels in Drosophila sperm. High H3K27me3 levels correspond with TADs and low levels with boundaries that are enriched for housekeeping genes.

62. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ: **Condensin-driven remodelling of X chromosome topology during dosage compensation**. *Nature* 2015, **523**:240-244.

63. Liu C, Wang C, Wang G, Becker C, Zaidem M, Weigel D: **Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution**. *Genome Res* 2016 http://dx.doi.org/10.1101/gr.204032.116.

64. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EEM: **Enhancer loops appear stable during development and are associated with paused polymerase**. *Nature* 2014, **512**:96-100.

65. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA *et al.*: **Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C**. *Nat Genet* 2015, **47**:598-606.

66. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J *et al.*: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation**. *Cell* 2012, **148**:84-98.

67. Hsieh T-HS, Fudenberg G, Goloborodko A, Rando OJ: **Micro-c XL: assaying chromosome conformation from the nucleosome to the entire genome**. *Nat Methods* 2016 http://dx.doi.org/10.1038/nMeth.4025.

68. Schmitt AD, Hu M, Jung I, Lin Y, Barr CL, Ren B: **A compendium of chromatin contact maps reveals spatially active regions in the human genome**. *Cell Rep* 2016 http://dx.doi.org/10.1016/j.celrep.2016.10.061.

69. Schoenfelder S, Furlan-magaril M, Mifsud B, Tavares-cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett SW *et al.*: **The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements**. *Genome Res* 2015 http://dx.doi.org/10.1101/gr.185272.114.

70. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA: **Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor**. *Cell* 2012, **149**:1233-1244.

71. Deng W, Rupon JW, Krivega I, Breda L, Motta I, Jahn KS, Reik A, Gregory PD, Rivella S, Dean A *et al.*: **Reactivation of developmentally silenced globin genes by forced chromatin looping**. *Cell* 2014, **158**:849-860.

Appendix B

Chromatin accessibility dynamics across *C. elegans* development and aging

This appendix introduces a publication entitled “Chromatin accessibility dynamics across *C. elegans* development and ageing”. During my first year, I work on this project initiated by Yan Dong, Michael Schoof and Jürgen Jänes, and I actively participated to the final submission of this publication as well as to the corrections. With inputs and comments from all the authors, I wrote the last two sections of the publication, entitled “Extensive regulation of chromatin accessibility in development” and “Analysis of ageing clusters”. The publication is presented in its final edited form hereafter and have been reprinted with permission from the publisher. It has been published as a “Tools and Resources” article in *eLife* in 2018.

DOI: https://doi.org/10.7554/eLife.37344

eLIFE elifesciences.org

TOOLS AND RESOURCES

Chromatin accessibility dynamics across *C. elegans* development and ageing

Jürgen Jänes[1,2†], Yan Dong[1,2†], Michael Schoof[1,2‡], Jacques Serizay[1,2‡], Alex Appert[1,2], Chiara Cerrato[1,2], Carson Woodbury[1,2], Ron Chen[1,2§], Carolina Gemma[1,2#], Ni Huang[1,2], Djem Kissiov[1,2¶], Przemyslaw Stempor[1,2], Annette Steward[1,2], Eva Zeiser[1,2], Sascha Sauer[3,4], Julie Ahringer[1,2*]

[1]The Gurdon Institute, University of Cambridge, Cambridge, United Kingdom; [2]Department of Genetics, University of Cambridge, Cambridge, United Kingdom; [3]Max Delbrück Center for Molecular Medicine, Berlin, Germany; [4]Max Planck Institute for Molecular Genetics, Otto-Warburg Laboratories, Berlin, Germany

Abstract An essential step for understanding the transcriptional circuits that control development and physiology is the global identification and characterization of regulatory elements. Here, we present the first map of regulatory elements across the development and ageing of an animal, identifying 42,245 elements accessible in at least one *Caenorhabditis elegans* stage. Based on nuclear transcription profiles, we define 15,714 protein-coding promoters and 19,231 putative enhancers, and find that both types of element can drive orientation-independent transcription. Additionally, more than 1000 promoters produce transcripts antisense to protein coding genes, suggesting involvement in a widespread regulatory mechanism. We find that the accessibility of most elements changes during development and/or ageing and that patterns of accessibility change are linked to specific developmental or physiological processes. The map and characterization of regulatory elements across *C. elegans* life provides a platform for understanding how transcription controls development and ageing.

DOI: https://doi.org/10.7554/eLife.37344.001

***For correspondence:** ja219@cam.ac.uk

[†]These authors contributed equally to this work

[‡]These authors also contributed equally to this work

Present address: [§]School of Molecular and Cellular Biology, Faculty of BiologicalSciences, University of Leeds, Leeds, United Kingdom; [#]Department of Surgery and Cancer, Imperial College London, London, United Kingdom; [¶]University of California, Berkeley, United States

Competing interest: See page 19

Funding: See page 19

Received: 07 April 2018
Accepted: 25 October 2018
Published: 26 October 2018

Reviewing editor: Siu Sylvia Lee, Cornell University, United States

Introduction

The genome encodes the information for organismal life. Because the deployment of genomic information depends in large part on regulatory elements such as promoters and enhancers, their identification and characterization is essential for understanding genome function and its regulation.

Regulatory elements are typically depleted for nucleosomes, which facilitates their identification using sensitivity to digestion by nucleases such as DNase I or Tn5 transposase, termed DNA accessibility (*Sabo et al., 2006*; *Crawford et al., 2006*; *Buenrostro et al., 2013*). In different organisms, large repertoires of regulatory elements have been determined by profiling DNA accessibility genome-wide in different cell types and developmental stages (*Thomas et al., 2011*; *Kharchenko et al., 2011*; *Thurman et al., 2012*; *Yue et al., 2014*; *Kundaje et al., 2015*; *Daugherty et al., 2017*; *Ho et al., 2017*). However, no study has yet investigated regulatory element usage across the life of an animal, from the embryo to the end of life. Such information is important, because different transcriptional programs operate in different periods of life and ageing. *Caenorhabditis elegans* is ideal for addressing this question, as it has a simple anatomy, well-defined cell types, and short development and lifespan. A map of regulatory elements and their temporal dynamics would facilitate understanding of the genetic control of organismal life.

Active regulatory elements have previously been shown to have different transcriptional outputs and chromatin modifications (*Andersson, 2015*; *Kim and Shiekhattar, 2015*). Transcription is initiated at both promoters and enhancers, with most elements having divergent initiation events from two independent sites (***Core et al., 2008***; ***Kim et al., 2010***; ***De Santa et al., 2010***; ***Koch et al., 2011***; ***Chen et al., 2013***). However, promoters and enhancers differ in the production of stable transcripts. At protein-coding promoters, productive transcription elongation produces a stable transcript, whereas enhancers and the upstream divergent initiation from promoters generally produce short, aborted, unstable transcripts (***Core et al., 2014***; ***Andersson et al., 2014***; ***Rennie et al., 2017***).

Promoters and enhancers have also been shown to be differently enriched for specific patterns of histone modifications. In particular, promoters often have high levels of H3K4me3 and low levels of H3K4me1, whereas enhancers tend to have the opposite pattern of higher H3K4me1 and lower H3K4me3 (***Heintzman et al., 2007***; ***Heintzman et al., 2009***). However, in human and *Drosophila* cell lines, it was observed that H3K4me3 and H3K4me1 levels correlate with levels of transcription at regulatory elements, rather than whether the element is a promoter or an enhancer (***Core et al., 2014***; ***Henriques et al., 2018***; ***Rennie et al., 2018***). Further, analyses of genes that are highly regulated in development showed that their promoters lacked chromatin marks associated with activity (including H3K4me3), even when the associated genes are actively transcribed (***Zhang et al., 2014***; ***Pérez-Lluch et al., 2015***). Therefore, stable elongating transcription, rather than histone modification patterns, appears to be the defining feature that distinguishes active promoters from active enhancers (reviewed in ***Andersson, 2015***; ***Andersson et al., 2015***; ***Kim and Shiekhattar, 2015***; ***Henriques et al., 2018***; ***Rennie et al., 2018***).

Regulatory elements have not been systematically mapped and annotated in *C. elegans*. Promoter identification has been hampered because the 5′ ends of ~70% of protein-coding transcripts are trans-spliced to a 22nt leader sequence (***Allen et al., 2011***). Because the region from the transcription initiation site to the trans-splice site (the 'outron') is removed and degraded, the 5′ end of the mature mRNA does not mark the transcription start site. To overcome this difficulty, previous studies identified transcription start sites for some genes through profiling transcription initiation and elongation in nuclear RNA or by inhibiting *trans*-splicing at a subset of stages (***Gu et al., 2012***; ***Chen et al., 2013***; ***Kruesi et al., 2013***; ***Saito et al., 2013***). In addition, two recent studies used ATAC-seq or DNAse I hypersensitivity to map regions of accessible chromatin in some developmental stages, and predicted element function by proximity to first exons or chromatin state (***Daugherty et al., 2017***; ***Ho et al., 2017***).

Toward building a comprehensive map of regulatory elements and their use during the life of an animal, here we used multiple assays to systematically identify and annotate accessible chromatin in the six *C. elegans* developmental stages and at five time points of adult ageing. Strikingly, most elements undergo a significant change in accessibility during development and/or ageing. Clustering the patterns of accessibility changes in promoters reveals groups that act in shared processes. This map makes a major step toward defining regulatory element use during *C. elegans* life.

Results and discussion

Defining and annotating regions of accessible DNA

To define and characterize regulatory elements across *C. elegans* life, we collected biological replicate samples from a developmental time course and an ageing time course (***Figure 1A***). The developmental time course consisted of wild-type samples from six developmental stages (embryos, four larval stages, and young adults). For the ageing time course, we used *glp-1(e2144ts)* mutants to prevent progeny production, since they lack germ cells at the restrictive temperature. Five adult ageing time points were collected, starting from the young adult stage (day 1) and ending at day 13, just before the major wave of death.

Figure 1A outlines the datasets generated. For all developmental and ageing time points, we used ATAC-seq to identify accessible regions of DNA. We also sequenced strand-specific nuclear RNA (>200 nt long) to determine regions of transcriptional elongation, because previous work demonstrated that this approach could capture outron signal linking promoters to annotated exons (***Chen et al., 2013***; ***Kruesi et al., 2013***; ***Saito et al., 2013***). For the development time course, we additionally sequenced short (<100 nt) capped nuclear RNA to profile transcription initiation, profiled four histone modifications to characterize chromatin state (H3K4me3, H3K4me1, H3K36me3, and H3K27me3), and performed a DNase I concentration course to investigate the relative

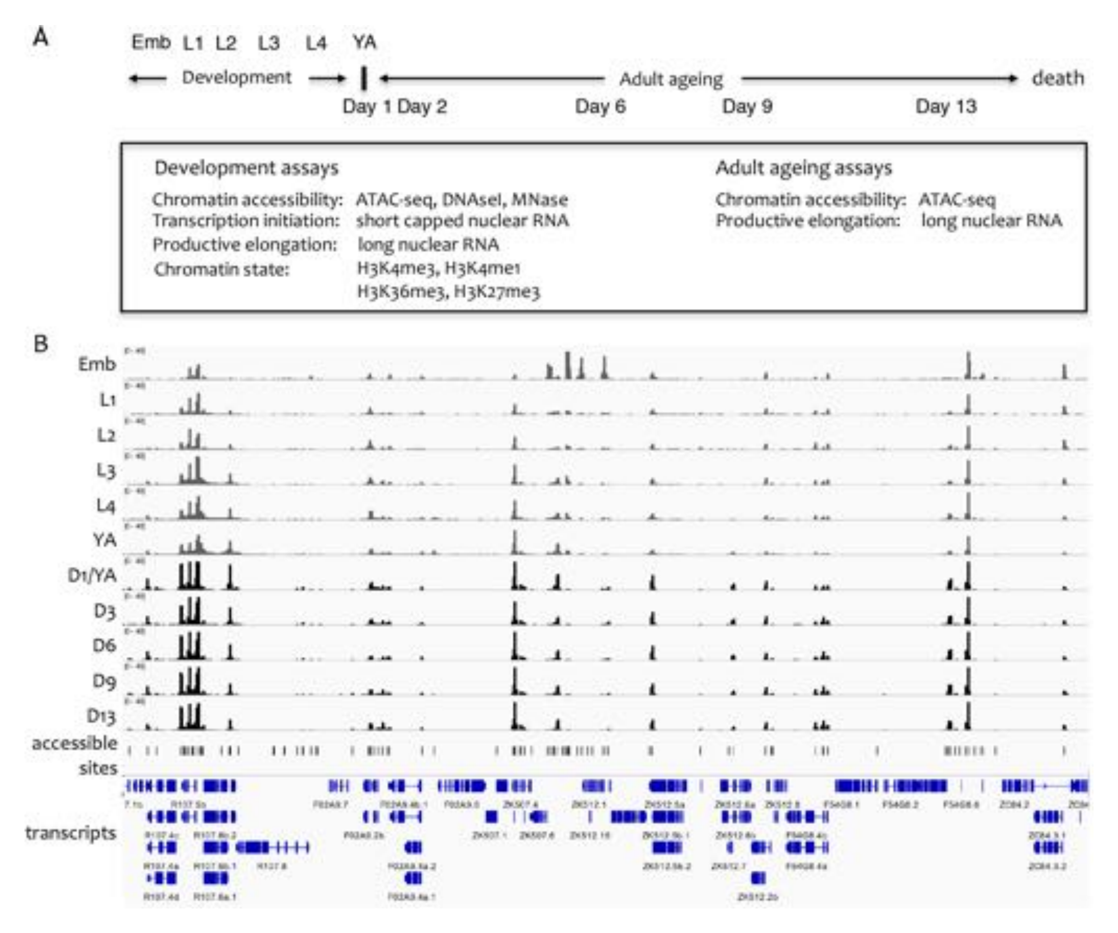


Figure 1. Overview of the project. (A) Overview of genome-wide assays and time points of developmental and ageing samples. For development samples, chromatin accessibility, transcription initiation, productive elongation, and chromatin state were profiled in six stages of wild-type animals (embryos, four larval stages, young adults). For ageing samples, chromatin accessibility and productive transcription elongation were profiled in five time points of sterile adult *glp-1* mutants (Day 1/Young adult, Day 2, Day 6, Day 9, Day 13). (B) Representative screen shot of normalized genome-wide accessibility profiles in the eleven samples (chrIII:9,041,700–9,196,700, 154 kb).

DOI: https://doi.org/10.7554/eLife.37344.002

The following source data and figure supplements are available for figure 1:

Source data 1. Accessible sites identified using ATAC-seq.

DOI: https://doi.org/10.7554/eLife.37344.006

Figure supplement 1. Comparison of ATAC-seq to concentration courses of DNase I-seq and MNase-seq.

DOI: https://doi.org/10.7554/eLife.37344.003

Figure supplement 2. Reproducibility and broad relatedness of ATAC-seq and RNA-seq data.

DOI: https://doi.org/10.7554/eLife.37344.004

Figure supplement 3. Reproducibility and broad relatedness of the histone modification data.

DOI: https://doi.org/10.7554/eLife.37344.005

accessibility of elements. Micrococcal nuclease (MNase) data were also collected for the embryo stage. As previously noted by others, we found that ATAC-seq accessibility signal is similar to that observed using a low-concentration DNase I or MNase, and that the ATAC-seq data has the highest signal-to-noise ratio (***Buenrostro et al., 2013***); ***Figure 1—figure supplement 1C***) (***Buenrostro et al., 2013***); ***Figure 1—figure supplement 1A***).

To define sites that are accessible in at least one developmental or ageing stage, focal peaks of significant ATAC-seq enrichment were identified across all developmental and ageing samples, yielding 42,245 individual elements (***Figure 1B***, ***Figure 1—source data 1***; see Materials and methods for details). Of these, 72.8% overlap a transcription factor binding site (TFBS) mapped by the modENCODE or modERN projects (***Araya et al., 2014***; ***Kudron et al., 2018***), supporting their potential regulatory functions (***Figure 2—figure supplement 1A***).

Two recent studies reported accessible regions in *C. elegans* identified using DNase I hypersensitivity or ATAC-seq (***Ho et al., 2017***; ***Daugherty et al., 2017***). The 42,245 accessible elements defined here overlap 33.7% of (***Ho et al., 2017***) DNase I hypersensitive sites and 47.9% of (***Daugherty et al., 2017***) ATAC-seq peaks (***Figure 2—figure supplement 1B,C***). Examining the non-overlapping sites from pairwise comparisons, it appears that differences in peak calling methods account for some of the differences. Accessible regions determined here required a focal peak of enrichment, whereas the other studies found both focal sites and broad regions with increased signal. Consistent with these differences in methods, sites unique to the two studies are enriched for exonic chromatin, depleted for both TFBS and transcription initiation sites, and often found in broad regions of increased accessibility across transcriptionally active gene bodies (***Figure 2—figure supplement 1B–E***). Similarly, using MACS2 to call peaks on the ATAC-seq data reported here, as used by ***Daugherty et al. (2017)***, identified a group of exon enriched sites not found using our peak calling method (***Figure 2—figure supplement 2A***). However, the fraction of such sites is relatively small indicating that other differences also contribute, such as signal-to-noise or nematode growth methods.

To functionally classify elements, we annotated each of the 42,245 elements for transcription initiation and transcription elongation signals on both strands (***Figure 2A,B***; ***Figure 2—source data 1***; see Materials and methods for details). Overall, 37.1% of elements had promoter activity, defined by a significant increase in transcription elongation signal originating at the element in at least one stage and one direction. Promoters were assigned to protein-coding or pseudogenes if continuous transcription elongation signal extended from the element to an annotated first exon (covering the outron). Promoters were unassigned if transcription elongation signal was not linked to an annotated gene. We observed detectable transcription initiation signal at 82.3% of elements (***Figure 2—source data 1***); those with no significant transcription elongation signal in either direction were annotated as putative enhancers (hereafter referred to as ’enhancers’). The remaining elements had no detectable transcriptional activity or overlapped ncRNAs (tRNA, snRNA, snoRNA, rRNA, or miRNA) (***Figure 2B***; ***Figure 2—source data 1***). We found that accessible sites are enriched for being located within outrons or intergenic regions (***Figure 2—figure supplement 3***).

Within the promoter class, we defined 15,572 protein-coding coding promoters: 11,478 elements are unidirectional promoters and 2118 are divergent promoters that drive expression of two oppositely oriented protein-coding genes (***Figure 2—source data 1***). In total, promoters were defined for 11,196 protein-coding genes, with 3000 genes having >1 promoter (***Figure 2C***). The protein-coding promoter annotations show good overlap with four sets of TSSs previously defined based on mapping transcription (***Chen et al., 2013***; ***Kruesi et al., 2013***; ***Saito et al., 2013***; ***Gu et al., 2012***); 76.8–85.1%; ***Figure 2—figure supplement 5***). Enhancers (n = 19,231) were assigned to a gene if they are located within the region from its most upstream promoter to its gene end; 6668 genes have at least one associated enhancer, and 3240 genes have >1 enhancer (***Figure 2C***).

The locations of unassigned promoters (n = 3106) suggest different potential functions. A large fraction (35.1%) generate antisense transcripts within the body of a protein coding gene, suggesting a possible role in regulating expression of the associated gene (***Figure 2—figure supplement 5***). Another large group (38.4%) produce antisense transcripts from an element that is a protein coding promoter in the sense direction, a pattern seen in many mammalian promoters, termed upstream antisense (uaRNA) or promoter upstream (PROMPT) transcripts (***Figure 2—figure supplement 5***; ***Preker et al., 2008***; ***Flynn et al., 2011***; ***Sigova et al., 2013***). Most of the rest (21.7%) are intergenic and may define promoters for unannotated transcripts.

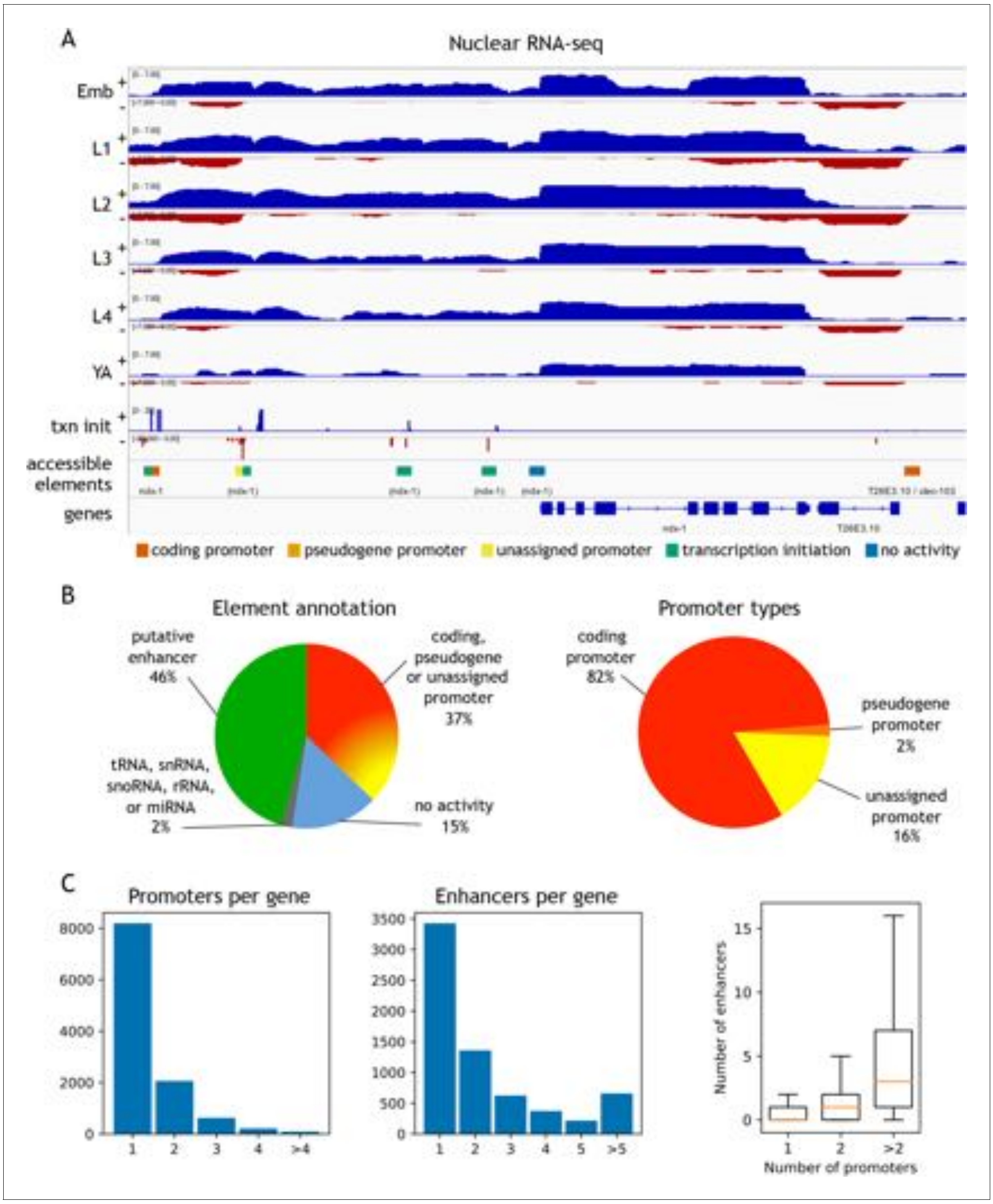


Figure 2. Annotation of accessible elements. (A) Top, strand-specific nuclear RNA in each developmental stage monitors transcription elongation; plus strand, blue; minus strand, red. Below is transcription initiation signal, accessible elements (colored by annotation), and gene models (chrI:12,675,000–

Figure 2 continued on next page

Figure 2 continued

12,683,400, 8.4 kb). The left side of each element is colored by the reverse strand annotation whereas the right side of an element is colored by the forward strand annotation (color key at bottom). (**B**) Left, distribution of accessible sites in four categories: promoters (one or both strands), putative enhancers, no activity, or overlapping a tRNA, snRNA, snoRNA, rRNA, or miRNA. Right, distribution of different types of promoter annotations. (**C**) Left, distribution of the number of promoters and enhancers per gene; right, boxplot shows that genes with more promoters also have more enhancers.

DOI: https://doi.org/10.7554/eLife.37344.007

The following source data and figure supplements are available for figure 2:

Source data 1. Regulatory annotation of accessible sites.

DOI: https://doi.org/10.7554/eLife.37344.014

Figure supplement 1. Comparisons to previous accessibility maps.

DOI: https://doi.org/10.7554/eLife.37344.008

Figure supplement 2. Effect of differences in peak calling methods on the types of identified accessible sites.

DOI: https://doi.org/10.7554/eLife.37344.009

Figure supplement 3. Genomic locations of accessible sites.

DOI: https://doi.org/10.7554/eLife.37344.010

Figure supplement 4. Comparison to published TSS maps.

DOI: https://doi.org/10.7554/eLife.37344.011

Figure supplement 5. Types of unassigned promoters.

DOI: https://doi.org/10.7554/eLife.37344.012

Figure supplement 6. Transgenic tests of annotated promoters and enhancers for promoter activity.

DOI: https://doi.org/10.7554/eLife.37344.013

Patterns of histone marks at promoters and enhancers

Promoters and enhancers show general differences in patterns of histone modifications, such as higher levels of H3K4me3 at promoters or H3K4me1 at enhancers, and chromatin states are frequently used to define elements as promoters or enhancers (***Heintzman et al., 2007***; ***Ernst and Kellis, 2010***; ***Ernst et al., 2011***; ***Kharchenko et al., 2011***; ***Hoffman et al., 2013***; ***Daugherty et al., 2017***). However, it has been shown that H3K4me3 levels correlate with transcriptional activity rather than with function (***Pekowska et al., 2011***; ***Core et al., 2014***; ***Andersson et al., 2014***; ***Henriques et al., 2018***; ***Rennie et al., 2018***), suggesting that defining regulatory elements solely based on chromatin state is likely to lead to incorrect annotations.

To further investigate the relationship between chromatin marking and element function, we mapped four histone modifications at each developmental stage (H3K4me3, H3K4me1, H3K27me3, H3K36me3) and examined their patterns around coding promoters and enhancers. As expected, many coding promoters had high levels of H3K4me3 and were depleted for H3K4me1 (***Figure 3A***). Moreover, enhancers had generally low levels of H3K4me3 and higher levels of H3K4me1 than promoters (***Figure 3A***). However, many elements did not have these patterns. For example, about 50% of coding promoters have a high level of H3K4me1 and no or low H3K4me3 marking (***Figure 3A***).

To investigate the nature of these patterns, we examined coefficients of variation of gene expression (CV; ***Gerstein et al., 2014***) of the associated genes. Genes with broad stable expression across cell types and development, such as housekeeping genes, have low variation of gene expression levels and hence a low CV value. In contrast, genes with regulated expression, such as those expressed only in particular stages or cell types have a high CV value. We found a strong inverse correlation between a gene's CV value and its promoter H3K4me3 level (−0.64, $p<10^{-15}$, Spearman's rank correlation; ***Figure 3***; ***Figure 3—figure supplement 1A***). Furthermore, promoters with low or no H3K4me3 marking are enriched for H3K27me3 (***Figure 3***; ***Figure 3—figure supplement 1A***), which is associated with regulated gene expression (***Tittel-Elmer et al., 2010***; ***Pérez-Lluch et al., 2015***; ***Evans et al., 2016***). These results support the view that H3K4me3 marking may be a specific feature of promoters with broad stable activity, consistent with the finding that active promoters of regulated genes lack H3K4me3 (***Pérez-Lluch et al., 2015***). The profiling here was done in whole animals, which may have precluded detecting modifications occurring in a small number of nuclei. Nevertheless, the results indicate that chromatin state alone is not a reliable metric for element annotation. Histone modification patterns at many promoters resemble those at enhancers, and vice versa.

Promoters and enhancers also share sequence features. Both are enriched for initiator INR elements, although enhancers have a slightly lower INR frequency (***Figure 3B*** and ***Figure 3—figure***

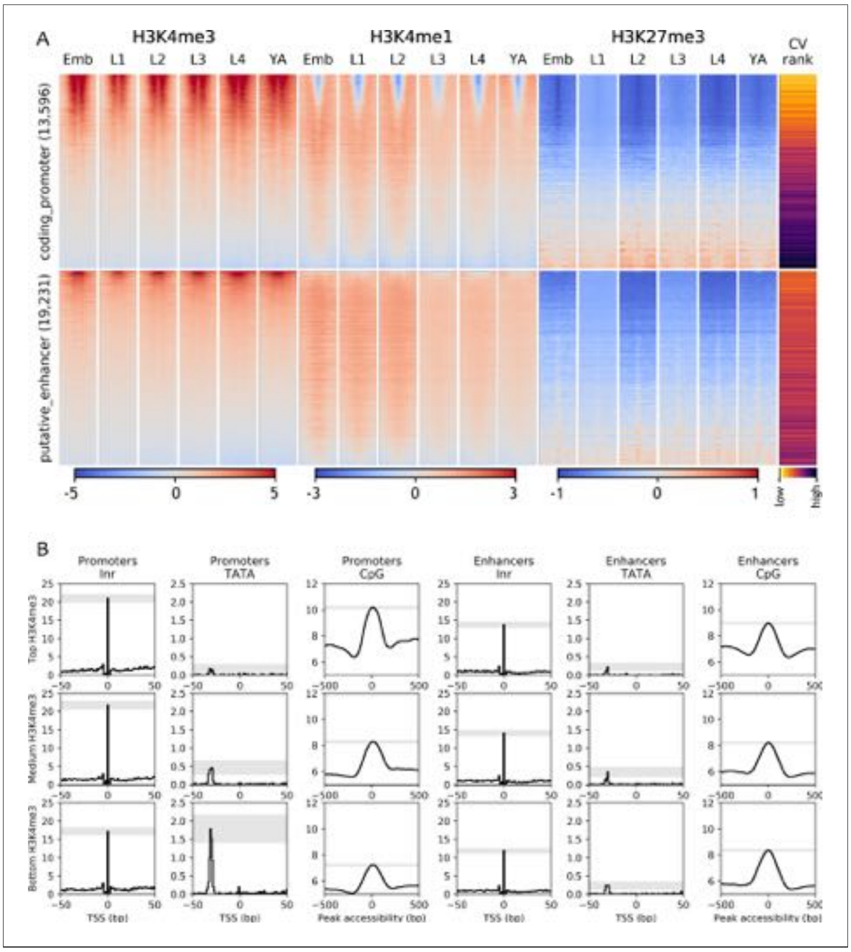


Figure 3. Chromatin state and sequence features of promoters and enhancers. (**A**) Heatmaps of indicated histone modifications and CV values at coding promoters (top), and enhancers (bottom), aligned at element midpoints. Elements are ranked by mean H3K4me3 levels. Low CV values indicate broad expression across development and cell types and high CV values indicate regulated expression. Promoters of genes with low CV values have high H3K4me3 levels. (**B**) Distribution of initiator Inr motif, TATA motif, and CpG content at coding promoters and enhancers, separated by H3K4me3 level (top, middle, and bottom thirds). Grey-shaded regions represent 95% confidence intervals of the sample mean at the genomic position with the highest signal.

DOI: https://doi.org/10.7554/eLife.37344.015

The following figure supplement is available for figure 3:

Figure supplement 1. Chromatin state and sequence features of promoters and enhancers sorted by CV value.

DOI: https://doi.org/10.7554/eLife.37344.016

supplement 1B). Promoters and enhancers are also both enriched for CpG dinucleotides (***Figure 3B*** and ***Figure 3—figure supplement 1B***). Promoters with high H3K4me3 and low CV values (broadly expressed genes) have the highest CpG content, whereas those with low H3K4me3 and high CV values have the lowest CpG content (***Figure 3B*** and ***Figure 3—figure supplement 1B***). Promoters also differ from enhancers by the presence of TATA motifs, which occur predominantly at genes with low

H3K4me3,and high CV values (i.e. with regulated expression; **Figure 3B** and **Figure 3—figure supplement 1B**).

Promoters and enhancers can drive gene expression in an orientation independent manner

To validate the promoter annotations, we compared them with studies where small regions of DNA had been defined as promoters using transgenic assays. These comprised 10 regions are defined based on transcription initiation signal (**Chen et al., 2014**), nine regions defined based on proximity to a germ line gene (**Merritt et al., 2008**), and four defined by proximity to the first exon of a muscle expressed gene (**Hunt-Newbury et al., 2007**). Of these 23 regions, 21 overlap an element in our set of accessible sites, 19 of which are annotated as protein coding promoters (**Figure 2—figure supplement 6A**). One of the remaining two is annotated as an enhancer and the other overlaps an accessible element for which no transcriptional signal was detected. We further directly tested three elements annotated as promoters (for *hlh-2, ztf-11* and *bed-3* genes), and found that all three drove robust expression of a histone-GFP reporter (**Figure 2—figure supplement 6A**). Overall, there is good concordance between promoter annotation and promoter activity.

Most of the elements annotated as protein-coding promoters are flanked by bidirectional transcription initiation signal (74.0%), similar to the pattern seen in mammals. Most (82.6%) are unidirectional promoters, producing a protein-coding transcript in one direction, but no stable transcript from the upstream initiation site. To test whether such upstream antisense initiation sites could function as promoters, we inverted the orientation of two active unidirectional promoters (*ztf-11* and F58D5.5). If the lack of in vivo transcription elongation was a property of the element or initiation site itself, the GFP fusion should not be expressed. However, we observed that the two inverted unidirectional promoters both drove GFP expression. The expression patterns generated were similar in both orientations, although the *ztf-11* promoter was weaker when inverted (**Figure 2—figure supplement 6B,C**). These results suggest that signals for productive elongation occur downstream of the transcription initiation site.

Similar to the upstream antisense transcription initiation observed at promoters, enhancers also show transcription initiation signals but generally do not produce stable transcripts (**Core et al., 2014**; **Andersson et al., 2014**). Previous studies have reported that some enhancers can function as promoters in transgenic assays and also at endogenous loci (**Kowalczyk et al., 2012**; **Leung et al., 2015**; **Nguyen et al., 2016**; **van Arensbergen et al., 2017**; **Mikhaylichenko et al., 2018**). To assess the potential promoter activities of *C. elegans* enhancers, we directly fused 12 putative enhancers that had transcription initiation signal in embryos to a histone-GFP reporter gene and assessed transgenic strains for embryo expression. Two of the tested enhancers are located in introns, and one of these, from the *bro-1* gene, has been previously validated as an enhancer (**Brabin et al., 2011**); most of the others are associated with the *hlh-2* or *ztf-11* genes. We found that 10 of 12 tested regions drove reporter expression in embryos, including the two intronic enhancers (**Figure 2—figure supplement 6B,C**). Whereas the *hlh-2* and *ztf-11* promoters drove strong, broad expression, the associated enhancers were active in a smaller number of cells and expression levels were overall lower (**Figure 2—figure supplement 6B,C**). We also tested two enhancers in inverted orientation and found that both showed similar activity in both orientations, as observed for the two tested promoters (**Figure 2—figure supplement 6B,C**). The percentage of enhancers that functioned as active promoters is higher than that observed in a cell-based assay (**Nguyen et al., 2016**), possibly because all cell types are tested in an intact animal. Episomal-based assays have also been reported to underestimate activity (**Inoue et al., 2017**).

Extensive regulation of chromatin accessibility in development

We observed extensive changes in chromatin accessibility across development, with most elements showing a significant difference within the developmental time course (71%,>=2 fold change, FDR < 0.01; **Figure 4—source data 1**; see Materials and methods). To investigate how accessibility relates to gene expression, we focused on the 13,596 elements annotated as protein-coding promoters. Of these, 10,199 displayed significant changes in accessibility in development, with the remaining 3397 promoters classified as having stable accessibility. We note that the detected changes could be due to regulation of accessibility, or alternatively to changes in cell number during development (e.g. the number of germ line nuclei increases from two in L1 larvae to ~2000 in young adults).

We reasoned that promoters having similar patterns of accessibility changes over development may regulate genes that function in shared processes and be regulated by shared sets of transcription factors. To investigate this, we applied *k*-medoid clustering to the 10,199 promoters with developmental changes in accessibility, defining 16 clusters (***Figure 4A***, ***Figure 4—figure supplement 1***, ***Figure 4—figure supplement 2***, and ***Figure 4—source data 1***; see Materials and methods). Within clusters, we observed that promoter accessibility and nuclear RNA levels are usually correlated (mean r = 0.47 (sd = 0.11) across all clusters), indicating that accessibility is a good metric of promoter activity and overall gene expression (***Figure 4—figure supplement 1*** and ***Figure 4—figure supplement 2***).

To investigate whether the shared patterns of accessibility changes over development identify promoters of genes involved in common processes, we took advantage of recent single-cell profiling data obtained from L2 larvae, which provides gene expression measurements in different tissues (***Cao et al., 2017***). We find that half of the developmental promoter clusters are enriched for genes with tissue biased expression (***Figure 4A***, ***Figure 4—figure supplement 1*** and ***Figure 4—figure supplement 2***). Based on these patterns of enrichment, we defined four gonad promoter clusters (G1-G4), two intestine clusters (I1, I2), one hypodermal cluster (H) and one cluster enriched for neural and muscle expression (N + M) (***Figure 4A***, ***Figure 4—figure supplement 1*** and ***Figure 4—figure supplement 2***). Genes associated with the remaining eight promoter clusters (Mix1–8) are generally expressed in multiple tissues, but predominantly in the soma (***Figure 4A***, ***Figure 4—figure supplement 1*** and ***Figure 4—figure supplement 2***). As expected, genes linked to the stable promoters are widely expressed. Interestingly, within a tissue, promoter clusters can exhibit similar variations in accessibility but with different amplitude. For instance, gonad clusters G1 and G2 both show a sharp increase in accessibility at the L3 stage; however, the increase is 1.5-fold larger in G2 than in G1. The gonad clusters are generally characterized by an increase of promoter accessibility starting in L3 when germ cell number strongly increases.

To further investigate promoter clusters sharing accessibility dynamics, we performed Gene Ontology analyses on the associated genes. As expected, we found that clusters containing genes enriched for expression in a particular tissue are also associated with GO terms related to that tissue (***Figure 4A***, ***Figure 4—figure supplement 1*** and ***Figure 4—figure supplement 2***). For instance, cluster H contains genes highly expressed in hypodermis and GO terms linked to cuticle development. Of note, the four accessibility clusters enriched for expression in germ line are associated with GO terms for different sets of germ line functions (***Figure 4—figure supplement 1*** and ***Figure 4—figure supplement 2***). Similarly, the two intestinal clusters also identify genes with different types of intestinal function. Furthermore, accessibility dynamics can reflect the temporal function of the associated promoters. For instance, cluster Mix4 has GO terms indicative of neuronal development and highest accessibility in the embryo, when many neurons develop. These results suggest that promoter clusters contain genes acting in a shared process and having a similar mode of regulation.

To identify potential transcriptional regulators, we asked whether the binding of particular transcription factors is enriched in any promoter clusters, using TF binding data from the modENCODE and modERN projects (***Boyle et al., 2014***; ***Kudron et al., 2018***). TFs with enriched binding were found for each cluster (***Figure 5A***), and the expression of such TFs was generally enriched in the expected tissue. For example, we found that ELT-2, an intestine-specific GATA protein (***Fukushige et al., 1998***), has enriched binding at promoters in intestinal clusters 1 and 2. Similarly, hypodermal transcription factors BLMP-1 (***Horn et al., 2014***), NHR-25 (***Gissendanner and Sluder, 2000***) and ELT-3 (***Gilleard et al., 1999***) are enriched in the hypodermal promoter cluster, and binding of the germ line XND-1 factor (***Wagner et al., 2010***) is enriched in the germ line clusters of promoters. We also identified novel tissue-specific associations for uncharacterized transcription factors, such as ZTF-18 and ATHP-1 with germ line promoter clusters and CRH-2 with the intestinal clusters (***Figure 5A***). These results agree and extend those of ***Cao et al. (2017)***, who identified TFs for which binding was correlated with cell-type-specific expression levels.

We also observed differences in TF-binding enrichments between promoter clusters associated with the same tissue. For example, Clusters G1-G4 all contain promoters associated with germline-enriched genes (***Figure 4A***). However, distinct binding enrichments are observed in promoters in G1-G2 compared to those in G3-G4, with the latter showing enrichment for LIN-35 and DPL-1, two

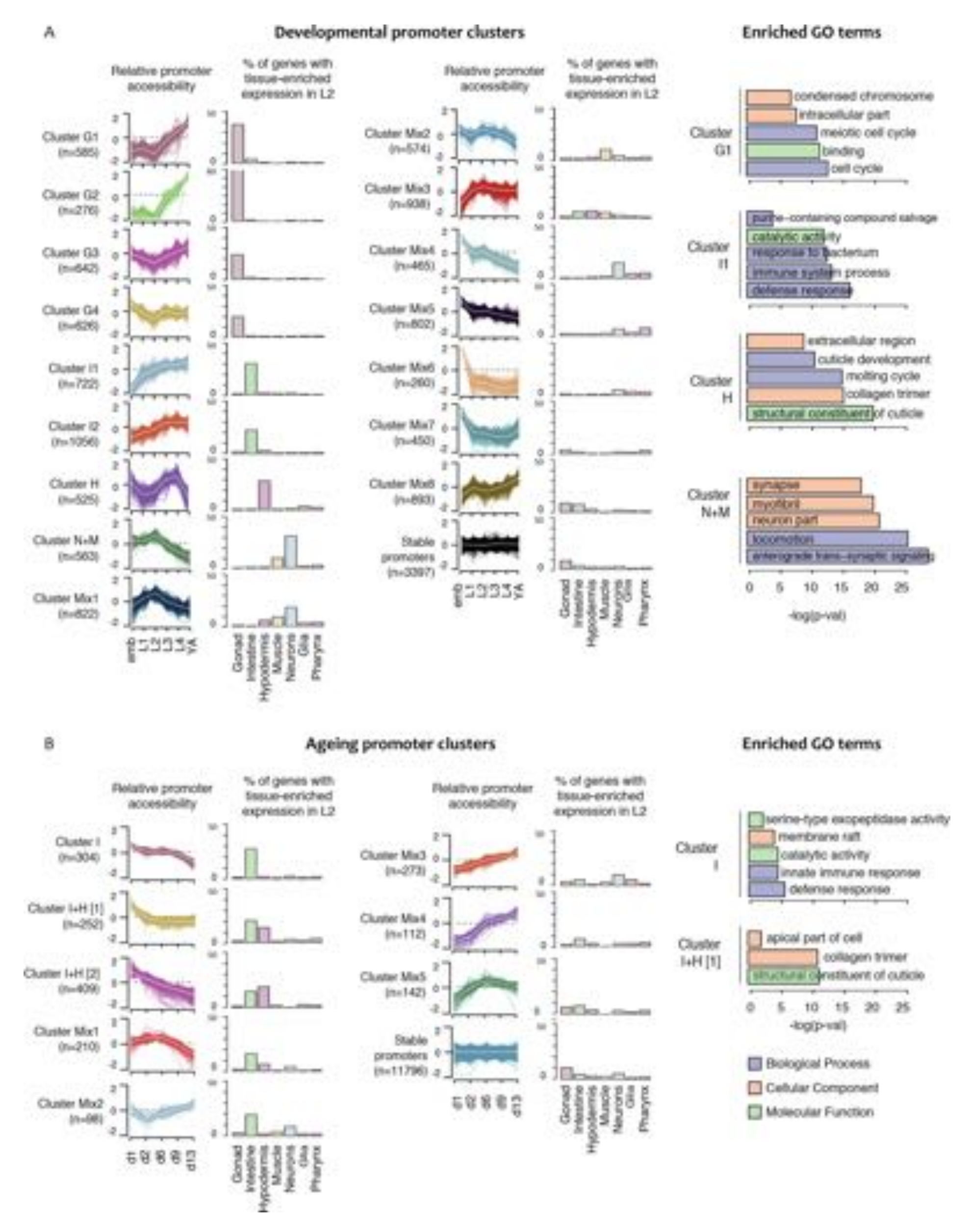


Figure 4. Shared dynamics of promoter accessibility in development and ageing. Clusters of promoters with shared relative accessibility patterns across (A) development or (B) ageing. Relative promoter accessibility is log2 of the depth-normalized ATAC-seq coverage at a given time point divided by the mean ATAC-seq coverage across the time series (see Materials and methods). The percentage of associated genes that have enriched expression in

Figure 4 continued on next page

Figure 4 continued

the indicated tissues was determined from single-cell L2 larval RNA-seq data (***Cao et al., 2017***; see Materials and methods). Right hand panels show examples of GO terms enriched in genes associated with development or ageing clusters.

DOI: https://doi.org/10.7554/eLife.37344.017

The following source data and figure supplements are available for figure 4:

Source data 1. Element accessibility dynamics and promoter accessibility clusters in development and ageing.

DOI: https://doi.org/10.7554/eLife.37344.021

Figure supplement 1. Characteristics of developmental promoter clusters (continued in ***Figure 4—figure supplement 2***).

DOI: https://doi.org/10.7554/eLife.37344.018

Figure supplement 2. Characteristics of developmental promoter clusters (continued from ***Figure 4—figure supplement 1***).

DOI: https://doi.org/10.7554/eLife.37344.019

Figure supplement 3. Characteristics of ageing promoter clusters.

DOI: https://doi.org/10.7554/eLife.37344.020

members of the DREAM complex, which controls cell cycle progression (***Figure 5A***). Taken together, the results suggest that promoters with shared accessibility patterns have shared cell- and process-specific activity, and they highlight potential regulators that are candidates for future studies.

Analysis of ageing clusters

We next focused on chromatin accessibility changes during ageing. In contrast to the development time course, the accessibility of most promoters is stable during ageing, with only 13% (n = 1,800) of promoters showing changes (***Figure 4—source data 1***). Interestingly, 75% of these also had regulated accessibility in development.

As for the development time course, we clustered accessibility changes in ageing. We identified eight clusters of promoters with similar accessibility changes across ageing and annotated them based on tissue biases in gene expression (***Figure 4B***; ***Figure 4—source data 1***). This defined one intestinal cluster (I), two clusters enriched for intestine or hypodermal biased expression (I + H) and five mixed clusters. Several mixed clusters show weak gene expression enrichments, such as intestine expression in Mix1-2 and neural expression in Mix3 (***Figure 4B***). As observed for the development clusters, enriched GO terms were consistent with gene expression biases (***Figure 4B***, ***Figure 4—figure supplement 3***).

We then evaluated the enrichment of transcription factors at each ageing promoter cluster. The binding of DAF-16/FoxO, a master regulator of ageing (***Lin et al., 2001***), is associated with five ageing promoter clusters (***Figure 5B***). Consistent with a prominent role in the intestine (***Figure 4B***; ***Kaplan and Baugh, 2016***), promoter clusters enriched for DAF-16 binding are also enriched for intestinal genes (***Figure 4B***). The binding enrichment patterns of five other TFs implicated in ageing (DVE-1, NHR-80, ELT-2, FOS-1 and PQM-1 (***Uno et al., 2013***; ***Folick et al., 2015***; ***Goudeau et al., 2011***; ***Mann et al., 2016***; ***Tian et al., 2016***; ***Mao et al., 2016***; ***Tepper et al., 2013***) are similar to DAF-16 (***Figure 5B***). These TFs and DAF-16 are also enriched in developmental intestine promoter clusters (***Figure 5A***), supporting cooperation between them in development and ageing. A group of hypodermal TFs including BLMP-1, ELT-1 and ELT-3 are found enriched at promoters in one of the two I + H ageing clusters (***Figure 5B***). Finally, CEBP-1 binding is enriched in clusters Mix3 and Mix4, which are characterized by a continuous increase of promoter accessibility across ageing. This suggests a potential role of CEBP-1 in activating a subset of genes during ageing, as it is the case for its homologue CEBP-β in mouse (***Sandhir and Berman, 2010***).

Conclusion

For the first time, we systematically map regulatory elements across the lifespan of an animal. We identified 42,245 accessible sites in *C. elegans* chromatin and functionally annotated them based on transcription patterns at the accessible site. This avoided the problems of histone-mark-based approaches for defining element function (***Core et al., 2014***; ***Henriques et al., 2018***; ***Rennie et al., 2018***). Our map identified promoters active across development and ageing, but we did not find promoters for every gene. Classes that would have been missed are those for genes expressed only in males or dauer larvae (which we did not profile) and genes not active under laboratory conditions. In addition, whole-animal profiling would miss promoters active in only a small number of cells. In

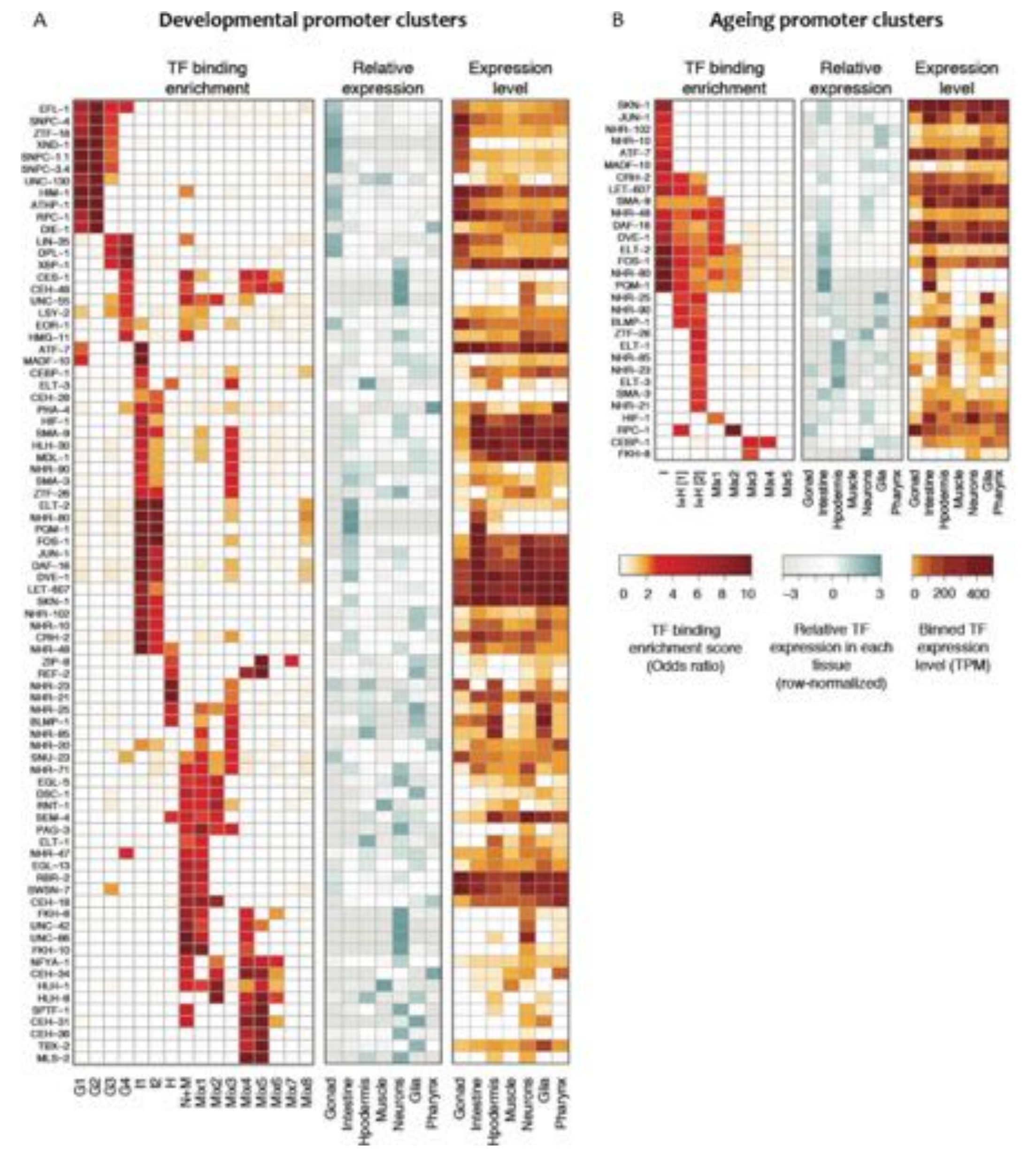


Figure 5. Transcription factor binding enrichment in developmental and ageing promoter clusters. Transcription factor (TF) binding enrichments in developmental (A) or ageing (B) promoter clusters from *Figure 4*. TF-binding data are from modENCODE/modERN (*Araya et al., 2014*; *Kudron et al., 2018*); peaks in HOT regions were excluded (see Materials and methods). Only TFs enriched more than twofold in at least one cluster are shown, and only enrichments with a p<0.01 (Fisher's exact test) are shown. Plots show TF binding enrichment odds ratio (left), expression of the TF

Figure 5 continued on next page

Figure 5 continued

in each tissue relative to its expression across all tissues (log2(TF tissue TPM/mean of the TF's TPMs across all tissues), middle), and the decile of expression of the TF in each tissue (right; TPMs < 1 are not taken into account when calculating TPMs deciles). Expression data are from ***Cao et al. (2017)***. Legends for Figure Supplements.

DOI: https://doi.org/10.7554/eLife.37344.022

The following source data is available for figure 5:

Source data 1. TF datasets used for analyses.

DOI: https://doi.org/10.7554/eLife.37344.023

the future, assaying accessible chromatin and nuclear transcription in specific cell types should identify many of these missed elements.

We found that accessibility of most elements changes during the life of the worm, supporting a key role played by chromatin structure. Despite the map being based on bulk profiling in whole animals, we find that regulatory elements with shared accessibility dynamics often share patterns of tissue-specific expression, GO annotation, and TF binding. The promoters with shared accessibility changes are therefore excellent starting points for studies of cell- and process-specific gene expression. In summary, our identification of regulatory elements across *C. elegans* life together with an initial characterization of their properties provides a key resource that will enable future studies of transcriptional regulation in development and ageing.

Materials and methods

Collection of developmental time series samples

Wild-type N2 were grown at 20˚C in liquid culture to the adult stage using standard S-basal medium with HB101 bacteria, animals bleached to obtain embryos, and the embryos hatched without food in M9 buffer for 24 hr at 20˚C to obtain synchronized starved L1 larvae. L1 larvae were grown in a further liquid culture at 20˚C to the desired stage, then collected, washed in M9, floated on sucrose, washed again in M9, then frozen into 'popcorn' by dripping embryo or worm slurry into liquid nitrogen. Popcorn were stored at −80˚C until use. Times of growth were L1 (4 hr), L2 (20 hr), L3 (30 hr), L4 (45 hr), young adults (60 hr). Mixed populations of embryos were collected by bleaching cultures of synchronized 1-day-old adults.

Collection of ageing time series samples

glp-1(e2144) were raised at 15˚C on standard NGM plates seeded with OP50 bacteria. Embryos were obtained by bleaching gravid adults and then approximately 6000 placed at 25˚C on 150 mm 2% NGM plates seeded with a 30X concentrated overnight culture of OP50. For harvest, worms were washed 3X in M9 and then worm slurry was frozen into popcorn by dripping into liquid nitrogen and stored at −80˚C. Harvest times after embryo plating were D1/YA (53 hr), D2 (71 hr), D6 (167 hr), D9 (239 hr), D13 (335 hr).

Nuclear isolation and ATAC-seq

Frozen embryos or worms (1–3 frozen popcorns) were broken by grinding in a mortar and pestle or smashing using a Biopulverizer, then the frozen powder was thawed in 10 ml Egg buffer (25 mM HEPES pH 7.3, 118 mM NaCl, 48 mM KCl, 2 mM CaCl2, 2 mM MgCl2). Ground worms were pelleted by spinning at 1500 g for 2 min, then resuspended in 10 ml working Buffer A (0.3M sucrose, 10 mM Tris pH 7.5, 10 mM MgCl2, 1 mM DTT, 0.5 mM spermidine 0.15 mM spermine, protease inhibitors (Roche complete, EDTA free) containing 0.025% IGEPAL CA-630. The sample was dounced 10X in a 14-ml stainless steel tissue grinder (VWR), then the sample spun 100 g for 6 min to pellet large fragments. The supernatant was kept and the pellet resuspended in a further 10 ml Buffer A, then dounced for 25 strokes. This was spun 100 g for 6 min to pellet debris and the supernatants, which contain the nuclei, were pooled, spun again at 100 g for 6 min to pellet debris, and transferred to a new tube. Nuclei were counted using a hemocytometer. One million nuclei were transferred to a 1.5-ml tube and spun 2000 g for 10 min to pellet. ATAC-seq was performed essentially as in ***Buenrostro et al. (2013)***. The supernatant was removed, the nuclei resuspended in 47.5 µl of tagmentation buffer, incubated for 30 min at 37˚C with 2.5 µl Tn5 enzyme (Illumina Nextera kit), and then tagmented DNA purified using a MinElute column (Qiagen) and converted into a library using the Nextera kit protocol. Typically, libraries were amplified using 12–16 PCR cycles. ATAC-seq was performed on two biological replicates for each developmental stage and each ageing time point.

DNAse I and MNase mapping

Replicate concentration courses of DNase I were performed for each stage as follows. Twenty million nuclei were digested in Roche DNAse I buffer for 10 min at 25˚C using 2.5, 5, 10, 25, 50, 100, 200, and 800 units/ml DNase I (Roche), then EDTA was added to stop the reactions. Embryo micrococcal nuclease (MNase) digestion concentration courses for embryos were made by digesting nuclei with 0.025, 0.05, 0.1, 0.25, 0.5, 1, 4, 8, or 16 units/ml MNase in 10 mM Tris pH 7.5, 10 mM MgCl2, 4 mM CaCl2 for 10 min at 37C. Reactions were stopped by the additon of EDTA. Following digestions, total DNA was isolated from the nuclei following proteinase K and RNase A digestion, then large fragments removed by binding to Agencourt AMPure XP beads (0.5 volumes). Small double cut fragments < 300 bp were isolated either using a Pippen prep gel (protocol 1) or using Agencourt AMPure XP beads (protocol 2). Libraries were prepared as described in the Sequencing library preparation section below.

Transcription initiation and nuclear RNA profiling

Nuclei were isolated and then chromatin associated RNA (development series) or nuclear RNA (ageing series) was isolated. Chromatin associated RNA was isolated as in (***Pandya-Jones and Black, 2009***), resuspending washed nuclei in Trizol for RNA extraction. To isolate nuclear RNA, nuclei were directly mixed with Trizol. Following purification, RNA was separated into fractions of 17–200nt and >200 nt using Zymo clean and concentrate columns. To profile transcription elongation ('long cap RNA-seq') in the nucleus, stranded libraries were prepared from the >200 nt RNA fraction using the NEB Next Ultra Directional RNA Library Prep Kit (#E7420S). Libraries were made from two biological replicates for each developmental stage and each ageing time point. To profile transcription initiation ('short cap RNA-seq'), stranded libraries were prepared from the 17–200nt RNA fraction. Non-capped RNA was degraded by first converting uncapped RNAs into 5'-monophosphorylated RNAs using RNA polyphosphatase (Epibio), then treating with 5' Terminator nuclease (Epibio). The RNA was treated with calf intestinal phosphatase to remove 5' phosphates from undegraded RNA, and decapped using Tobacco Acid Pyrophosphatase (Epicentre), Cap-Clip Acid Pyrophosphatase (CellScript, for one L2 and one L3 replicate) or Decapping Pyrophosphohydrolase (Dpph tebu-bio, for one L3 replicate) and then converted into sequencing libraries using the Illumina TruSeq Small RNA Preparation Kit kit. Libraries were size selected to be 145–225 bp long on a 6% acrylamide gel, giving inserts of 20–100 bp long. Libraries were made from two biological replicates for each developmental stage. During the course of this work, the TAP enzyme stopped being available; the Cap-Clip and Dpph enzymes perform less well than TAP. One L3 and one YA replicate was made using a slightly different protocol. Embryo short cap RNA-seq data from ***Chen et al. (2013)*** was also included in the analyses (GSE42819).

ChIP-seq

Balls of frozen embryos or worms were ground to a powder using a mortar and pestle or a Retch Mixer Mill to break animals into pieces. Frozen powder was thawed into 1% formaldehyde in PBS, incubated 10 min, then quenched with 0.125M glycine. Fixed tissue was washed 2X with PBS with protease inhibitors (Roche EDTA-free protease inhibitor cocktail tablets 05056489001), once in FA buffer (50 mM Hepes pH7.5, 1 mM EDTA, 1% TritonX-100, 0.1% sodium deoxycholate, and 150 mM NaCl) with protease inhibitors (FA+), then resuspended in 1 ml FA +buffer per 1 ml of ground worm powder and the extract sonicated to an average size of 200 base pairs with a Diagenode Bioruptor or Bioruptor Pico for 25 pulses of 30 s followed by 30 s pause. For ChIP, 500 ug protein extract was incubated 2 ug antibody in FA +buffer with protease inhibitors overnight at 4˚C, then incubated with magnetic beads conjugated to secondary antibodies for 2 hr at 4˚C. Magnetic beads bound to immunoprecipitate were washed at room temperature twice in FA+, then once each in FA with 0.5M NaCl, FA with 1M NaCl, 0.25M LiCl (containing 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris pH8) and finally twice with TE pH8. Immunoprecipitated DNA was then eluted twice with

1% SDS, 250 mM NaCl, 10 mM Tris pH8, 1 mM EDTA at 65˚C. Eluted DNA was treated with RNase for 1 hr at 37C and crosslinks reversed by overnight incubation at 65˚C with 200 ug/ml proteinase K, and the DNA purified using a Qiagen column. Libraries were prepared as described in the Sequencing library preparation section below. Two biological replicate ChIPs were conducted for each histone modification at each developmental time point (Embryo, L1, L2, L3, L4, YA). Antibodies used were: anti-H3K4me3 (Abcam ab8580), anti-H3K4me1 (Abcam ab8895), anti-H3K36me3 (Abcam ab9050), and anti-H3K27me3 (Wako 309–95259).

Sequencing library preparation

DNA was converted into sequencing libraries using a modified Illumina Truseq protocol based on https://ethanomics.files.wordpress.com/2012/09/chip_truseq.pdf. Briefly DNA fragments are first repaired with an End repair enzyme mix (New England Biolabs, cat E5060) for 30 min at 20C in 50 µl, then all DNA fragments were recovered using 1 vol of AMPure XP beads and 1 vol of 30% PEG_{8000} in 1.25M NaCl, and eluted in 16.5 µl of H_2O. The DNA was 3' A-tailed in 1X NEB buffer 2 using 2.5 units of Klenow 3' to 5' exo(minus) (New England Biolabs, cat M0212) and 0.2 mM ATP for 30 min at 37C in 20 µl. Illumina Truseq adaptors were then directly ligated to the DNA fragments by adding 25 µl 2X buffer, 1 µl of 0.06 nM adaptors (1 µl of 1:250 dilution of Illumina stock solution), 2.5 µl water and 1.5 µl of NEB Quick ligase (cat M2200). After 20 min at room temperature, 5 µl of 0.5M EDTA pH8 was added to inactivate the enzyme and DNA was purified using AMPure XP beads. For DNAse and MNase libraries, 1.3 volumes of beads were used; for ChIP libraries, 0.9 volumes of beads were used. DNA fragments were eluted in 20 µl of H_2O. We used 1 µl to determine the number of cycles needed to get amplification to 50% of the plateau as in https://ethanomics.wordpress.com/ngs-pcr-cycle-quantitation-protocol/. Libraries were amplified by PCR by adding 20 µl of the KAPA Hifi Hotstart Ready Mix (Kapabiosystem cat KK2601) and 1 µl of 25 uM Illumina Universal primers. Libraries were then size selected. DNAse and MNase libraries were purified using 1.3 volumes of beads. For ChIP libraries, 0.7 volumes of beads were added to bind large DNA. Beads were discarded and DNA recovered from the supernatant by adding 0.75 volumes of beads and 0.75 volumes of 30% PEG_{8000} in 1.25M NaCl. DNA was eluted in 40 µl water and 0.8 volumes of beads used to bind the library, leaving adaptor dimers in the supernatant. DNA was eluted in 10–15 µl water, quantified using a Qubit, and analyzed using a Agilent Tapestation.

Data processing

Reads were aligned using bwa-backtrack (***Li and Durbin, 2009***) in single-end (ATAC-seq, short cap RNA-seq, ChIP-seq) or paired-end mode (ATAC-seq - developmental only, DNase-seq, MNase-seq, long cap RNA-seq). Low-quality ($q < 10$), mitochondrial and modENCODE-blacklisted (***Boyle et al., 2014***) reads were discarded at this point.

For ATAC-seq, normalized genome-wide accessibility profiles from single-end reads were then calculated with MACS2 (***Zhang et al., 2008***) using the parameters –format BAM –bdg –SPMR –gsize ce –nolambda –nomodel –extsize 150 –shift -75 –keep-dup all. Developmental ATAC-seq was also processed in paired-end mode (ATAC-seq libraries of ageing samples were single-end). We did not observe major differences between accessible sites identified from paired-end, and single-end profiles, and therefore use single-end profiles throughout the study for consistency.

Short-cap and long-cap data were processed essentially as in ***Chen et al. (2013)***. Following alignment, and filtering, transcription initiation was represented using strand-specific coverage of 5' ends of short-cap reads. Transcription elongation was represented as strand-specific coverage of long-cap reads, with regions between read pairs filled in. For browsing, transcription elongation signal was normalized between samples by sizeFactors calculated from gene-level read counts using DESeq2 (***Love et al., 2014***). Normalized (linear) coverage signal was then further log-transformed with $log_2(normalised\_coverage + 1)$.

ChIP-seq data was processed as in ***Chen et al. (2014)***. After alignment and filtering, the BEADS algorithm was used to generate normalized ChIP-seq coverage tracks (***Cheung et al., 2011***).

Stage-specific tracks used in downstream analyses were obtained by averaging normalized signal across two biological replicates. Manipulations of genome-wide signal were performed using bedtools (***Quinlan and Hall, 2010***), UCSC utilities (***Kent et al., 2010***), and wiggleTools (***Zerbino et al., 2014***). Computationally intensive steps were managed and parallelized using snakemake

(**Köster and Rahmann, 2012**). Genome-wide data was visualized using the Integrative Genomics Viewer (**Robinson et al., 2011**; **Thorvaldsdóttir et al., 2013**).

To assess the reproducibility of replicate datasets, we performed PCA using the plotPCA() function in DESeq2 (**Love et al., 2014**) on peak accessibility at promoters (ATAC-seq), read counts at annotated genes (long cap RNA-seq), 5′ end read counts at promoters (short cap RNA-seq), and genic regions, from the most upstream promoter to the annotated 3′ end, excluding genes with no annotated promoter (histone modifications). Replicates agreed well as shown in **Figure 1—figure supplements 2** and **3**.

Identification of accessible sites

Accessible sites were identified as follows. We first identified concave regions (regions with negative smoothed second derivative) from ATAC-seq coverage averaged across all stages and replicates. This approach is extremely sensitive, identifying a large number (>200,000) of peak-like regions. We then scored all peaks in each sample using the magnitude of the sample-specific smoothed second derivative. We used IDR (**Li et al., 2011**) on the scores to assess stage-specific signal levels and biological reproducibility, setting a conservative cutoff at 0.001. Final peaks boundaries were set to peak accessibility extended by 75 bp on both sides. We found that calling peaks using paired end or single end data were highly similar, but some regions were captured better by one or the other. Developmental ATAC-seq datasets were sequenced paired-end and ageing datasets single-end. Peaks were therefore called separately using developmental paired-end data, developmental single-end data extended to 150 bp and shifted 75 bp upstream, and ageing (single-end only) data, and then merged. This was achieved by successively including peaks from the three sets if they did not overlap a peak already identified in an earlier set. **Figure 1—source data 1** gives peak calls and ATAC peak heights at each stage.

Datasets and genome versions

Throughout this study, we used the WBcel215/ce10 (WS220) version of the *C. elegans* genome, and WormBase WS260 genome annotations - with coordinates backlifted to WBcel215/ce10 (WS220). For convenience, **Figure 2—source data 1** also contains WBcel235/ce11 coordinates of accessible sites and representative transcription initiation modes.

For motif analyses, Inr and TATA consensus sequences were obtained from **Sloutskin et al. (2015)**, and mapped with zero mismatches using homer (**Heinz et al., 2010**). CpG density was defined as in **Chen et al. (2014)**.

modENCODE (**Araya et al., 2014**) and modERN (**Kudron et al., 2018**) transcription factor binding datasets used in this paper were obtained from http://www.encodeproject.org or http://data.modencode.org (EOR-1). ChIP-seq profiles were manually inspected and 227 high quality datasets selected, covering 176 transcription factors (given in **Figure 5—source data 1**). To define TFBS clusters (**Figure 1—figure supplement 1C,D**; **Figure 2—figure supplement 1**), TF peak calls were extended to 200 bp on either side of the summit, and clustered using a single-linkage approach. To analyze enrichment of individual factors (**Figure 5**), TF peaks were assigned to a regulatory element if their summits overlapped with the 400 bp region centered at the element midpoint. Factors associated with each regulatory element via this approach are given in **Figure 4—source data 1**. We excluded binding at so-called 'HOT' (highly occupied target) regions from enrichment analyses in **Figure 5**, as these are thought to represent non-sequence-specific TF binding or ChIP artifacts (**Gerstein et al., 2010**; **Kudron et al., 2018**). HOT regions were defined here as accessible sites with binding of 19 or more of the analyzed 176 TFs (sites in the top 20% of binding, excluding sites with no binding).

Coefficients of variation of gene expression (CV) are from (**Gerstein et al., 2014**); processed table was kindly provided by Burak Alver).

Annotation of regulatory elements

Patterns of nuclear transcription were used to annotate elements. At each stage, separately on both strands, we assessed 1) initiating and elongating transcription at the site, 2) continuity of transcription from the site to the closest downstream gene, and 3) positioning of nearby exons (on the matching strand).

To assess for transcription elongation at an accessible site, we counted 5′ ends of long cap reads upstream (−250:−75), and downstream (+75:+250) of peak accessibility. We then used two approaches to identify sites with a local increase in transcription elongation. First, we used DESeq2 to test for an increase in downstream vs upstream counts ('jump' method). Statistical significance was called at log2FoldChange > 1.5, and adjusted p-value<0.1 (one-sided test). To capture additional regions with weak signal ('incr' method), we accepted sites with 0 reads upstream, at least one read in both biological replicates downstream, and three reads total when summed across both biological replicates.

To assess transcription initiation, we pooled short cap across all six wild-type stages, and included two additional embryo replicates from ***Chen et al. (2013)***. The pooled signal was filtered for reproducibility by only keeping signal at base pairs with non-zero transcription initiation in at least two replicates. We then required the presence of at least one base pair with reproducible signal within 125 bp of peak accessibility to designate an accessible site as having transcription initiation. For every site, we also defined a representative transcription initiation mode as the position with maximum short-cap signal within 125 bp of peak accessibility. For sites without reproducible short-cap signal, we used an extrapolated, 'best-guess' position at 60 bp downstream of peak accessibility.

We annotated accessible sites as coding_promoter or pseudogene_promoter if they fulfilled the following four criteria. (1) The accessible site had transcription initiation, and passed at least one of the elongation tests (jump or incr), or passed both elongation tests (jump and incr). (2) Transcription initiation mode at the accessible site was either upstream of the closest first exon, or, in the presence of a UTR, up to 250 bp downstream within the UTR. (The closest first exon was chosen based on the distance between the 5′ end of the first exon and peak accessibility at the accessible site, allowing the 5′ end of the exon to be up to 250 bp upstream or anywhere downstream of peak accessibility). (3) The region from peak accessibility to the closest first exon did not contain the 5′ end of a non-first exon. (4) Distal sites (peak accessibility >250 bp from the closest first exon) were additionally required to (a) have continuous long-cap coverage from 250 bp downstream of peak accessibility to the closest first exon, and (b) be further than 250 bp away from any non-first exon.

We then further attempted to assign a single, lower-confidence promoter to genes that were not assigned a promoter so far. For every gene without promoter assignments, we re-examined sites that fulfilled criteria (2-4), and were either intergenic, or within 250 bp of the closest first exon. We then annotated the site with the largest jump test log2FoldChange as the promoter, if it was also larger than 1.

Next, sites within 250 bp of the 5′ end of an annotated tRNA, snRNA, snoRNA, miRNA, or rRNA were annotated as non-coding_RNA. Intergenic sites more than 250 bp away from annotated exons that had initiating transcription, and passed the jump test were annotated as unassigned_promoter. All remaining sites were annotated as transcription_initiation or no_transcription based on whether they had transcription initiation.

Elements were then annotated on each strand based on aggregating transcription patterns across stages by determining the 'highest' annotation using the ranking of: coding_promoter, pseudogene_promoter, non-coding_RNA, unassigned_promoter, transcription_initiation, no_transcription. Element type and coloring was then defined using the following ranking: coding_promoter on either strand => coding_promoter (red); pseudogene_promoter on either strand => pseudogene_promoter (orange); non-coding_RNA on either strand => non-coding_RNA (black); unassigned_promoter on either strand => unassigned_promoter (yellow); transcription_initiation on either strand => putative_enhancer (green); all remaining sites => other_element (blue). ***Figure 2—source data 1*** gives annotation information.

Clustering of promoter accessibility

Accessible elements with regulated accessibility were determined as follows. All elements (n = 42,245) were tested for a difference in ATAC-seq coverage between any two developmental time points or between any two ageing time points using DESeq2 (***Love et al., 2014***). Sites with >= 2 absolute fold change and adjusted p-value<0.01 were defined as 'regulated' (n = 30,032 in development and n = 6590 in ageing; ***Figure 4—source data 1***); regulated promoters (n = 10,199 in development and n = 1800 in ageing) were used in clustering analyses.

For clustering analyses, depth-normalized ATAC-seq coverage of each promoter was calculated at each time point in development or ageing. Relative accessibility was calculated at each time point in development or ageing by applying the following formula: $log_2\left(ATACseq\ coverage_{time\ point\ i} + 1\right) - log_2(mean\ ATACseq\ coverage\ across\ time\ points + 1)$. Mean ATAC-seq coverage across time points was calculated separately for the developmental and ageing time courses. Clustering was performed using *k*-medoids, as implemented in the pam() method of the cluster R package (***Maechler et al., 2017***). Different numbers of clusters were tested for clustering of regulatory elements in developmental and ageing datasets; 16 was chosen for developmental data and 10 for ageing data as the normalized changes in promoter ATAC-seq signals within each cluster were relatively homogeneous. We manually merged two ageing clusters showing comparable accessibility and tissue-specific gene enrichment (resulting in the cluster I + H [2]). Clusters labels were determined based on enrichment for tissue-biased gene expression within each cluster (see below).

To compare accessibility and gene expression, FPM-normalized gene-level read counts were calculated using DESeq2, and then averaged across biological replicates. For visualisation, relative expression levels were calculated using the approach described above for relative promoter accessibility (see formula above), with FPM values instead of ATAC-seq coverage values.

Using single-cell RNA-seq data from ***Cao et al. (2017)***, we defined tissue-biased gene expression as follows: Gene expression was considered enriched in a given tissue if it had a fold-change >= 3 between expression in the tissues with highest and second highest levels and an adjusted p-value<0.01. This defined 5315 genes with tissue-biased expression (1432 in Gonad, 553 in Hypodermis, 799 in Intestine, 352 in Muscle, 1218 in Neurons, 447 enriched in Glia, 514 in Pharynx). For each developmental or ageing cluster of promoters, we calculated the percentage of genes with biased expression in a given tissue relative to the total number of genes in the cluster. These values were plotted in ***Figure 4A and B*** (bar plots).

GO enrichments were evaluated using the R package gProfileR (***Reimand et al., 2016***) against *C. elegans* GO database. Significant enrichment was set at an adjusted p-value of 0.05, and hierarchically redundant terms were automatically removed by gProfileR.

Enrichment for transcription factor binding in promoter clusters

Prior to analysis of TF peak enrichment at annotated promoters, accessible elements considered 'HOT' (see above) were removed, resulting in 10,086 to be assessed by enrichment analysis. Only transcription factors with more than 200 peaks overlapping 'non-hot' regulatory elements were kept, to ensure sufficient data for analysis. Following this stringent filtering, 89 transcription factors could be assayed for binding enrichment. Transcription factor binding enrichment in each cluster was estimated using the odds ratio and enrichments with an associated p-value<0.01 (Fisher's exact test) were kept. Transcription factors which did not show enrichment higher than two in any cluster were discarded. ***Figure 5*** summarizes the transcription factor binding enrichment in each cluster during development or ageing. Relative tissue expression profiles of each transcription factor at the L2 stage (data from ***Cao et al., 2017***) was calculated in each tissue by taking the log2 of its expression (TPM) in the tissue divided by its mean expression across all tissues. A pseudo-value of 0.1 was first added to all the TPM values before calculation of the relative levels of expression.

Construction of transgenic lines

Transgene constructs were made using three-site Gateway cloning (Invitrogen) as in ***Chen et al. (2014)***. Site one has the regulatory element sequence to be tested, site two has a synthetic outron (OU141; ***Conrad et al., 1995***) fused to *his-58* (plasmid pJA357), and site three has gfp-tbb-2 3'UTR (pJA256; ***Zeiser et al., 2011***) in the MosSCI compatible vector pCFJ150, which targets Mos site Mos1(ttTi5605); MosSCI lines were generated as described (***Frøkjaer-Jensen et al., 2008***).

Data access

ATAC-seq, ChIP-seq, DNase/MNase-seq, long/short cap RNA-seq data from this study, including processed tracks are available at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE114494.

Acknowledgements

We thank C Bradshaw for bioinformatics support, K Harnish for sequencing, B Alver for providing processed data, and F Carelli, C Gal, and A Frapporti for comments on the manuscript. The work was supported by Wellcome Trust Senior Research Fellowships to JA (054523 and 101863), a Wellcome Trust PhD fellowship to JJ (097679), a Sir Robert Edwards Scholarship from Churchill College, an English Speaking Union Graduate Scholarship, and funding from the Cambridge Trust to MS, a Medical Research Council DTP studentship to JS, and a Thouron award to CW. This study was also supported by the European Sequencing and Genotyping Infrastructure (funded by the EC, FP7/2007-2013) under Grant Agreement 26205 (ESGI) as part of the transnational access program. We thank Drs. Hans Lehrach and Marie-Laure Yaspo for generous support of the ESGI project, Dr. Marc Sultan for setting up sequencing technology platforms, and Mathias Linser and the rest of the sequencing team of the Department of Vertebrate Genomics at the Max Planck Institute for Molecular Genetics for technical assistance. We also acknowledge core support from the Wellcome Trust (092096) and Cancer Research UK (C6946/A14492).

Additional information

Competing interests

Julie Ahringer: Reviewing editor, *eLife*. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Wellcome	101863	Jürgen Jänes Yan Dong Alex Appert Chiara Cerrato Ron Chen Carolina Gemma Ni Huang Przemyslaw Stempor Annette Steward Eva Zeiser Julie Ahringer
Medical Research Council		Jacques Serizay
European Commission	FP7/2007-2013	Sascha Sauer Julie Ahringer
Wellcome	097679	Jürgen Jänes

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Jürgen Jänes, Conceptualization, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Yan Dong, Investigation, Methodology, Writing—review and editing; Michael Schoof, Alex Appert, Formal analysis, Investigation, Methodology; Jacques Serizay, Software, Formal analysis, Visualization, Writing—original draft, Writing—review and editing; Chiara Cerrato, Carson Woodbury, Ron Chen, Carolina Gemma, Djem Kissiov, Annette Steward, Eva Zeiser, Formal analysis, Investigation; Ni Huang, Software, Formal analysis; Przemyslaw Stempor, Data curation, Software, Formal analysis; Sascha Sauer, Funding acquisition, Project administration; Julie Ahringer, Conceptualization, Formal analysis, Supervision, Funding acquisition, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Jürgen Jänes https://orcid.org/0000-0002-2540-1236
Ni Huang https://orcid.org/0000-0001-8849-038X
Przemyslaw Stempor https://orcid.org/0000-0002-9464-7475
Julie Ahringer https://orcid.org/0000-0002-7074-4051

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.37344.038
Author response https://doi.org/10.7554/eLife.37344.039

Additional files

Supplementary files

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.37344.024

Data availability

Sequencing data have been deposited in as a SuperSeries in GEO under accession code GSE114494.

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Julie Ahringer, Jürgen Jänes	2018	Chromatin accessibility dynamics across C. elegans development and ageing [DNase, MNase]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114481	Gene Expression Omnibus, GSE114481
Ahringer J, Jürgen Jänes	2018	Chromatin accessibility dynamics across C. elegans development and ageing [scap]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114490	NCBI Gene Expression Omnibus, GSE114490
Julie Ahringer, Jürgen Jänes	2018	Chromatin accessibility dynamics across C. elegans development and ageing [lcap]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114483	Gene Expression Omnibus, GSE114483
Julie Ahringer, Jürgen Jänes	2018	Chromatin accessibility dynamics across C. elegans development and ageing [ChIP-seq]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114440	Gene Expression Omnibus, GSE114440
Julie Ahringer, Jürgen Jänes	2018	Chromatin accessibility dynamics across C. elegans development and ageing [ATAC-seq]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114439	Gene Expression Omnibus, GSE114439

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Down TA	2013	The landscape of RNA polymerase II transcription initiation in C. elegans reveals a novel regulatory architecture	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42819	NCBI Gene Expression Omnibus, GSE42819

References

Allen MA, Hillier LW, Waterston RH, Blumenthal T. 2011. A global analysis of C. elegans trans-splicing. *Genome research* **21**:255–264. DOI: https://doi.org/10.1101/gr.113811.110, PMID: 21177958

Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, Sandelin A. 2014. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Communications* **5**:5336. DOI: https://doi.org/10.1038/ncomms6336, PMID: 25387874

Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends in Genetics* **31**:426–433. DOI: https://doi.org/10.1016/j.tig.2015.05.007, PMID: 26073855

Andersson R. 2015. Promoter or enhancer, what's the difference? deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37**:314–323. DOI: https://doi.org/10.1002/bies.201400162, PMID: 25450156

Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D, Niu W, Boyle AP, Xie D, Ma L, Murray JI, Reinke V, Waterston RH, Snyder M. 2014. Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature* **512**:400–405. DOI: https://doi.org/10.1038/nature13497, PMID: 25164749

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**:453–456. DOI: https://doi.org/10.1038/nature13668, PMID: 25164757

Brabin C, Appleford PJ, Woollard A. 2011. The Caenorhabditis elegans GATA factor ELT-1 works through the cell proliferation regulator BRO-1 and the fusogen EFF-1 to maintain the seam stem-like fate. *PLoS Genetics* **7**: e1002200. DOI: https://doi.org/10.1371/journal.pgen.1002200, PMID: 21829390

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**:1213–1218. DOI: https://doi.org/10.1038/nmeth.2688, PMID: 24097267

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**:661–667. DOI: https://doi.org/10.1126/science.aam8940, PMID: 28818938

Chen RA, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, Ahringer J. 2013. The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. *Genome Research* **23**:1339–1347. DOI: https://doi.org/10.1101/gr.153668.112, PMID: 23550086

Chen RA, Stempor P, Down TA, Zeiser E, Feuer SK, Ahringer J. 2014. Extreme HOT regions are CpG-dense promoters in C. Elegans and humans. *Genome Research* **24**:1138–1146. DOI: https://doi.org/10.1101/gr.161992.113, PMID: 24653213

Cheung MS, Down TA, Latorre I, Ahringer J. 2011. Systematic Bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research* **39**:e103. DOI: https://doi.org/10.1093/nar/gkr425, PMID: 21646344

Conrad R, Lea K, Blumenthal T. 1995. SL1 trans-splicing specified by AU-rich synthetic RNA inserted at the 5′ end of Caenorhabditis elegans pre-mRNA. *RNA* **1**:164–170. PMID: 7585246

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**:1845–1848. DOI: https://doi.org/10.1126/science.1162228, PMID: 19056941

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics* **46**:1311–1320. DOI: https://doi.org/10.1038/ng.3142, PMID: 25383968

Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods* **3**:503–509. DOI: https://doi.org/10.1038/nmeth888, PMID: 16791207

Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. 2017. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Research* **27**:2096–2107. DOI: https://doi.org/10.1101/gr.226233.117, PMID: 29141961

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biology* **8**:e1000384. DOI: https://doi.org/10.1371/journal.pbio.1000384, PMID: 20485488

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology* **28**:817–825. DOI: https://doi.org/10.1038/nbt.1662, PMID: 20657582

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**:43–49. DOI: https://doi.org/10.1038/nature09906, PMID: 21441907

Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J. 2016. Stable *Caenorhabditis elegans* chromatin domains separate broadly expressed and developmentally regulated genes. *PNAS* **113**. DOI: https://doi.org/10.1073/pnas.1608162113, PMID: 27791097

Flynn RA, Almada AE, Zamudio JR, Sharp PA. 2011. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *PNAS* **108**:10460–10465. DOI: https://doi.org/10.1073/pnas.1106630108, PMID: 21670248

Folick A, Oakley HD, Yu Y, Armstrong EH, Kumari M, Sanor L, Moore DD, Ortlund EA, Zechner R, Wang MC. 2015. Aging. Lysosomal signaling molecules regulate longevity in Caenorhabditis elegans. *Science* **347**:83–86. DOI: https://doi.org/10.1126/science.1258857, PMID: 25554789

Frøkjaer-Jensen C, Davis MW, Hopkins CE, Newman BJ, Thummel JM, Olesen SP, Grunnet M, Jorgensen EM. 2008. Single-copy insertion of transgenes in Caenorhabditis elegans. *Nature Genetics* **40**:1375–1383. DOI: https://doi.org/10.1038/ng.248, PMID: 18953339

Fukushige T, Hawkins MG, McGhee JD. 1998. The GATA-factor elt-2 is essential for formation of the Caenorhabditis elegans intestine. *Developmental Biology* **198**:286–302. DOI: https://doi.org/10.1016/S0012-1606(98)80006-7, PMID: 9659934

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**:1775–1787. DOI: https://doi.org/10.1126/science.1196914, PMID: 21177976

Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, et al. 2014. Comparative analysis of the transcriptome across distant species. *Nature* **512**:445–448. DOI: https://doi.org/10.1038/nature13424, PMID: 25164755

Gilleard JS, Shafi Y, Barry JD, McGhee JD. 1999. ELT-3: a Caenorhabditis elegans GATA factor expressed in the embryonic epidermis during morphogenesis. *Developmental Biology* **208**:265–280. DOI: https://doi.org/10.1006/dbio.1999.9202, PMID: 10191044

Gissendanner CR, Sluder AE. 2000. nhr-25, the Caenorhabditis elegans ortholog of ftz-f1, is required for epidermal and somatic gonad development. *Developmental Biology* **221**:259–272. DOI: https://doi.org/10.1006/dbio.2000.9679, PMID: 10772806

Goudeau J, Bellemin S, Toselli-Mollereau E, Shamalnasab M, Chen Y, Aguilaniu H. 2011. Fatty acid desaturation links germ cell loss to longevity through NHR-80/HNF4 in C. elegans. *PLoS Biology* **9**:e1000599. DOI: https://doi.org/10.1371/journal.pbio.1000599, PMID: 21423649

Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell* **151**:1488–1500. DOI: https://doi.org/10.1016/j.cell.2012.11.023, PMID: 23260138

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**:311–318. DOI: https://doi.org/10.1038/ng1966, PMID: 17277777

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**:108–112. DOI: https://doi.org/10.1038/nature07829, PMID: 19295514

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**:576–589. DOI: https://doi.org/10.1016/j.molcel.2010.05.004, PMID: 20513432

Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. 2018. Widespread transcriptional pausing and elongation control at enhancers. *Genes & Development* **32**:26–41. DOI: https://doi.org/10.1101/gad.309351.117, PMID: 29378787

Ho MCW, Quintero-Cadena P, Sternberg PW. 2017. Genome-wide discovery of active regulatory elements and transcription factor footprints in *Caenorhabditis elegans* using DNase-seq. *Genome Research* **27**. DOI: https://doi.org/10.1101/gr.223735.117, PMID: 29074739

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research* **41**:827–841. DOI: https://doi.org/10.1093/nar/gks1284, PMID: 23221638

Horn M, Geisen C, Cermak L, Becker B, Nakamura S, Klein C, Pagano M, Antebi A. 2014. DRE-1/FBXO11-dependent degradation of BLMP-1/BLIMP-1 governs C. elegans developmental timing and maturation. *Developmental Cell* **28**:697–710. DOI: https://doi.org/10.1016/j.devcel.2014.01.028, PMID: 24613396

Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, Fang L, Halfnight E, Lee D, Lin J, Lorch A, McKay S, Okada HM, Pan J, Schulz AK, Tu D, Wong K, Zhao Z, Alexeyenko A, Burglin T, Sonnhammer E, et al. 2007. High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. *PLoS Biology* **5**:e237. DOI: https://doi.org/10.1371/journal.pbio.0050237, PMID: 17850180

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research* **27**:38–52. DOI: https://doi.org/10.1101/gr.212092.116, PMID: 27831498

Kaplan RE, Baugh LR. 2016. L1 arrest, daf-16/FoxO and nonautonomous control of post-embryonic development. *Worm* **5**:e1175196. DOI: https://doi.org/10.1080/21624054.2016.1175196, PMID: 27383290

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**:2204–2207. DOI: https://doi.org/10.1093/bioinformatics/btq351, PMID: 20639541

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**:480–485. DOI: https://doi.org/10.1038/nature09725, PMID: 21179089

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**:182–187. DOI: https://doi.org/10.1038/nature09033, PMID: 20393465

Kim TK, Shiekhattar R. 2015. Architectural and functional commonalities between enhancers and promoters. *Cell* **162**:948–959. DOI: https://doi.org/10.1016/j.cell.2015.08.008, PMID: 26317464

Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, Eick D, Gut I, Ferrier P, Andrau JC. 2011. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Structural & Molecular Biology* **18**:956–963. DOI: https://doi.org/10.1038/nsmb.2085, PMID: 21765417

Köster J, Rahmann S. 2012. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* **28**:2520–2522. DOI: https://doi.org/10.1093/bioinformatics/bts480, PMID: 22908215

Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D, Brown JM, Gray NE, Collavin L, Gibbons RJ, Flint J, Taylor S, Buckle VJ, Milne TA, Wood WG, Higgs DR. 2012. Intragenic enhancers act as alternative promoters. *Molecular cell* **45**:447–458. DOI: https://doi.org/10.1016/j.molcel.2011.12.021, PMID: 22264824

Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. 2013. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *eLife* **2**:e00808. DOI: https://doi.org/10.7554/eLife.00808, PMID: 23795297

Kudron MM, Victorsen A, Gevirtzman L, Hillier LW, Fisher WW, Vafeados D, Kirkey M, Hammonds AS, Gersch J, Ammouri H, Wall ML, Moran J, Steffen D, Szynkarek M, Seabrook-Sturgis S, Jameel N, Kadaba M, Patton J, Terrell R, Corson M, et al. 2018. The ModERN resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* Transcription Factors. *Genetics* **208**:937–949. DOI: https://doi.org/10.1534/genetics.117.300657, PMID: 29284660

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**:317–330. DOI: https://doi.org/10.1038/nature14248, PMID: 25693563

Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y, Xie W, Yue F, Hariharan M, Ray P, Kuan S, Edsall L, Yang H, Chi NC, Zhang MQ, Ecker JR, et al. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**:350–354. DOI: https://doi.org/10.1038/nature14217, PMID: 25693566

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI: https://doi.org/10.1093/bioinformatics/btp324, PMID: 19451168

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**:1752–1779. DOI: https://doi.org/10.1214/11-AOAS466

Lin K, Hsin H, Libina N, Kenyon C. 2001. Regulation of the Caenorhabditis elegans longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nature Genetics* **28**:139–145. DOI: https://doi.org/10.1038/88850, PMID: 11381260

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:550. DOI: https://doi.org/10.1186/s13059-014-0550-8, PMID: 25516281

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2017. *Cluster: Cluster Analysis Basics and Extensions.* Scientific Research Publisher.

Mann FG, Van Nostrand EL, Friedland AE, Liu X, Kim SK. 2016. Deactivation of the GATA Transcription Factor ELT-2 Is a Major Driver of Normal Aging in C. elegans. *PLoS Genetics* **12**:e1005956. DOI: https://doi.org/10.1371/journal.pgen.1005956, PMID: 27070429

Mao XR, Kaufman DM, Crowder CM. 2016. Nicotinamide mononucleotide adenylyltransferase promotes hypoxic survival by activating the mitochondrial unfolded protein response. *Cell Death & Disease* **7**:e2113. DOI: https://doi.org/10.1038/cddis.2016.5, PMID: 26913604

Merritt C, Rasoloson D, Ko D, Seydoux G. 2008. 3' UTRs are the primary regulators of gene expression in the C. elegans germline. *Current Biology* **18**:1476–1482. DOI: https://doi.org/10.1016/j.cub.2008.08.013, PMID: 18818082

Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, Furlong EEM. 2018. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & Development* **32**:42–57. DOI: https://doi.org/10.1101/gad.308619.117, PMID: 29378788

Nguyen TA, Jones RD, Snavely AR, Pfenning AR, Kirchner R, Hemberg M, Gray JM. 2016. High-throughput functional comparison of promoter and enhancer activities. *Genome Research* **26**:1023–1033. DOI: https://doi.org/10.1101/gr.204834.116, PMID: 27311442

Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**:1896–1908. DOI: https://doi.org/10.1261/rna.1714509, PMID: 19656867

Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, Imbert J, Andrau JC, Ferrier P, Spicuglia S. 2011. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO Journal* **30**:4198–4210. DOI: https://doi.org/10.1038/emboj.2011.295, PMID: 21847099

Pérez-Lluch S, Blanco E, Tilgner H, Curado J, Ruiz-Romero M, Corominas M, Guigó R. 2015. Absence of canonical marks of active chromatin in developmentally regulated genes. *Nature Genetics* **47**:1158–1167. DOI: https://doi.org/10.1038/ng.3381, PMID: 26280901

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**:1851–1854. DOI: https://doi.org/10.1126/science.1164096, PMID: 19056938

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033, PMID: 20110278

Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J. 2016. G:profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research* **44**:W83–W89. DOI: https://doi.org/10.1093/nar/gkw199, PMID: 27098042

Rennie S, Dalby M, Lloret-Llinares M, Bakoulis S, Vaagenso CD, Jensen TH, Andersson R. 2017. Transcription start site analysis reveals widespread divergent transcription in D. Melanogaster and core promoter encoded enhancer activities. *bioRxiv*. https://www.biorxiv.org/content/early/2017/11/18/221952.

Rennie S, Dalby M, Lloret-Llinares M, Bakoulis S, Dalager Vaagensø C, Heick Jensen T, Andersson R. 2018. Transcription start site analysis reveals widespread divergent transcription in D. Melanogaster and core promoter-encoded enhancer activities. *Nucleic Acids Research* **46**:5455–5469. DOI: https://doi.org/10.1093/nar/gky244, PMID: 29659982

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature biotechnology* **29**:24–26. DOI: https://doi.org/10.1038/nbt.1754, PMID: 21221095

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods* **3**:511–518. DOI: https://doi.org/10.1038/nmeth890, PMID: 16791208

Saito TL, Hashimoto S, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, Morishita S. 2013. The transcription start site landscape of C. elegans. *Genome Research* **23**:1348–1361. DOI: https://doi.org/10.1101/gr.151571.112, PMID: 23636945

Sandhir R, Berman NE. 2010. Age-dependent response of CCAAT/enhancer binding proteins following traumatic brain injury in mice. *Neurochemistry International* **56**:188–193. DOI: https://doi.org/10.1016/j.neuint.2009.10.002, PMID: 19833158

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *PNAS* **110**:2876–2881. DOI: https://doi.org/10.1073/pnas.1221904110, PMID: 23382218

Sloutskin A, Danino YM, Orenstein Y, Zehavi Y, Doniger T, Shamir R, Juven-Gershon T. 2015. ElemeNT: a computational tool for detecting core promoter elements. *Transcription* **6**:41–50. DOI: https://doi.org/10.1080/21541264.2015.1067286, PMID: 26226151

Tepper RG, Ashraf J, Kaletsky R, Kleemann G, Murphy CT, Bussemaker HJ. 2013. PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. *Cell* **154**:676–690. DOI: https://doi.org/10.1016/j.cell.2013.07.006, PMID: 23911329

Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, Biggin MD, Stamatoyannopoulos JA. 2011. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biology* **12**:R43. DOI: https://doi.org/10.1186/gb-2011-12-5-r43, PMID: 21569360

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**:178–192. DOI: https://doi.org/10.1093/bib/bbs017, PMID: 22517427

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**:75–82. DOI: https://doi.org/10.1038/nature11232, PMID: 22955617

Tian Y, Garcia G, Bian Q, Steffen KK, Joe L, Wolff S, Meyer BJ, Dillin A. 2016. Mitochondrial Stress Induces Chromatin Reorganization to Promote Longevity and UPR(mt). *Cell* **165**:1197–1208. DOI: https://doi.org/10.1016/j.cell.2016.04.011, PMID: 27133166

Tittel-Elmer M, Bucher E, Broger L, Mathieu O, Paszkowski J, Vaillant I. 2010. Stress-induced activation of heterochromatic transcription. *PLoS Genetics* **6**:e1001175. DOI: https://doi.org/10.1371/journal.pgen.1001175, PMID: 21060865

Uno M, Honjoh S, Matsuda M, Hoshikawa H, Kishimoto S, Yamamoto T, Ebisuya M, Yamamoto T, Matsumoto K, Nishida E. 2013. A fasting-responsive signaling pathway that extends life span in C. elegans. *Cell Reports* **3**:79–91. DOI: https://doi.org/10.1016/j.celrep.2012.12.018, PMID: 23352664

van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, van Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology* **35**:145–153. DOI: https://doi.org/10.1038/nbt.3754, PMID: 28024146

Wagner CR, Kuervers L, Baillie DL, Yanowitz JL. 2010. xnd-1 regulates the global recombination landscape in Caenorhabditis elegans. *Nature* **467**:839–843. DOI: https://doi.org/10.1038/nature09429, PMID: 20944745

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:355–364. DOI: https://doi.org/10.1038/nature13992, PMID: 25409824

Zeiser E, Frøkjær-Jensen C, Jorgensen E, Ahringer J. 2011. MosSCI and gateway compatible plasmid toolkit for constitutive and inducible expression of transgenes in the C. elegans germline. *PLoS One* **6**:e20082. DOI: https://doi.org/10.1371/journal.pone.0020082, PMID: 21637852

Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek P. 2014. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* **30**:1008–1009. DOI: https://doi.org/10.1093/bioinformatics/btt737, PMID: 24363377

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**:R137. DOI: https://doi.org/10.1186/gb-2008-9-9-r137, PMID: 18798982

Zhang H, Gao L, Anandhakumar J, Gross DS. 2014. Uncoupling transcription from covalent histone modification. *PLoS Genetics* **10**:e1004202. DOI: https://doi.org/10.1371/journal.pgen.1004202, PMID: 24722509

Appendix C

Introduction to VplotR

VplotR: a visualization package to generate fragment density plots from high-throughput sequencing data

Jacques Serizay[*]
Julie Ahringer[†]

March 2020

Contents

1 Introduction

This R package makes the process of generating fragment density plots (also known as "V-plots") straightforward. V-plots have been introduced for the first time by the Henikoff lab in 2011 (Henikoff et al. 2011). More recently, V-plots have proven to be very instructive to understand the local organization of the chromatin at regulatory elements. For instance, the nucleoATAC package relies on cross-correlation of ATAC-seq fragment density plots to accurately map nucleosome occupancy along the genome (Schep et al. 2015).

[*] The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, United Kingdom, jacquesserizay@gmail.com

[†] The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, United Kingdom, ja219@cam.ac.uk

VplotR aim is to streamline the process of generating V-plots. It contains wrapping functions to import paired-end sequencing bam files and generate V-plots around genomic loci of interest. VplotR is designed around ggplot2 and makes full use of its potential (Wickham et al. 2019). As such, it is easy to generate V-plots in batch and combine them with other plots to make publication-ready figures.

VplotR is aimed toward investigating ATAC-seq datasets but can be used to plot other types of paired-end sequencing datasets such as DNase-seq or MNase-seq. This vignettes illustrates how VplotR can be leveraged to investigate the local arrangment of nucleosomes around different sets of promoters.

2 Installation

VplotR can be installed as follows:

```r
install.packages("devtools")
devtools::install_github("js2264/VplotR")
library(VplotR)
```

3 Quick use

A V-plot of ATAC-seq fragments (`fragments`) over loci of interest (`granges`) can be generated using the `plotVmat()` function:

```r
plotVmat(fragments, granges)
```

Multiple V-plots can be generated in parallel as follow:

```r
list_params <- list(
    "sample_1" = list("bam" = fragments_1, "granges" = granges_1),
    "sample_2" = list("bam" = bam_2, "granges" = granges_2),
    ...,
    "sample_N" = list("bam" = bam_N, "granges" = granges_N)
)
plotVmat(
    list_params,
    cores = length(list_params)
) + ggplot2::facet_wrap(~Cond.)
```

4 Detailed use of VplotR: Positioning of nucleosomes flanking ubiquitous or tissue-specific promoters

4.1 Importing data

VplotR use requires several objects. First, genomic loci of interest should be stored in a GRanges object. In this example, we will fetch promoter and enhancer annotations stored on the Ahringer server.

```
ce_REs <- readRDS(
    url('https://ahringerlab.com/VplotR/ce11_annotated_REs.rds')
)
ce_proms <- ce_REs[ce_REs$is.prom]
proms_list <- list(
    "Ubiq._proms" = ce_proms[ce_proms$which.tissues == 'Ubiq.'],
    "Germline_proms" = ce_proms[ce_proms$which.tissues == 'Germline'],
    "Neurons_proms" = ce_proms[ce_proms$which.tissues == 'Neurons'],
    "Muscle_proms" = ce_proms[ce_proms$which.tissues == 'Muscle'],
    "Hypod._proms" = ce_proms[ce_proms$which.tissues == 'Hypod.'],
    "Intest._proms" = ce_proms[ce_proms$which.tissues == 'Intest.']
)
```

Secondly, ATAC-seq bam files can be imported using the `importPEBamFiles()` function.

```
# Do not run
bam_files <- paste0('~/bam-files/ATAC_',
    c('mixed', 'germline', 'neuron',
        'muscle', 'hypodermis', 'intestine'
    ), '.bam'
)
bam_list <- importPEBamFiles(
    bam_files,
    where = GenomicRanges::reduce(
        GenomicRanges::resize(ce_REs, width = 2000, fix = 'center')
    ),
    shift_ATAC_fragments = TRUE
) %>% setNames(c(
    'mixed', 'Germline', 'Neurons',
    'Muscle', 'Hypod.', 'Intest.'
))
# // Do not run
```

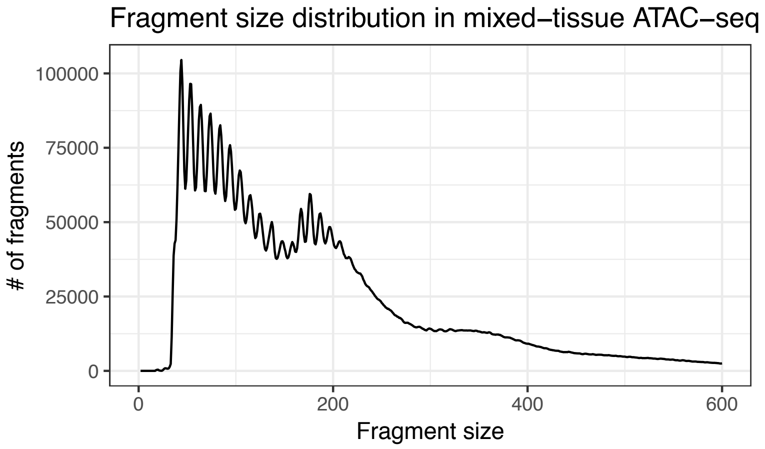
In this vignette, ATAC-seq bam files already imported are fetched from the Ahringer server:

```
bam_list <- readRDS(
    url('https://ahringerlab.com/VplotR/ATAC_PE_fragments.rds')
)
```

4.2 Plotting fragment size distribution

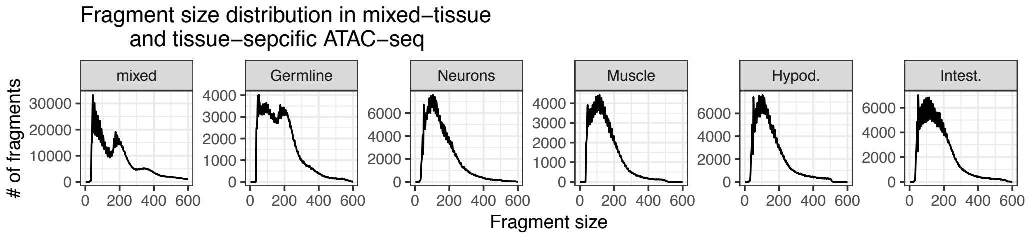
VplotR allows investigation of the distribution of fragments relative to the centers of loci of interest by separating them according to their length. Before generating fragment density plots, one might want to visualize the distribution of fragment lengths in a given sample. The following command plots the distribution of fragment found in a mixed-tissue ATAC-seq sample and mapping over annotated *C. elegans* promoters:

```
sizes <- getFragmentsDistribution(
    bam_list[['mixed']],
    ce_proms,
    limits = c(0, 1200)
)
p_distr <- ggplot(sizes, aes(x = x, y = y), color = '#991919') +
    geom_line() +
    theme_bw() +
    labs(
        title = 'Fragment size distribution in mixed-tissue ATAC-seq',
        x = 'Fragment size',
        y = '# of fragments'
    ) +
    xlim(c(0, 600))
```



This plot highlights the multi-modal distribution of ATAC-seq fragments found over promoters. However, when splitting the fragments mapping to each set of ubiquitous or tissue-specific promoters, a striking difference in fragment length distribution is observed between germline and somatic-tissue-specific promoters.

```r
sizes <- parallel::mclapply(mc.cores = 6,
    c('mixed', 'Germline', 'Neurons', 'Muscle', 'Hypod.', 'Intest.'),
    function(TISSUE) {
        idx <- paste0(ifelse(TISSUE == 'mixed', 'Ubiq.', TISSUE), '_proms')
        d <- getFragmentsDistribution(
            bam_list[[TISSUE]],
            proms_list[[idx]],
            limits = c(0, 1200)
        )
        return(data.frame(d, tissue = TISSUE))
    }
) %>% do.call(rbind, .)
p_distr_split <- ggplot(sizes, aes(x = x, y = y), color = '#991919') +
    geom_line() +
    theme_bw() +
    labs(
        title = 'Fragment size distribution in mixed-tissue
        and tissue-sepcific ATAC-seq',
        x = 'Fragment size',
        y = '# of fragments'
    ) +
    facet_wrap(~tissue, scales = 'free', nrow = 1) +
    xlim(c(0, 600))
```



4.3 Plotting ATAC-seq fragment density plots

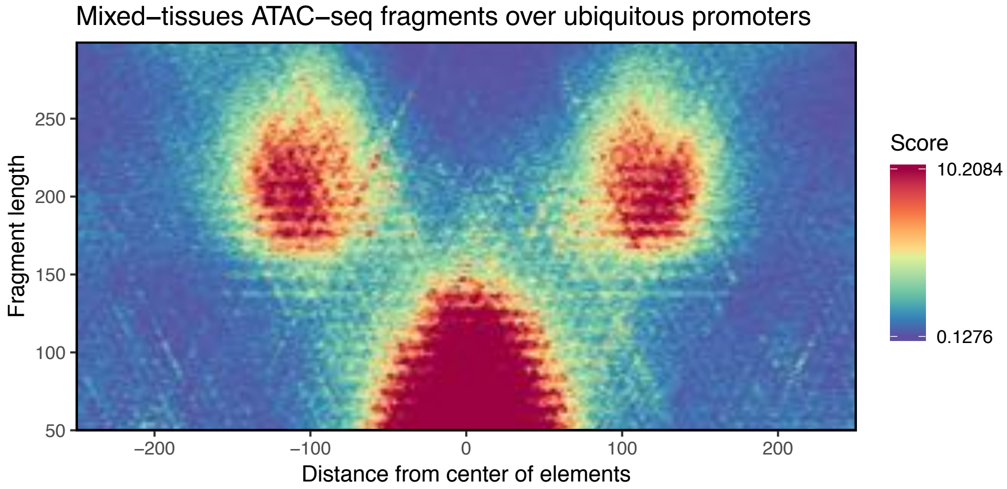
The fragment length distribution plots generated above do not show where shorter or longer ATAC-seq fragments respectively map. ATAC-seq fragment density plots, a.k.a. V-plots, can be used to show the distribution of ATAC-seq fragments of variable lengths over a set of regions of interest. In such plot, the horizontal axis represents the position of the fragments relative to the center of the loci of interest, and the vertical axis represents the length of the fragments. The color code symbolizes the density of fragments.

For example, a V-plot of ATAC-seq fragments over ubiquitous promoters (from a mixed-tissue ATAC-seq sample) can be generated as follows:

```
Vplot_ubiq <- bam_list[['mixed']] %>%
    computeVmat(proms_list[['Ubiq._proms']]) %>%
    normalizeVmat(normFun = 'pctsum', roll = 3) %>%
    plotVmat(
        main = 'Mixed-tissues ATAC-seq fragments over ubiquitous promoters'
    )
```

The different steps to generate a V-plot over a set of genomic loci can be wrapped in a single call to the `plotVmat()` function:

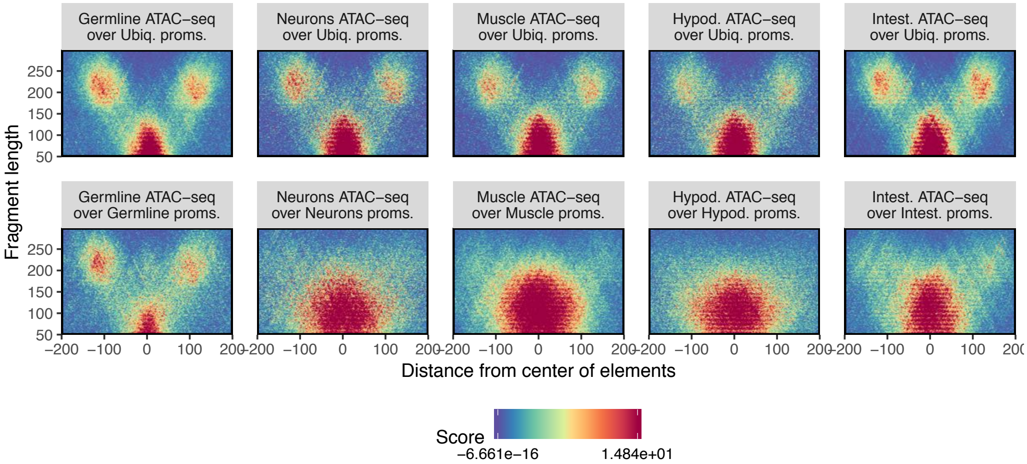
```
Vplot_ubiq <- plotVmat(
    bam_list[['mixed']],
    proms_list[['Ubiq._proms']],
    main = 'Mixed-tissues ATAC-seq fragments over ubiquitous promoters'
)
```



To compute many different V-plots simultaneously, one can pass the two main arguments (bam_granges and granges) to the plotVmat function using a named list. Let's generate V-plots for each of the five tissue-specific ATAC-seq datasets; for each tissue-specific ATAC-seq dataset, two V-plots will be generated:

- One for fragments mapping over ubiquitous promoters;
- One for fragments mapping over the corresponding tissue-specific promoters;

```
list_params <- list(
    "Germline ATAC-seq\nover Ubiq. proms." = list(
        bam_list[['Germline']], proms_list[['Ubiq._proms']]),
    "Germline ATAC-seq\nover Germline proms." = list(
        bam_list[['Germline']], proms_list[['Germline_proms']]),
    "Neurons ATAC-seq\nover Ubiq. proms." = list(
        bam_list[['Neurons']], proms_list[['Ubiq._proms']]),
    "Neurons ATAC-seq\nover Neurons proms." = list(
        bam_list[['Neurons']], proms_list[['Neurons_proms']]),
    "Muscle ATAC-seq\nover Ubiq. proms." = list(
        bam_list[['Muscle']], proms_list[['Ubiq._proms']]),
    "Muscle ATAC-seq\nover Muscle proms." = list(
        bam_list[['Muscle']], proms_list[['Muscle_proms']]),
    "Hypod. ATAC-seq\nover Ubiq. proms." = list(
        bam_list[['Hypod.']], proms_list[['Ubiq._proms']]),
    "Hypod. ATAC-seq\nover Hypod. proms." = list(
        bam_list[['Hypod.']], proms_list[['Hypod._proms']]),
    "Intest. ATAC-seq\nover Ubiq. proms." = list(
        bam_list[['Intest.']], proms_list[['Ubiq._proms']]),
    "Intest. ATAC-seq\nover Intest. proms." = list(
        bam_list[['Intest.']], proms_list[['Intest._proms']])
)
plots <- plotVmat(
    list_params,
    xlim = c(-200, 200),
    estimate_background = TRUE,
    cores = length(list_params)
)
Vplots <- plots +
    facet_wrap(~Cond., nrow = 2, dir = 'v') +
    theme(legend.position = 'bottom') +
    theme(panel.spacing = unit(1, "lines"))
```

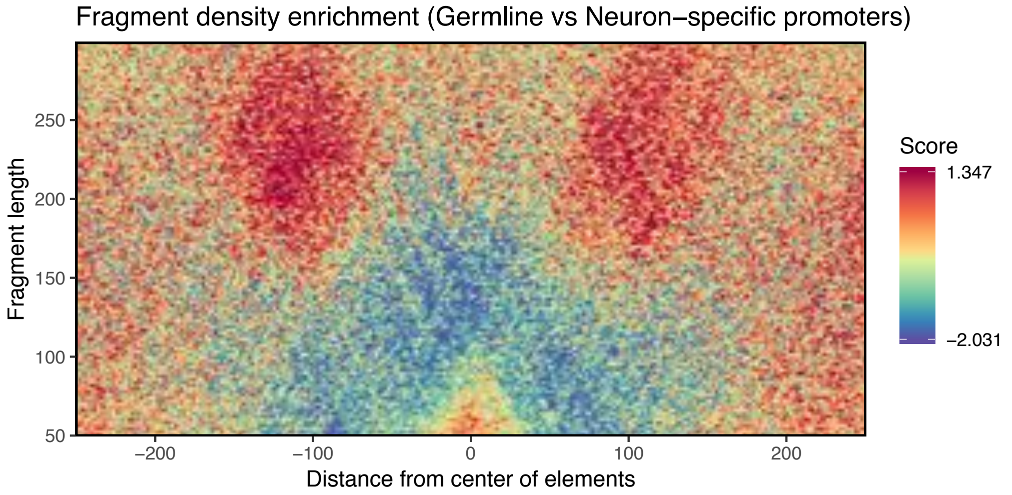


In the five tissue-specific ATAC-seq datasets, ubiquitous promoters are flanked by nucleosomes in all tissues. Germline promoters are also characterized by prominent flanking nucleosomes, while somatic promoters are not.

4.4 Direct comparison of two V-plots

To better understand the differences in nucleosome organization in different contexts, two V-plots can be directly compared to each other. Here, let's compare the nucleosome positioning at neuron-specific promoters (using neuron-specific ATAC-seq data)

```r
Vmat_neurons <- plotVmat(
    bam_list[['Neurons']],
    proms_list[['Neurons_proms']],
    estimate_background = FALSE,
    ylim = c(50, 200),
    xlim = c(-200, 200),
    cores = 10,
    roll = 3,
    return_Vmat = TRUE
)
Vmat_germline <- plotVmat(
    bam_list[['Germline']],
    proms_list[['Germline_proms']],
    estimate_background = FALSE,
    ylim = c(50, 200),
    xlim = c(-200, 200),
    cores = 10,
    roll = 3,
    return_Vmat = TRUE
)
Vmat_comp <- log2((Vmat_germline+1)/(Vmat_neurons+1))
Vplot_comp <- plotVmat(
    Vmat_comp,
    cores = 10,
    roll = 2,
    main = 'Fragment density enrichment (Germline vs Neuron-specific promoters)'
)
```



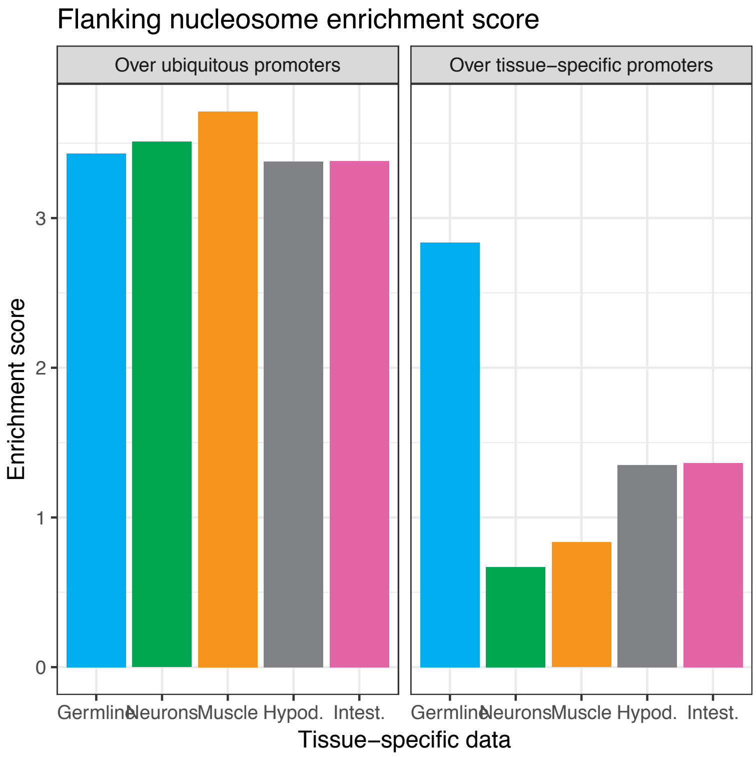
4.5 Nucleosome enrichment score

The enrichment of flanking nucleosomes around a nucleosome-depleted central region can be estimated from a ATAC-seq fragment density plot using the `nucleosomeEnrichment()` function. This function is used to quantify the local enrichment of nucleosomal fragments compared to background noise. Several flanking nucleosome enrichment scores can be computed at once as follows:

```
list_scores <- parallel::mclapply(
    c("Germline", "Neurons", "Muscle", "Hypod.", "Intest."),
    function(TISSUE) {
        message('>> ', TISSUE)
        nucenrich_tissue_spe_proms <- nucleosomeEnrichment(
            bam_granges = bam_list[[TISSUE]],
            granges = proms_list[[paste0(TISSUE, '_proms')]],
            estimate_background = TRUE,
            verbose = TRUE
        )
        nucenrich_ubiq_proms <- nucleosomeEnrichment(
            bam_granges = bam_list[[TISSUE]],
            granges = proms_list[['Ubiq._proms']],
            estimate_background = TRUE,
            verbose = TRUE
        )
        return(list(
            'tissue-spe-proms' = nucenrich_tissue_spe_proms,
            'ubiq-proms' = nucenrich_ubiq_proms
        ))
    },
    mc.cores = 5
) %>% setNames(c("Germline", "Neurons", "Muscle", "Hypod.", "Intest."))
```

Once the flanking nucleosome enrichment scores are computed, they can all be plotted together:

```
nucenrich_scores <- data.frame(
    tissue = factor(
        rep(names(list_scores), each = 2),
        levels = names(list_scores)
    ),
    promoters = c(rbind(
        paste0(names(list_scores), ' promoters'), rep('Ubiq. promoters', 5)
    )),
    promoters2 = factor(rep(c(0.5, 0.8), 5)),
    score = unlist(lapply(
        list_scores, function(Vmat) {
            c(
                Vmat[[1]]$fisher_test$estimate,
                Vmat[[2]]$fisher_test$estimate
            )
        })),
    is_ubiq = factor(rep(
        c('Over tissue-specific promoters', 'Over ubiquitous promoters'), 5),
        levels = c('Over ubiquitous promoters', 'Over tissue-specific promoters')
    )
)
nucenrich_plots <- ggplot(nucenrich_scores, aes(
    x = tissue,
    y = score,
    fill = tissue,
    group = is_ubiq
)) +
    geom_col() +
    facet_wrap(~is_ubiq, nrow = 1) +
    scale_fill_manual(
        values = c("#00AEEF", "#00A651", "#F7941D", "#808285", "#E365A6")
    ) +
    labs(
        title = 'Flanking nucleosome enrichment score',
        y = 'Enrichment score',
        x = 'Tissue-specific data'
    ) +
    theme_bw() +
    theme(legend.position = 'none')
```



According to these results, flanking nucleosomes are significantly enriched at ubiquitous promoters in nuclei from all tissues. Among tissue-specific promoters, only germline promoters have a significant enrichment of flanking nucleosome. The promoters only active in the soma (*e.g.* neuron-specific promoters) do not show any strong nucleosome enrichment flanking their central NDR.

5 Session info

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.2 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] magrittr_1.5  ggplot2_3.1.1 VplotR_0.4.0  knitr_1.22
##
## loaded via a namespace (and not attached):
##  [1] SummarizedExperiment_1.12.0      tinytex_0.20
##  [3] zoo_1.8-5                        tidyselect_0.2.5
##  [5] xfun_0.5                         purrr_0.3.2
##  [7] reshape2_1.4.3                   lattice_0.20-38
##  [9] colorspace_1.4-1                 htmltools_0.3.6
## [11] stats4_3.5.2                     rtracklayer_1.42.2
## [13] yaml_2.2.0                       XML_3.98-1.19
## [15] rlang_0.4.2                      pillar_1.3.1
## [17] glue_1.3.1                       withr_2.1.2
## [19] BiocParallel_1.16.6              BiocGenerics_0.28.0
## [21] RColorBrewer_1.1-2               matrixStats_0.54.0
## [23] GenomeInfoDbData_1.2.0           plyr_1.8.4
## [25] stringr_1.4.0                    zlibbioc_1.28.0
## [27] Biostrings_2.50.2                munsell_0.5.0
## [29] gtable_0.3.0                     evaluate_0.13
## [31] labeling_0.3                     Biobase_2.42.0
## [33] IRanges_2.16.0                   GenomeInfoDb_1.18.2
## [35] parallel_3.5.2                   Rcpp_1.0.1
## [37] scales_1.0.0                     BSgenome_1.50.0
## [39] DelayedArray_0.8.0               S4Vectors_0.20.1
## [41] XVector_0.22.0                   Rsamtools_1.34.1
## [43] digest_0.6.18                    stringi_1.3.1
## [45] bookdown_0.9.2                   dplyr_0.8.0.1
## [47] GenomicRanges_1.34.0             grid_3.5.2
## [49] tools_3.5.2                      bitops_1.0-6
## [51] lazyeval_0.2.2                   RCurl_1.95-4.12
## [53] tibble_2.1.1                     crayon_1.3.4
## [55] pkgconfig_2.0.2                  BSgenome.Celegans.UCSC.ce11_1.4.2
## [57] Matrix_1.2-17                    assertthat_0.2.1
## [59] rmarkdown_1.12.6                R6_2.4.0
## [61] GenomicAlignments_1.18.1         compiler_3.5.2
```

References

Henikoff, Jorja G, Jason A Belsky, Kristina Krassovsky, David M MacAlpine, and Steven Henikoff. 2011. "Epigenome Characterization at Single Base-Pair Resolution." *Proc. Natl. Acad. Sci. U. S. A.* 108 (45): 18318–23.

Schep, Alicia N, Jason D Buenrostro, Sarah K Denny, Katja Schwartz, Gavin Sherlock, and William J Greenleaf. 2015. "Structured Nucleosome Fingerprints Enable High-Resolution Mapping of Chromatin Architecture Within Regulatory Regions." *Genome Res.*

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *JOSS* 4 (43): 1686.

Appendix D

periodicDNA: an R package to investigate nucleotide periodicity

This appendix introduces a manuscript entitled “periodicDNA: an R package to investigate nucleotide periodicity”. During my thesis, I designed a methodology to measure the periodicity of oligonucleotides in DNA sequences. This manuscript presents periodicDNA, the package I created to investigate dinucleotide periodicity in different classes of regulatory elements. This manuscript will soon be submitted to Journal of Open-Source Software for peer-reviewed publication.

periodicDNA: an R package to investigate oligonucleotide periodicity

Jacques Serizay[*]
Julie Ahringer[†]

March 2020

Contents

1 Statement of Need

periodicDNA provides a framework to quantify oligonucleotide periodicity over individual or multiple DNA genomci loci.

2 Summary

periodicDNA is an R package offering a set of functions to identify local periodic elements in short sequences such as regulatory elements. Many oligonucleotides are periodically occurring in genomes across eukaryotes, and some are impacting the physical properties of DNA. Notably, DNA bendability is modulated by 10-bp periodic occurrences of WW (W = A/T) dinucleotides. The package relies on Biostrings and GenomicRanges packages to handle DNA sequences and genome assemblies. It uses the Fourier Transform to measure the periodicity of a given oligonucleotide in sets of sequences. It also provides methods to generate continuous tracks of oligonucleotide periodicity over genomic loci, as well as visualization tools to interpret these tracks. The use of periodicDNA has already shed light on fundamental differences in sequence features in functional classes of promoters (Serizay et al. (2020)). We hope that the integration of this open-source package into genomic assay analysis workflows will help further improve our understanding of chromatin organization.

3 Methodology

periodicDNA can be used to estimate the power spectral density (PSD) of a given dinucleotide (`motif` argument) at specific periods (`period` argument) in a set of sequences of interest (`seqs` argument), using a simple wrapper function:

[*] The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, United Kingdom, jacquesserizay@gmail.com

[†] The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, United Kingdom, ja219@cam.ac.uk

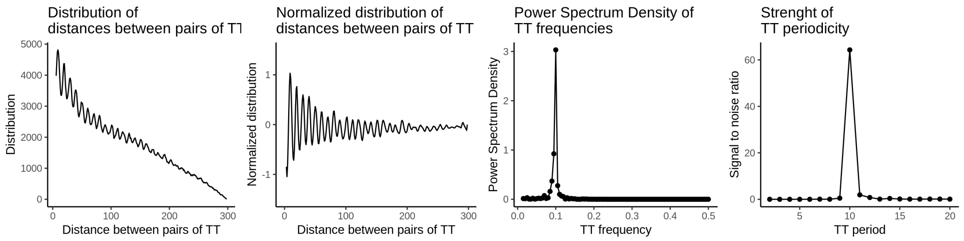


Figure 1: Output of plotPeriodicityResults() function. A- Distribution of all the TT pairwise distances. B- Normalized TT pairwise distance frequency. C- Power Spectral Density (PSD) of TT occurrences. D- PSD signal-to-noise ratio.

```r
library(periodicDNA)
library(magrittr)
library(ggplot2)
## The periodicity can be calculated from DNAStringSet objects:
data(ce_proms_seqs)
score <- getPeriodicity(
    seqs = ce_proms_seqs,
    motif = 'TT',
    cores = 4
)
## Alternatively, the periodicity can be calculated
## from a GRanges object in combination with a genome:
data(ce_proms)
score <- getPeriodicity(
    granges = ce_proms[ce_proms$which.tissues == 'Ubiq.'] %>%
        '['(strand(.) == '+') %>%
        resize(width = 1, fix = 'end') %>%
        resize(width = 300, fix = 'start'),
    genome = 'ce11',
    motif = 'TT',
    cores = 120
)
## Results can be plotted using the plotPeriodicityResults() function:
## See Figure 1
plots <- plotPeriodicityResults(score) %>%
    cowplot::plot_grid(plotlist = ., nrow = 1)
```

The intermediate steps internally performed when calling this function are the following (Figure 2):

1. In each sequence of a set of n sequences (the `seqs` argument), all the pairs of the dinucleotide of interest (the `motif` argument, e.g. `TT`) are identified and their pairwise distances are measured.
2. The distribution of the all the resulting pairwise distances (also called “distogram”) is generated.
3. The following normalization steps are then performed:
    - The distogram is transformed into a frequency histogram and then normalized by the following steps:
    - The frequency histogram follows a marked overall decrease of frequencies with increased pairwise distances. Indeed, for a 200-bp long sequence containing 20 WW dinucleotides exactly distant from each other by 10 base pairs, there are 19 pairs with a pairwise distance of 10 but only 1 pair of dinucleotides with a pairwise distance of 190. To overcome this distance decay, the frequency histogram is smoothed using a moving average window of 10 and the resulting smoothed frequency histogram is substracted from the frequency histogram. This effectively transforms the decreasing frequency histogram into a dampened oscillating signal and improves the PSD estimation by Fourier Transform.
    - The dampened oscillating signal is then scaled (i.e. mean-centered and normalized) and smoothed using a moving average window of 3. This effectively removes the effect of the latent 3-bp periodicity

of most dinucleotides found in eukaryote genomes (Gutiérrez et al., 1994).

4. A Fast Fourier Transform (FFT) is then used to estimate the power spectral density (PSD) of the normalized oscillating distribution at different periods (the `period` argument).

The PSD can be used in itself to identify which dinucleotide frequencies are enriched in the provided set of sequences. Its amplitude at a given frequency can also be used to compare dinucleotide frequencies across samples.

4 Leveraging periodicDNA to understand chromatin organization

Soon after solving the structure of nucleosomes, Kornberg raised a fundamental question: whether the positioning of nucleosomes in vivo in regard to a DNA locus was "specific" or "statistical" (Kornberg (1981)). Nucleosome "specific" positioning implies that the physicochemical properties of DNA sequences are enough to explain how nucleosomes are arranged along a DNA double-helix (e.g. described in Segal et al. (2006)). On the contrary, a "statistical" positioning postulates the presence of a "boundary" nucleosome (either a protein or a strong intrinsic positioning sequence, or both) which specifies one end of a nucleosomal array not determined by the physicochemical properties of DNA sequence (e.g. described in Mavrich et al. (2008)). Later on, biochemists and computational biologists found out that periodic dinucleotide sequences were associated with positioned nucleosomes, suggesting that the "specific" model is contributing to nucleosome positioning - at least to a certain extent (Jiang & Pugh (2009); Struhl & Segal (2013) for review). To test whether specific periodic sequences were associated with the positioning of nucleosomes directly flanking regulatory elements, I leveraged periodicDNA. I focused on ubiquitous and tissues-specific promoters and enhancers, splitting each element into core (-70 to +70 base pairs around the center of the regulatory element), flanking (-210 to -70 base pairs and +70 to +210 base pairs) and distal sequences (-350 to -210 base pairs and +210 to +350 base pairs) (Figure 3A). Ubiquitous and germline-specific promoters exhibit a high TT 10-bp periodicity in the flanking sequences which decreases remarkably in the neighboring distal sequences (Figure 3B). In contrast, ubiquitous and germline enhancers both show a mild increase of TT 10-bp periodicity in flanking sequences as well as in distal sequences of enhancers (Figure 3B). This suggests that the 10bp TT periodicity can act as a local positioning signal at ubiquitous and germline-specific promoters, but not at ubiquitous and germline-specific enhancers (Figure 3C). This 10-bp TT periodicity is absent in other somatic tissue-specific promoters, as expected from previous results (Serizay et al. (2020)), as well as in somatic tissue-specific enhancers (Figure 3B). This supports a model whereby nucleosome organization at soma-restricted regulatory elements does not primarily depend on the underlying DNA sequence (Figure 3C).

5 Acknowledgments

The work was supported by a Wellcome Trust Senior Research Fellowship to J.A. (101863) and a Medical Research Council DTP studentship to J.S..

References

Jiang, C., & Pugh, B. F. (2009). Nucleosome positioning and gene regulation: Advances through genomics. *Nat. Rev. Genet.*, *10*(3), 161–172. doi:10.1038/nrg2522

Kornberg, R. (1981). The location of nucleosomes in chromatin: Specific or statistical. *Nature*, *292*(5824), 579–580. doi:10.1073/pnas.78.2.1095

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., et al. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, *18*(7), 1073–1083. doi:10.1101/gr.078261.108

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., et al. (2006). A genomic code for nucleosome positioning. *Nature*. doi:10.1038/nature04979

Serizay, J., Dong, Y., Jänes, J., Chesney, M., Cerrato, C., & Ahringer, J. (2020). Tissue-specific profiling reveals distinctive regulatory architectures for ubiquitous, germline and somatic genes. *bioRxiv*, 2020.02.20.958579. doi:10.1101/2020.02.20.958579

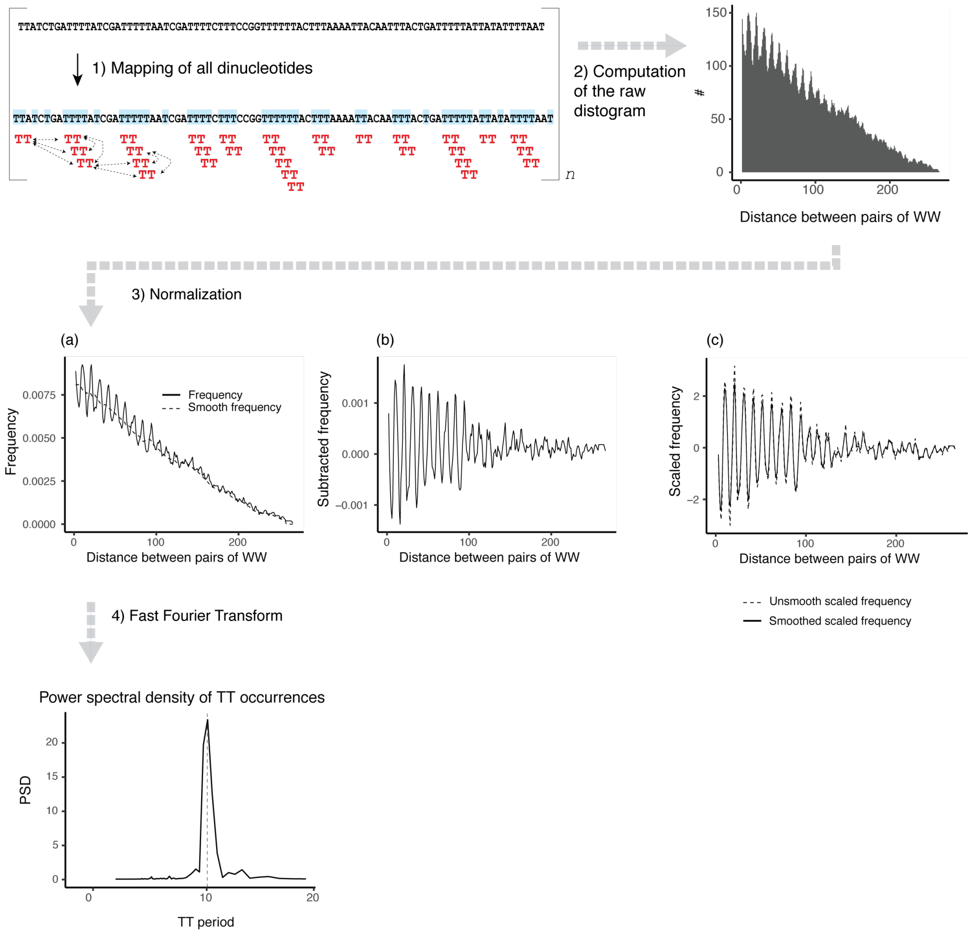

TTATCTGATTTTATCGATTTTTAATCGATTTTCTTTCCGGTTTTTTACTTTAAAATTACAATTTACTGATTTTTATTATATTTTAAT
1) Mapping of all dinucleotides
TTATCTGATTTTATCGATTTTTAATCGATTTTCTTTCCGGTTTTTTACTTTAAAATTACAATTTACTGATTTTTATTATATTTTAAT
TT
n
2) Computation of the raw distogram
#
150
100
50
0
0
100
200
Distance between pairs of WW
3) Normalization
(a)
Frequency
0.0075
0.0050
0.0025
0.0000
Frequency
Smooth frequency
0
100
200
Distance between pairs of WW
(b)
Subtracted frequency
0.001
0.000
−0.001
0
100
200
Distance between pairs of WW
(c)
Scaled frequency
2
0
−2
0
100
200
Distance between pairs of WW
Unsmooth scaled frequency
Smoothed scaled frequency
4) Fast Fourier Transform
Power spectral density of TT occurrences
PSD
20
15
10
5
0
0
10
20
TT period

Figure 2: Internal steps of the getPeriodicity() function. Each step is further described in the main text. The dotted double-arrows in the first step represent the distances measured by periodicDNA between some of the pairs of TT. For the single sequence shown here, there are 31 individual TT dinucleotides, resulting in $\binom{31}{2} = 465$ different pairs in total.

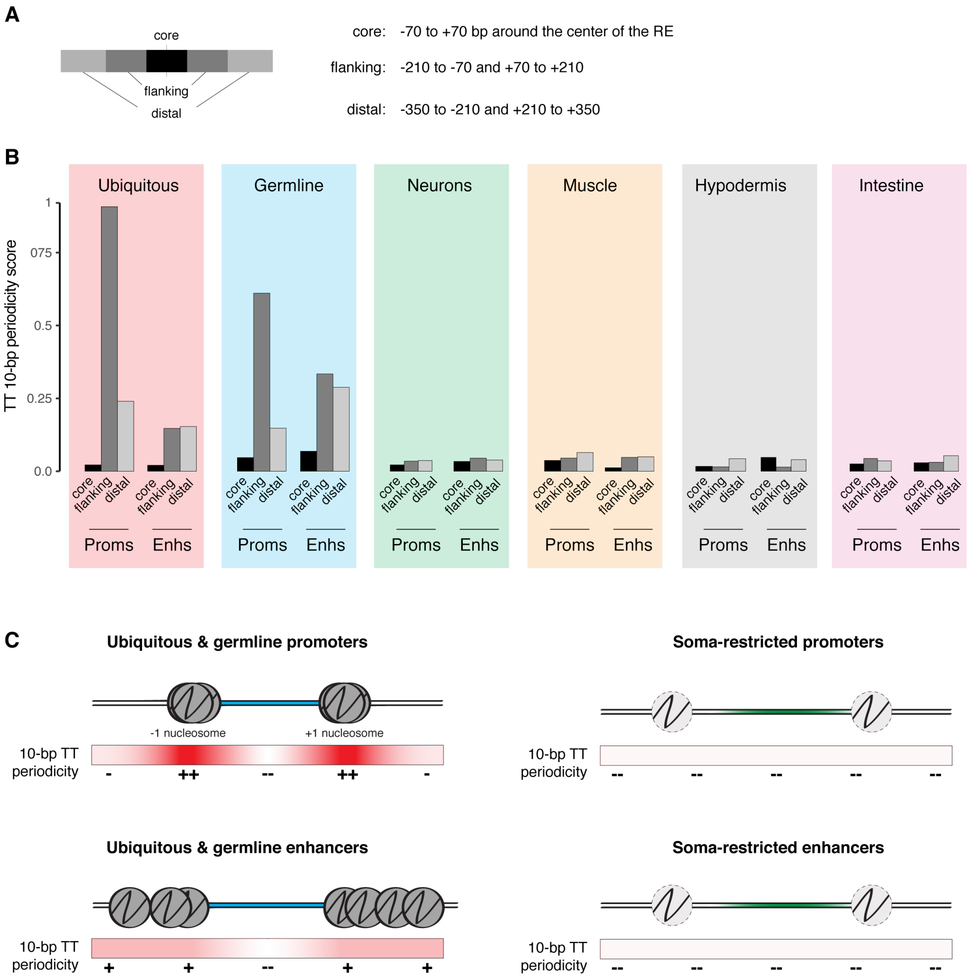


Figure 3: TT periodicity in promoters and enhancers. A- Pictogram representing how regulatory elements were divided into core, flanking and distal regions. The core sequence is the 140-bp long sequence at the center of the regulatory element; the flanking sequences range from -210 to -70 and from +70 to +210; the distal sequences range from -350 to -210 and from +210 to +350 (with the center of the regulatory element being the reference). B- TT 10-bp periodicity scores obtained from periodicDNA. C- Model of sequence-driven nucleosome positioning at different sets of promoters or enhancers. Three different situations are observed: (1) a decrease of TT periodicity on both sides of the flanking nucleosomes favors their precise positioning, (2) a weaker widespread TT periodicity favors nucleosome positioning without local enrichment and (3) absence of TT periodicity does not favor nucleosome positioning. Note that these models do not illustrate the role of other factors such as chromatin remodelers.

Struhl, K., & Segal, E. (2013). Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, *20*. doi:10.1038/nsmb.2506