Imperial College London

Department of Life Sciences

Identification of transcription factor targets, gene expression profiles and accessible chromatin regions in the *Caenorhabditis elegans* epidermis using targeted DamID

Dimitrios Katsanos

Statement of Originality

I declare that the content of my thesis is product of research and illustration conducted by myself in the lab of Dr Michalis Barkoulas at Imperial College London, except when otherwise stated. The contributions to the work presented in this thesis by students under my supervisions is acknowledged in the context of the corresponding experiments.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Licence (CC BY NC-SA).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that; you credit the author, do not use it for commercial purposes and share any derivative works under the same licence.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Development is an exceptionally complex process that is performed with exquisite control. A series of developmental programmes allow the orchestrated and tightly-regulated deployment of the genomic information, governing events like cell division, cell fate maintenance and differentiation. Understanding the complete regulatory states that instruct a selective decoding of the genome capable of bringing about morphogenetic events is central to developmental biology. Among all cells, stem cells maintain the potential to produce cells that undergo transitions in developmental trajectories and thus are particularly interesting. In this study, I have used the postembryonic development of the *Caenorhabditis elegans* epidermis driven by the stem cell-like seam cells, to begin exploring the gene regulatory network, transcriptional states and epigenomic regulation involved in cell fate patterning. To that end, I have adapted and present here the first application of the targeted DamID (TaDa) methodology in *C. elegans,* for assaying protein-DNA interactions, to use as a single technique in approaching all of the above objectives. I show that TaDa requires little starting material, is reproducible and tissue-specific. Using TaDa I identify targets for the transcription factors LIN-22 and NHR-25 that propose new biological functions for these regulators in epidermal development. I acquire gene expression profiles for the seam cells and hypodermis that lead to the discovery of novel transcription and chromatin factors, as well as new miRNAs. Finally, I produce the first cell-type-specific chromatin accessibility maps in *C. elegans* for the seam cells and hypodermis and use them to identify tissue-specific enhancers. These findings expand our knowledge of the mechanisms underlying fate decisions in epidermal patterning and provide a proof-of-concept for the application of TaDa in *C. elegans*.

Acknowledgements

The road to a PhD in the Michalis Barkoulas Lab at the department of Life Sciences at Imperial College, was what I hoped it would have been. A fundamentally creative experience, challenging at times, gratifying at others but motivating, intriguing and interesting throughout. For the work that led to this thesis and for the scientist that I have become, I first and foremost would like to thank my supervisor Dr Michalis Barkoulas. Michalis recognised my enthusiasm in research and gave me the opportunity to apply it in exciting questions and interesting problems in his lab, working towards a doctoral degree. He was a tireless mentor, always open to discussion, with insightful advice and guidance. Michalis never stifled creativity and allowed for ideas to be investigated and interests to develop, being attentive but trusting. He provided and promoted every opportunity for me to grow both as a scientist and as professional in general. His ambition is contagious and motivating and I am grateful I was his student.

I would also like to thank Dr Tony Southall and his lab members Dr Colin McClure and Dr Gabriel Aughey for offering me their expertise in TaDa, even at a moment's notice. This thesis would not have been as thorough if it wasn't for their insights. I am also thankful to my students Ziqi Lu, Jia Woo, Fangzi Zha, Fu Xiang Quah, Minge Pan and Max Biguet that have contributed in various ways to the research here and the experience of the degree. In particular, I would like to thank Mar Ferrando-Marco whose contribution towards this thesis was substantial. I am grateful to the Imperial College President's Scholarship scheme that supported me financially and along with the European Research Council funded my research. I also thank the Sir Alexander Fleming building Tech team that supported my laboratory research and SAF for feeling like home.

My sincere gratitude to all the past and present members of the Barkoulas Lab for their company, advice, valuable discussion and for making the lab a creative, fun environment, always cultivating team spirit. Michael Fasseas, Ritobrata Ghose, Guled Osman, Iqrah Razzaq, Sophie Gilbert, Florence Drury, Manish Grover, Mark Hintze, all contributed with their feedback and comments to my research, shaped my work ethic and helped me develop valuable skills. They are collectively responsible for making lab meetings the most intellectually challenging and creative time of the week. Especially, the discussions with my fellow PhD student Sneha Koneru were always enlightening and stimulating and I am grateful for her friendship.

I would like to thank my friends in London for providing both relaxing and exciting times, to recharge and enjoy every aspect of life and my friends in Athens who always make me feel like I have never left 6 years now and running. My thanks also go to the friends and co-players of Monday and Thursday night football that helped me stay physically active as well throughout my PhD.

I am indebted to my family. To my grandparents whose stories put our challenges into perspective and to my aunt, who gifted me my first microscope, for being my second mother and friend; their support was continuous. I cannot articulate the gratitude I have towards my parents. They have taught me both

Table of contents

List of figures

Abbreviations

ANOVA	Analysis Of Variance
ATAC-seq	Assay for Transposase-Accessible Chromatin identification by sequencing
CA	Chromatin Accessibility
CATaDa	Chromatin Accessibility Targeted DamID
CGC	*Caenorhabditis* Genetics Centre
ChIP	Chromatin ImmunoPrecipitation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Dam	DNA adenine methyltransferase
DamID	DNA adenine methyltransferase Identification
DAPI	4′,6-DiAmidino-2-PhenylIndole
DNA	DeoxyriboNucleic Acid
dNTP	deoxyNucleoside TriPhosphate
dsRNA	double-stranded RNA
DTT	DiThioThreitol
EA	Early-Adult
EDTA	EthyleneDiamineTetraacetic Acid
FACS	Fluorescence-Activated Cell Sorting
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements identified by sequencing
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
FPPMR	Fragment Pile-up Per Million Reads
GO	Gene Ontology
HOT	High-Occupancy Target
hpRNA	hairpin-RiboNucleic Acid
IES	Isoform End Site
IPTG	IsoPropyl β- d-1-ThioGalactopyranoside
ISS	Isoform Start Site
kb	kilo-basepairs
LB	Lysogeny Broth
LED	Light Emitting Diode
Mb	Mega-basepairs
miRNA	micro-RiboNucleic Acid
MosSCI	Mos1-mediated Single Copy Insertion
NAD	Nicotinamide Adenine Dinucleotide
NGM	Nematode Growth Medium

NGS	Next-Generation Sequencing
NLS	Nuclear Localisation Signal
NR	Nuclear Receptor
oligo	oligonucleotide
ORF	Open Reading Frame
PAT-seq	Poly(A)-Test RNA-sequencing
PBS	Phosphate-Buffered Saline
PCR	Polymerase Chain Reaction
PDE	postdeirid neuron
PEG	Polyethylene glycol
POI	Protein Of Interest
RNA	RiboNucleic Acid
RNAi	RiboNucleic Acid interference
RNApol	RiboNucleic Acid polymerase
RNA-seq	RiboNucleic Acid identification by sequencing
ROI	Region Of Interest
RPM	Revolutions Per Minute
rpm	reads per million
sci-RNA-seq	single-cell combinatorial indexing RNA identification by sequencing
SCM	Seam Cell Marker
scn	seam cell number
SD	Standard Deviation
SEM	Standard Error of the Mean
seq	sequencing
smFISH	single-molecule Fluoresncence *In Situ* Hybridization
SPMR	Signal Per Million Reads
SSC	Saline-Sodium Citrate
TaDa	Targeted DNA adenine methyltransferase identification
TBE	Tris/Borate/Ethylenediaminetetraacetic acid
TES	Transcriptional End Site
TF	Transcription Factor
TPM	Transcripts Per Million
TSS	Transcriptional Start Site
WLB	Worm Lysis Buffer
WT	Wild-Type
Wβa	Wnt/β-catenin asymmetry pathway

Chapter 1

General Introduction

1.1 Introductory concepts

1.1.1 Genomic regulation of development as the source of complexity and diversity in multicellular organisms

Multicellular organisms have conquered almost every habitat of our planet. They have achieved that throughout millions of years of selective pressure that allowed them to acquire the forms and the traits required for survival and reproduction. Metazoans exemplify this to the extreme, with a diversity of adaptive forms that underline the diversity of life as a whole. This wide variety of forms is a result of the specification of cells, in these multicellular organisms, to differentiated cell-types that carry out different functions. These cell-types constitute focal points for evolution to act upon and a majority of them are conserved across taxa (Arendt, 2008).

All this vast variety of differentiated cells that make up and define an organism arises through the process of development. Both the specification of cell-types and their organisation in functional morphologies, like tissues and organs, occurs through this process (Slack, 2006). The complexity and importance of the developmental process can be appreciated in the intuitive paradigm of the single-celled zygote, the unit from which all metazoans are formed, which develops to establish a complete organism that is several degrees more complex.

Development brings about this complexity by combining maternal instructions, cell-signalling and environmental cues, in decoding genomic information essential to execute developmental phenomena. Therefore, it is largely the process through which genotype is transformed to a developmental phenotype, typically in a very reproducible fashion, regardless of genetic variation, environmental or stochastic noise (Waddington, 1942; Félix & Barkoulas, 2015). The resulting gene expression is tightly regulated to establish expression programmes that instruct a variety of morphogenetic and patterning events where cells divide, acquire specialised fates and organise in specific structures.

Classic embryological experiments had shown that regardless of the final highly varied developmental outcomes, the first few patterning events are very similar in all organisms, with all eumetazoans for example, going through a gastrula stage (Slack, 2006; Gilbert, 2000). However, it wasn't until the mid-80s when we began to identify that key genes controlling these developmental events were actually conserved, performing equivalent developmental functions in those animals that possessed them (McGinnis *et al.*, 1984; Halder, Callaerts & Gehring, 1995). Therefore, the diversity of developmental outcomes that generate the myriads of forms we come across in nature, is driven to a great extent by a common core developmental toolkit of genes, with the key to diversity being the different regulation of when, where and at what levels are these genes expressed, as well as how they interact with each other and with more evolutionarily novel developmental factors (Cañestro, Yokoi & Postlethwait, 2007; Levine & Davidson, 2005). All these dimensions of which genes are expressed, at which developmental timepoint,

within what domain or tissue of an animal and to what quantitative extent towards leading to definition of the body plan, have been and still are central questions in the study of developmental biology.

1.1.2 Gene regulatory networks govern developmental decisions

A large proportion of such key developmental genes encode for transcription factor (TF) proteins that control the expression of multiple genes, by binding *cis*-regulatory elements often in their proximity and acting to activate or repress their expression, thus promoting or inhibiting specific developmental events (Zeitlinger & Stark, 2010; Spitz & Furlong, 2012). TFs with the input of signalling pathways and in combination with epigenomic regulation, control the precise spatiotemporal expression of a selection of genes and shape the transcriptional state of a cell which drive it towards specific fates (Ben-Tabou De-Leon & Davidson, 2007).

Epigenomic regulation refers to the collection of epigenetic control mechanism of gene expression across the genome (Callinan & Feinberg, 2006). The mechanisms of epigenetic regulation rely primarily on modifications of the DNA, like methylation of the bases, or on chemical modifications of the histones that result in changes in the chromatin compactness that can promote or inhibit gene expression (Jaenisch & Bird, 2003). This is achieved by closing or opening DNA sequences that can recruit TFs to initiate transcription (Tsompana & Buck, 2014). These somewhat stable alterations can be maternally defined or arise during development as a mode of cell specification, as they determine the selection of potential permissible targets for TFs and are heritable through mitosis, permiting the maintenance of defined transcriptional programs in differentiated cell-types and their progeny (Jaenisch & Bird, 2003; Wilson & Filipp, 2018). Therefore, epigenomic regulation and transcriptional regulation by transcription factors are linked in generating differential gene expression programs that specify cell fates during development. Even more so, transcription factors often act synergistically with the epigenetic machinery to establish new epigenomic states that will permit the expression of required batteries of genes to promote further cell specification (Wilson & Filipp, 2018).

Generally, determinants of cell fate including developmentally important TFs and their plethora of regulated targets can be characteristic of a cell-type and are often conserved providing a definition of cell-type across phylogenies (Arendt *et al.*, 2016; Zeitlinger & Stark, 2010; Cañestro, Yokoi & Postlethwait, 2007). Studying how transcription factors controlling developmental programmes achieve their phenotypic outcomes in model organisms, has revealed interconnected networks of regulatory interactions (Oliveri & Davidson, 2007). These networks that can be illustrated as logic architectures, are made up of links between the transcription factors and other genes, that in essence correspond to binding of *cis*-regulatory modules that control the expression of a target gene. These links possess specific directionality of which factor exerts the control and the type of regulatory relationship (activating or repressive). These networks describe and perform the precise quantitative and spatiotemporal regulation of gene expression programs that control the correct execution of the developmental event to which they correspond. This for example could be the formation of a tissue or organ, a pattern of cell divisions or a differentiation program.

Such gene networks have been found to be pervasive in animal development, with a range of potential levels of complexity, from a few factors to multi-layered networks made up of sub-circuits of factors with more focused functions (Oliveri & Davidson, 2007; Davidson, 2010). A large number of the traditional developmental control genes, which are major regulators of specific developmental events or cell-fates that are often conserved across species, have been found to participate and generally have central roles in more intricate networks that govern the developmental program for which they were identified (Davidson, 2010; Stathopoulos & Levine, 2005). Therefore, the network view of the unfolding of developmental processes, as opposed to single gene regulators, provides a better explanation for the observable strict control of developmental phenomena and allows better understanding of the underlying regulatory complexity that generates the observable complexity in developing organisms.

As is the case for important developmental genes, multiple developmental gene network modules are conserved across phyla and despite often having been layered with evolutionary novel factors and sub-circuits, they perform similar developmental functions in diverse body plans (Cañestro, Yokoi & Postlethwait, 2007; Oliveri & Davidson, 2007; Davidson & Erwin, 2006). Consequently, the identification and dissection of developmental gene networks in model organisms has the potential to translate to better understanding of disease in humans in those cases where critical components of gene networks fail.

Modern genomics methodologies have allowed us to decipher gene networks controlling developmental programmes of interest to a degree that had not been previously possible at this scale (St Johnston, 2015). Elucidation of complete gene expression profiles of developing tissues can uncover key developmental regulators that were previously unknown and high-throughput functional approaches can expose functions controlled by existing networks. At the same time, the ability to assay the genome-wide binding of transcription factors can expand the connections in existing networks to a previously unattainable scale. Methods probing chromatin state can further enrich gene regulatory networks with epigenomic regulation information. Tissue-specific and single-cell genomics can provide unrivalled detail in the study of fate decisions (St Johnston, 2015; Marioni & Arendt, 2017). Therefore, questions in developmental biology about the genetic mechanisms underlying patterning programmes can now be approached to great detail by dissecting the gene regulatory networks that govern them.

1.1.3 Cell division symmetry, differentiation and stem cell behaviour

As discussed above development is responsible for generating multiple different cell-types that make up a multicellular organism. A principal manifestation of a differentiation phenomenon, as well as a mediating mechanism for the successful acquisition of the differentiated fate, is asymmetric cell division. As opposed to a symmetric division that generates two identical daughter cells, a cell division can be defined as asymmetric based on a few different potential criteria. These are: whether the two resulting daughters differ in size, have asymmetric localisation of cellular content between the two or if the cells evidently possess different capacity to acquire a differentiated fate (Horvitz & Herskowitz, 1992).

Mechanisms explaining how the establishment of asymmetric divisions is achieved have been proposed and studied thoroughly in model organisms and can be summarised to external signals and intrinsic mechanisms (Neumüller & Knoblich, 2009; Horvitz & Herskowitz, 1992). External signals refer to cell signalling that acts on the two daughters differently, depending on the source of the signal or the orientation of the dividing cell within the tissue, while intrinsic mechanisms correspond to asymmetric segregation of fate determinants or positioning of the mitosis machinery prior to division (Horvitz & Herskowitz, 1992; Betschinger & Knoblich, 2004; Neumüller & Knoblich, 2009). However, even though there is wealth of information when it comes to how asymmetric divisions are executed, complete mechanistic understanding of how the differentiation to a specified fate is achieved in most cases is lacking (Neumüller & Knoblich, 2009).

Differentiation appears to be more complex, requiring specific regulatory states at the level of TF networks, chromatin architecture and the interface of how these are assembled in the regulatory apparatus of the differentiated cell (Newman, 2020). A question of great interest in developmental biology is how decisions between differentiation and asymmetric division, as opposed to symmetric division and maintenance of fate, are taken in those cells that have the capacity to perform both.

This behaviour is a defining characteristic of stem cells. Stem cells are central to the developmental process as they can be maintained in an undifferentiated state while possessing the potential to proliferate or differentiate, conditional to specific regulatory cues (Morrison & Spradling, 2008). More specifically, a fundamental stem cell trait is the ability to perform asymmetric self-renewal divisions. These divisions produce a daughter cell that will commit to a differentiated fate and one that is self-renewed, in that it retains the stem cell fate of the precursor cell, allowing stem cells to generate differentiated tissues while perpetuating themselves (Knoblich, 2008; Morrison & Kimble, 2006). Alternatively, stem cells can undergo a symmetric mode of division that is proliferative and expands their population as daughter cells are identical in fate to the mother (Morrison & Kimble, 2006). Therefore, stem cells are crucial for tissue morphogenesis, as well as replenishing tissues and mediating regeneration (Morrison & Kimble, 2006; Klein & Simons, 2011).

Strict regulation of stem cell fate decisions and numbers is vital to prevent tissue hyperplasia and cancer (Neumüller & Knoblich, 2009; Morrison & Kimble, 2006). Understanding the switch between symmetric and asymmetric modes of division can elucidate broader mechanisms underlying long-term repair capacity and how it can facilitate longer lifespan (Klein & Simons, 2011; Morrison & Kimble, 2006; Morrison & Spradling, 2008). Uncovering these mechanisms can also provide invaluable insight into how development progresses in general and have ample potential for medical translation.

In our lab we are interested in questions relating to the development of the *C. elegans* epidermis that revolve around the concepts presented above. Work on model organisms has informed most of our current understanding of development and stem cell biology (St Johnston, 2015) and we aim to contribute to that direction.

1.2 Introduction to *C. elegans*

1.2.1 General information and life-cycle

Caenorhabditis elegans is a free-living nematode that can be found in nature growing in rotting plant material by feeding on the flourishing microbe communities that decompose vegetation (Schulenburg & Félix, 2017). The lab reference strain N2 was isolated in Bristol, UK but since then *C. elegans* has been sampled around the globe in all continents, from Hawaii to New Zealand (Cook *et al.*, 2017). It was established as a model organism by Sydney Brenner, with remarkable foresight for its use in the study of genetics of development, neurobiology and behaviour (Brenner, 1973, 1974).

C. elegans possesses a series of traits that make it ideal as a model organism. Adult animals grow up to a size of ~1 mm and thus large populations can be easily cultivated in petri dishes, on agar based nematode growth medium (NGM), monoxenically by feeding on lawns of the uracil auxotroph *Escherichia coli* OP50 strain, that has limited growth allowing for clear observation of animals on the plate (Corsi, 2006; Stiernagle, 2006). Laboratory *C. elegans* populations are primarily made up of self-fertilising hermaphrodites, with each individual producing approximately 300 progeny (Brenner, 1974). The single-celled zygotes require approximately 3.5 days at 20 °C to reach adulthood, with the potential to decrease or increase developmental speed by shifting the populations of developing animals between 15 °C and 25 °C respectively (Corsi, 2006). Overall, the potential for quick growth of vast populations for experimentation is unparalleled.

The *C. elegans* life cycle, in presence of food, is made up of 4 distinct larval stages (L1-L4) post hatching (Figure 1.1). At each larval stage molting of the collagen-based cuticle takes place by shedding and replacing by a newly secreted one (Page, 2007). The larval stages can be confidently determined by developmental landmarks, such as the degree of gonadal arms extension (Altun & Hall, 2009). In the absence of food or under stress, hatched individuals can arrest their development at L1 or divert their development towards the life-cycle of the starvation and draught resistant dauer form (Golden & Riddle, 1984). Upon reintroduction of food or other stress relief, dauers exit to the L4 stage to grow and reproduce. In nature dauers are the most common form, likely serving survival between burst of population growth when food becomes available (Schulenburg & Félix, 2017).

As a final point, *C. elegans* are transparent permitting fine microscopic observations of any part of their anatomy. This has allowed the elucidation of the complete cell lineage of the developing *C. elegans,* from the zygote to the adult, demonstrating highly stereotypical patterns of cell divisions and an almost invariant terminal number of 959 somatic cells (Sulston *et al.*, 1983; Sulston & Horvitz, 1977; Kimble & Hirsh, 1979; Cunha *et al.*, 1999). This discovery has been an invaluable tool for anyone wishing to study development in *C. elegans*.
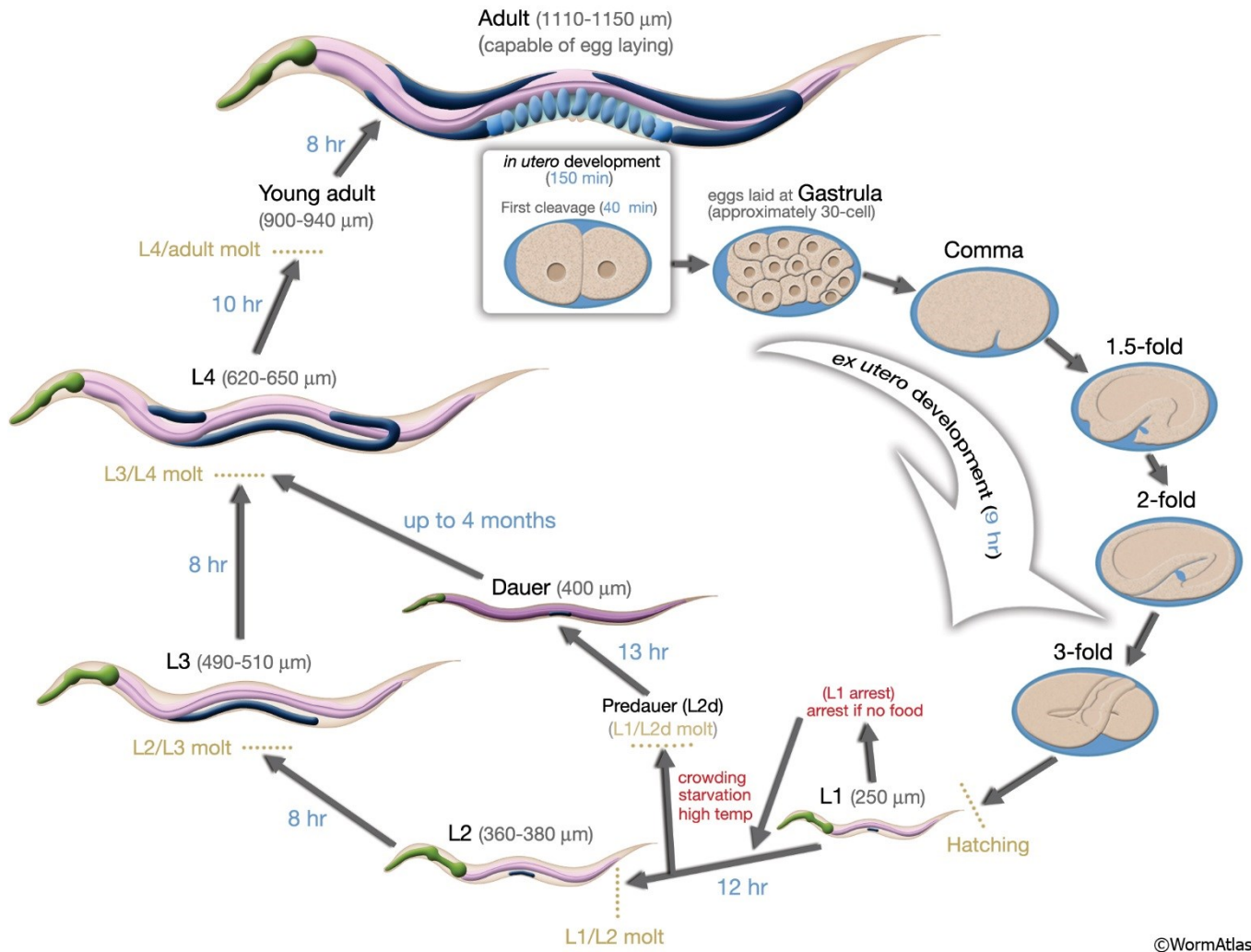
Figure 1. 1 Life cycle of *C. elegans* from zygote to adult at 22 °C Illustration of the main developmental stages of the *C. elegans* reproductive and dauer life cycle from the first cleavage in the egg to the fully developed adult. Numbering in blue indicates the time needed from one stage to the next and in parentheses is the normal size range in μm. The intestine is indicated in pink, the pharynx in green and the gonad in blue. The figure has been reproduced from WormAtlas (Altun *et al.*, 2020)

1.2.2 Genome and genetics of *C. elegans*

Aside of the ease of growth and manipulation there are also elements of *C. elegans* genetics that have made it the attractive model it is. The self-fertilising hermaphroditic mode of reproduction means that a population can be established by a single individual that will create a clonal isogenic population, a characteristic that allows the study of the effect of mutations or perturbations in general, in the absence of genetic variation in the population (Corsi, 2006). However, it does not preclude the capacity to perform genetic crosses to combine mutations or backgrounds of interest, as rare spontaneous males occur in populations (~1/1000) or protocols can be used to induce them (Corsi, 2006; Stiernagle, 2006). Sex-determination is dependent on the sex chromosomes (X) with hermaphrodites possessing two (XX), while males, due to non-disjunction during meiosis, have only one (XO)(Brenner, 1974).

A particularly significant trait is that *C. elegans* is amenable to transgenesis by simple microinjection (Evans, 2006). It can be transformed both transiently, with multi-copy transgenes and stably with genome-inserted multi-or single-copy transgenes, permitting the utilisation of a vast toolkit of available methodologies and experimentation approaches (Evans, 2006; Nance & Frøkjær-Jensen, 2019). Addiitionally, the capacity to grow large numbers of animals and screen for phenotypes with ease has facilitated the performance of countless powerful forward genetics screens to identify genes controlling phenomena of interest in *C. elegans* (Jorgensen & Mango, 2002).

The *C. elegans* haploid genome is approximately 100 Mega-bases (Mb) in size and is organised in 6 chromosomes: 5 autosomes (I, II, III, IV, V) and the sex chromosome (X). Both hermaphrodites and males are diploid for the autosomes and all chromosomes are holocentric (Spieth *et al.*, 2014). The genome encodes for 20191 protein-coding genes (WBcel235 assembly ensemble.org) approximately as many as the human genome with a size almost 31x smaller. The compactness of the *C. elegans* genome aids the discovery of regulatory regions controlling genes, as they tend to be proximal to their location (Gaudet & McGhee, 2010; Araya *et al.*, 2014).

Protein coding genes have a median size of ~2 kilo-bases (kb) and contain on average 6.4 exons (Spieth *et al.*, 2014; Tourasse, Millet & Dupuy, 2017). Approximately 15% of them are organised in operons and 70% have a 22-nucleotide leader sequence post-transcriptionally trans-spliced to the 5' of their transcript, while up to 94% of them encode more than one isoform (Tourasse, Millet & Dupuy, 2017). The genome also contains 24791 non-coding genes, 256 of which are annotated as miRNAs, a class of crucial post-transcriptional regulatory molecules that were discovered first in *C. elegans* (Lee, Feinbaum & Ambros, 1993; Reinhart *et al.*, 2000). Of the protein-coding genes ~41% have predicted orthologues in humans which is reciprocated by ~52.6% of the human genes having orthologues in *C. elegans* (Kim *et al.*, 2018). A large number of these orthologues participate in important conserved signalling pathways like TGF-β or Wnt, or are known to be implicated in human pathologies, illustrating how *C. elegans* research has the potential to lead to discoveries with direct medical translation (Apfeld & Alper, 2018).

1.3 Postembryonic development of the *C. elegans* epidermis: An overview of seam cell patterning

The *C. elegans* epidermis is ectodermal in origin and is a simple epithelium with an apical surface secreting and tightly affixed to the collagenous cuticle, while the basal is anchored to a basal lamina (Chisholm & Hsiao, 2012). Primary functions of the epidermis are to secrete the cuticle and together act as a tough barrier and exoskeleton as well as facilitate growth (Chisholm & Xu, 2012). A key cell type in the epidermis, which is the major focus of this work, are the lateral seam cells and their patterning is discussed below.

During embryogenesis the *C. elegans* epidermis is born entirely from progenitor cells of the AB and C lineages of the embryo (Sulston *et al.*, 1983). In particular, the lateral linearly-aligned seam cells of the

epidermis, arise exclusively from the AB lineage (Sulston *et al.*, 1983). In the epidermis of the newly hatched L1 larvae there are 10 seam cells on each lateral side (Figure 1.2, 1.3 top). Three of those occupy the head region, namely H0, H1 and H2 and extend from the anterior up to after the posterior pharyngeal bulb, six are in the midbody up to the rectum, namely V1, V2, V3, V4, V5, V6 and one on the tail, the T seam cell.

The syncytial hypodermis hyp7, which is born during embryogenesis, covers the largest area of the dorsal side and extends ventrally to surround the anteroposterior axis of the animal, running behind the seam cells on the medial side and forming two continuous rings, one in the anterior and one in the posterior (Figure 1.2A, B, C) (Altun & Hall, 2002). The anterior head region is covered by the syncytial hypodermal cells hyp1-hyp6, while the tail by hyp8-hyp11. The ventral side is largely occupied by the P-cells at L1 (Figure 1.2) a proportion of which will give rise to the vulva in following stages (Lints & Hall, 2004; Sulston & Horvitz, 1977).

The seam cells have an apical surface in contact with the cuticle, anchored and linked to the adjacent hypodermis also on the apical side with tight adherens junctions, while also being laterally connected by gap junctions (Michaux, Legouis & Labouesse, 2001). Seam cells are linked to the P-cells with adherens junctions as well. Apical junctions between cells can be visualised by labelling AJM-1, seen in figure 1.2D.

Throughout post-embryonic development the seam cells perform a series of stem cell-like divisions, outlined in figure 1.3. They divide both symmetrically to proliferate and asymmetrically, in a self-renewal manner, where one daughter differentiates while the other maintains the seam cell fate (Joshi *et al.*, 2010). Of all the seam cells the H0 is the only one that does not divide at any point but retains the seam cell fate until the end of postembryonic development. At L1, approximately 5 h post-hatching, seam cells divide asymmetrically. Divisions are not synchronous for all seam cells, with V5 dividing first, followed by V1-V4 and then the rest of the seam cells, a trend that persist in further larval stages (Gritti *et al.*, 2016). The anterior daughters of V1-V6 and the posterior daughter of H1 differentiate directly to hyp7 fate and fuse to the hyp7 syncytium mediated by the fusogen EFF-1 (Podbilewicz, 2006). The non-differentiating daughters retain the seam cell fate. For H2, the anterior daughter will divide once more and its posterior progenitor differentiates to the hyp7 while the anterior to a neuronal fate (Sulston & Horvitz, 1977). In the case of T, the posterior daughter differentiates towards a neuronal fate and the anterior divides once more to produce an anterior hyp7 daughter and a seam cell in the posterior.

The daughters that differentiate to hyp7 endoreduplicate, becoming tetraploid before fusing to hyp7, while the embryonic nuclei of the hyp7 syncytium are diploid (van Rijnberk *et al.*, 2017). The endoreduplication, which continues at the adult stage, is essential to facilitate growth in *C. elegans* (Lozano *et al.*, 2006). The hyp7 daughters of V2-V6 at L1 intercalate ventrally between the P cells, which lose their junctions with the seam cells and fuse to the dorsal hypodermis, creating a continuous ventral hyp7 and a complete cylindrical syncytium which covers most of the body (Altun & Hall, 2002). The seam cell-destined daughters of divisions extend processes along the anteroposterior axis to re-establish contacts between them, which are important for normal patterning (Austin & Kenyon, 1994).
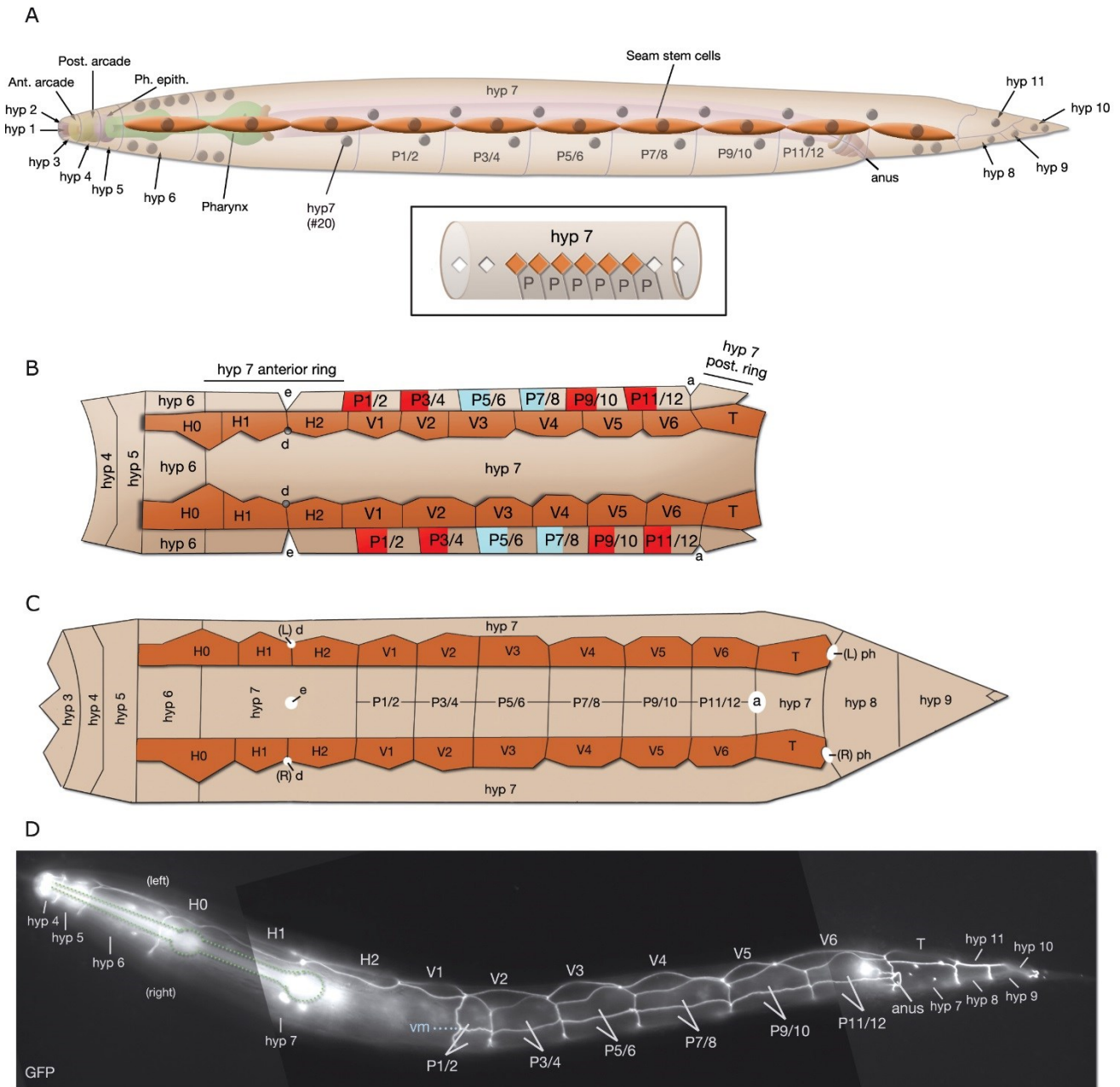
Figure 1. 2 Anatomy of the *C. elegans* epidermis at the L1 larval stage (A) Illustration of the whole L1 animal along its anteroposterior axis with a view of the left side, depicting the position of the 10 seam cells in orange (H0-H2, V1-V6, T). The hypodermal cells of the head (hyp1-hyp6), the tail (hyp8-hyp11), the dorsal hyp7 and the ventrolateral P cells are in beige. For the seam cells, P cells and hyp6-hyp11 the nuclei are indicated as brown circles. The connections between seam cells and P cells are also depicted below for clarity. **(B-C)** L1 epidermis sectioned at the ventral midline (B) or the at the dorsal midline (C) showing both seam cell lines extending from the junction between hyp6 and hyp5 in the anterior up to the junction between hyp7and hyp8 in the posterior. hyp3, hyp4, hyp5, hyp6 form rings around the head. P cells in (B) are labelled according to the fate their progenitors acquire with red depicting neurons, blue vulva and beige hyp7. Orifices of the epidermis are indicated by letters: (a) anus, (d) anterior deirid, (e) excretory pore, (ph) phasmid. Some head and tail hyp cells are omitted for clarity **(D)** Representative fluorescent image of L1 animals carrying the *ajm-1p::AJM-1:GFP* translational reporter indicating the apical junctions between seam cells, hypodermis and P cells. The epidermal cells visible are indicated, so is ventral midline (vm). The view is oblique ventral. The figure has been reproduced from WormAtlas (Altun *et al.*, 2020).

At the L2 stage the H1, V1-V4 and V6 seam cells undergo a proliferative symmetric division, generating two daughter cells each, that both maintain the seam cell fate, increasing the total number of seam cells to 16 per lateral side (Figure 1.3)(Sulston & Horvitz, 1977). The H1 anterior seam cell daughter does not divide again until the end of postembryonic development. The V1-V4 and V6 daughters divide once more asymmetrically after the symmetric division, with anterior daughters from all the divisions differentiating and fusing to hyp7, while posterior remain seam cells (Sulston & Horvitz, 1977). The T cell also performs a set of two divisions, generating two hyp7 daughters from the anterior second division and one seam cell and a neuroblast from the posterior. As in the case of the anterior H1, the T seam cell does not divide again but remains a seam cell until the end of postembryonic development. V5 performs two asymmetric divisions at L2. The anterior daughter of the first division acquires a neuronal fate and goes on to generate the posteirid (PDE) and PVD neurons, while the posterior divides again to produce a hyp7 and a seam cell daughter (Sulston & Horvitz, 1977). In contrast, H2 performs only one asymmetric division at L2, with the anterior daughter differentiating to hyp7. In the remaining two larval stages (L3 and L4),
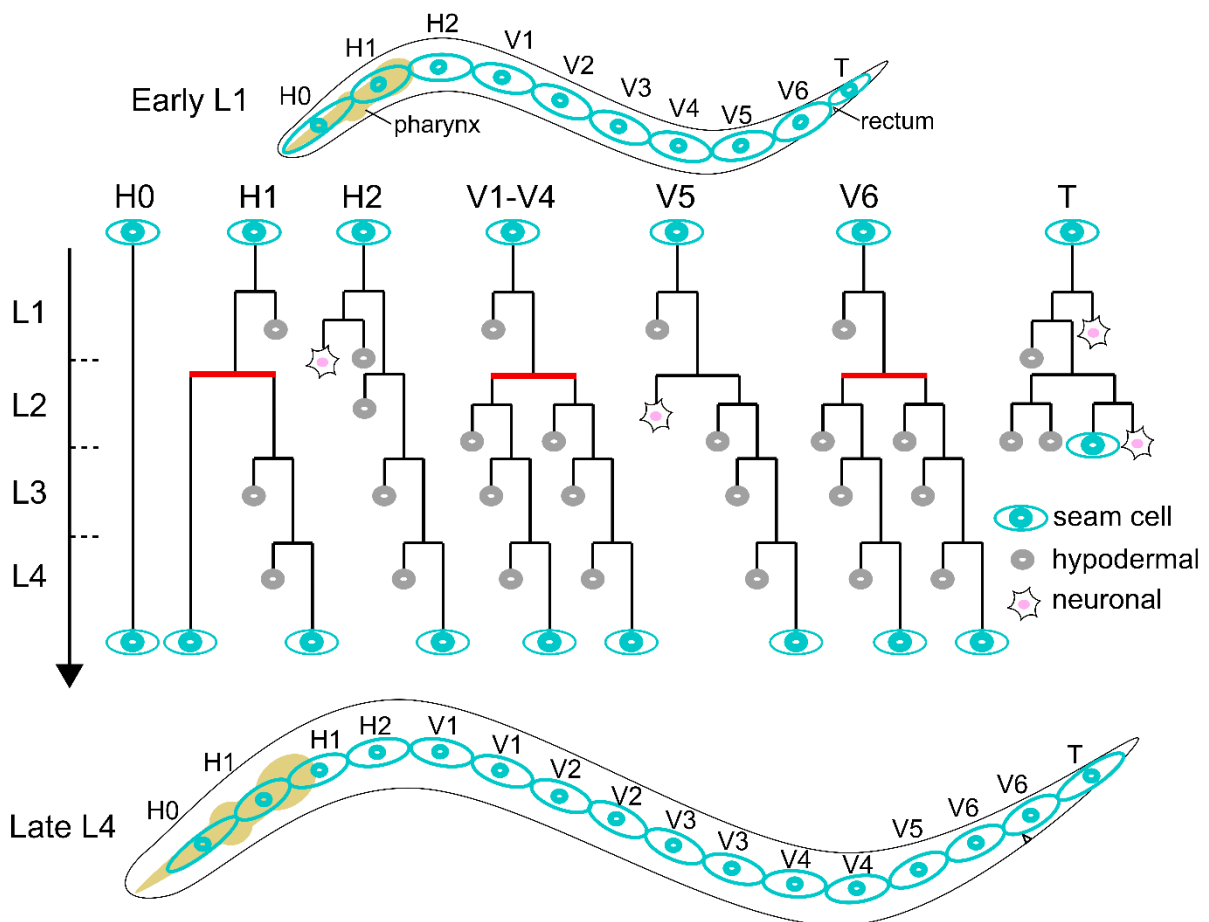
Figure 1. 3 Postembryonic development of the seam cells Illustration of the postembryonic lineages for each of the initial 10 seam cells of the hatched L1 animal, depicted on top, through to the 16 terminal seam cell number of the late L4 stage, depicted at the bottom. Seam cells are illustrated by teal eye-shapes, representing the membrane and nucleus, hypodermal daughters are grey and neuronal-fated daughters are shown in pink. Seam cells divide either symmetrically, denoted here by horizontal red lines or asymmetrically shown with black lines. The pharynx and rectum are indicated.

seam cells reiterate these asymmetric divisions once per stage, with the anterior daughters differentiating and fusing to hyp7 and the posterior remaining seam cells.

At the end of postembryonic development the terminal seam cell number is 16 per lateral side, which is a very robust phenotype with minimal variation in the population (Boukhibar & Barkoulas, 2016; Katsanos *et al.*, 2017) and it has been used in this study to probe developmental errors in seam cell patterning that accumulate throughout development.

As the animals are transitioning towards adulthood, the seam cells of each lateral side terminally differentiate and fuse together to form a syncytium, a process mediated by the fusogen AFF-1 (Sapir *et al.*, 2007; Podbilewicz, 2006). The syncytial seam cells secrete the alae which are longitudinal cuticular ridge-like structures along each lateral side of the animal, with a yet unclear function. Throughout postembryonic development the seam cells contribute 98 endoreduplicated nuclei to the hyp7 out of the 139 in total in the adult syncytium (Altun & Hall, 2002), largely mediating the potential for growth of the animal (Lozano *et al.*, 2006). The stem cell-like behaviour of the seam cells in combination with the ease of study in *C. elegans,* make them a very attractive *in vivo* model to investigate questions regarding fate maintenance, differentiation and symmetric/asymmetric divisions. This is further substantiated from the plethora of conserved factors and pathways that are known so far to be involved in the control of their development.

1.4 Genetic control of seam cell postembryonic development

1.4.1 Transcription factors participating in seam cell patterning

Several transcription factors and transcription factor families have been implicated with seam cell development and a large proportion of them are conserved, with orthologues participating in stem cell development in mammalian systems. Central to seam cell fate and patterning are GATA transcription factors. *C. elegans* possesses 11 GATA transcription factors, which take their name from their binding of GATA-centred DNA motifs, out of which ELT-1, EGL-18, ELT-6 and ELT-3 are involved in epidermal development (Block & Shapira, 2015).

ELT-1*,* orthologue of the human GATA1, is generally thought to be a master epidermal fate regulator and is necessary and sufficient to establish epidermal lineages during embryogenesis (Page *et al.*, 1997; Gilleard & Mcghee, 2001). One of its likely primary targets for activation in that process is ELT-3*,* another GATA factor that is also sufficient to specify the hypodermal fate and is only absent from the precursors of the seam cells (Gilleard & Mcghee, 2001; Chisholm & Hsiao, 2012). The *elt-1* expression persists postembryonically primarily in the seam cells where it is essential for seam cell fate determination and maintenance (Smith, McGarr & Gilleard, 2005; Katsanos *et al.*, 2017). In contrast, *elt-3* is expressed in the hypodermis and is considered to be a regulator of terminal differentiation towards the hypodermal fate (Gilleard & Mcghee, 2001; Block & Shapira, 2015).

ELT-1 is also thought to act upstream of the GATA factors *egl-18* and *elt-6*, potentially activating their expression in the seam cells (Koh & Rothman, 2001). EGL1-8 is a likely orthologue of GATA4, and along with its paralog, ELT-6, acts redundantly to specify seam cell fate (Koh & Rothman, 2001; Gorrepati, Thompson & Eisenmann, 2013). More specifically, they supress the hypodermal fate factor *elt-3* and the fusogen *eff-1* in the seam cells, activate various seam cell specific marker genes and during asymmetric divisions they are essential (particularly *egl-18*) for seam cell fate maintenance in the non-differentiating, usually posterior daughters of the divisions (Koh & Rothman, 2001; Gorrepati, Thompson & Eisenmann, 2013). This is achieved by the asymmetric activation of *egl-18* and *elt-6* expression, by the Wnt/β-catenin asymmetry pathway, only in the posterior daughters of asymmetric divisions that as a result acquire the seam cell fate (Gorrepati, Thompson & Eisenmann, 2013).

egl-18 has also been hypothesised to be activated in the seam cells by the homeodomain transcription factor CEH-16, the *C. elegans* orthologue of *engrailed* (Cassata *et al.*, 2005)*. ceh-16* is also expressed in the seam cells and their embryonal precursors and postembryonically is thought to be involved in the decisions between symmetric and asymmetric divisions (Cassata *et al.*, 2005; Huang *et al.*, 2009; Katsanos *et al.*, 2017). It potentially interacts with the Wnt signalling pathway in mediating the symmetric division of L2, likely through activation of *egl-18* and can be functionally substituted by the human orthologue *En2* (Huang *et al.*, 2009). *ceh-16*, as multiple of the above factors, has been proposed to be activated by ELT-1 but so far concrete evidence is lacking (Cassata *et al.*, 2005; Thompson *et al.*, 2016).

In contrast, ELT-1 has been shown to directly activate the expression of *bro-1* in the seam cells (Brabin, Appleford & Woollard, 2011). BRO-1 is the *C. elegans* homologue of CBFβ that forms a complex with the homologue of Runx, RNT-1 to promote the seam cell fate (Kagoshima *et al.*, 2007; Nimmo, Antebi & Woollard, 2005). Specifically, both factors of the complex have been shown to be expressed in the seam cells and cause seam cell hyperplasia if overexpressed, which is in keeping with the context-dependent tumorigenic capacity of the complex in humans and its role in cancer (Kagoshima *et al.*, 2007; Nimmo & Woollard, 2008; Cameron & Neil, 2004). The RNT-1/BRO-1 complex is required for the correct execution of the L2 symmetric division, by overwriting the asymmetry established by the Wnt signalling, through suppression of the TCF homologue POP-1 (van der Horst *et al.*, 2019). It is also thought to promote proliferation by supressing the cell cycle inhibitor *cki-1* (Nimmo, Antebi & Woollard, 2005; Kagoshima *et al.*, 2007). The seam cell promoting potential of *rnt-1* is thought to be supressed within the domain of the differentiating daughter of an asymmetric seam cell division by the transcription factors CEH-20 and UNC-62, homologues of Pbx and Meis respectively, which repress *rnt-1* expression thus acting as regulators of division asymmetry (Hughes *et al.*, 2013).

Of the much expanded family of more than 250 nuclear receptor (NRs) transcription factors of *C. elegans,* multiple are expressed in the seam cells, but it is NHR-25 that has been found to be the most consequential for postembryonic epidermal development, along with NHR-23 for molting (Miyabayashi *et al.*, 1999; Gissendanner & Sluder, 2000; Chisholm & Hsiao, 2012). This is likely an effect of functional

redundancy between the members of the expanded group (Miyabayashi *et al.*, 1999; Koh & Rothman, 2001).

NHR-25 is the *C. elegans* orthologue of the well-studied *fushi tarazu* transcription factor 1 (FTZ-F1) of *Drosophila* that acts on aspects of cuticle formation (Gissendanner & Sluder, 2000). NHR-25 is pleiotropic, participating in molting and vulva development along with seam cell patterning (Gissendanner & Sluder, 2000; Chen, Eastburn & Han, 2004). Animals with *nhr-25* silencing or mutations show additional seam cells and NHR-25 is thought to exert its developmental role in seam cell patterning by mediating the re-establishment of cell-to-cell contacts between seam cells and fusion of the differentiating daughters to hyp7 after divisions (Chen, Eastburn & Han, 2004; Šilhánková, Jindra & Asahina, 2005; Hajduskova *et al.*, 2009). It is also possible that it acts directly on the differentiation process as it has been shown to do in the T seam cell in collaboration with Wnt signalling (Hajduskova *et al.*, 2009).

Another TF that participates in seam cell patterning is the Hes-related bHLH factor LIN-22, homologue of the mammalian Hes1. Hes factors have been implicated in multiple developmental events and stem cell behaviour in mammalian systems, often controlling differentiation decisions between epithelial and neuronal fates as well as proliferation (Kageyama & Ohtsuka, 1999; Murata *et al.*, 2005; Kobayashi *et al.*, 2009; Kay *et al.*, 2017). In *C. elegans, lin-22* is expressed specifically in the seam cells and was first studied in the context of neurogenesis, for its ability to supress ectopic formation of PDE neurons in V1-V4 seam cells (Wrischnik & Kenyon, 1997; Katsanos *et al.*, 2017). We recently found that LIN-22 regulates seam cell patterning by also antagonising Wnt signalling, likely by supressing the *frizzled* receptor *lin-17,* to establish correct division patterns (Katsanos *et al.*, 2017).

For the factors presented above there are already known links of mainly genetic interactions, most of which were described above, which form simplified core networks, like those presented in (Koh & Rothman, 2001; Thompson *et al.*, 2016), that attempt to explain the genetic mechanism underlying epidermal patterning. It is still unclear which of these interactions are truly direct and how many factors make up the complete gene network.

1.4.2 The Wnt/β-catenin asymmetry pathway controls the polarity of asymmetric seam cell divisions

In the seam cell postembryonic development the polarity of the self-renewal asymmetric divisions that the seam cells undergo is dictated by Wnt signalling. More specifically, by a divergent version of the canonical Wnt signalling pathway called Wnt/β-catenin asymmetry pathway (Wβa) (Sawa & Korswagen, 2013; Gleason & Eisenmann, 2010).

The conserved canonical Wnt signalling, which also acts in *C. elegans,* utilises the β-catenin BAR-1, which in the absence of signal is continually targeted for degradation by the destruction complex (Sawa & Korswagen, 2013; Gleason & Eisenmann, 2010). The destruction complex is made up of the conserved *C. elegans* homologues of Axin, APC, GSK3 and CK1, namely PRY-1, APR-1, GSK-3 and KIN-19,

respectively, which capture and phosphorylate BAR-1 resulting to degradation (Sawa & Korswagen, 2013). In the case of activation of the pathway by binding of one of the Wnt ligands (MOM-2, LIN-44, EGL-20, CWN-1, CWN-2) by one of the transmembrane *frizzled* receptors (MOM-5, LIN-17, MIG-1, CFZ-2), a *C. elegans* orthologue of Dishevelled (MIG-5, DSH-1, DSH-2) sequesters the destruction complex, preventing degradation of BAR-1, which translocates to the nucleus and forms a complex with the TCF orthologue POP-1 to activate expression of Wnt target genes. The canonical pathway activity via BAR-1-mediated gene activation was not thought to have a significant role in seam cell patterning but may play a role within the seam cells upon temperature increase (Hintze *et al.*, 2020).

In contrast, the Wβa pathway that is essential for correct seam cell patterning, is reliant on two divergent β-catenins, WRM-1 and SYS-1 and does not rely solely on stabilization of β-catenin in signalled cells (Sawa & Korswagen, 2013; Gleason & Eisenmann, 2010). Rather it is dependent on asymmetric distribution of Wnt components within dividing cells, with the final output being the regulation of the relative levels of nuclear POP-1 and SYS-1 in the dividing daughters. POP-1 individually acts as a reppressor, whereas in complex with SYS-1 as an activator of gene expression and the relative ratio between the two determines the effect on gene expression and is controlled by Wnt signalling (Takeshita & Sawa, 2005; Mizumoto & Sawa, 2007a; Sawa & Korswagen, 2013; Lam & Phillips, 2017; Lin, Hill & Priess, 1998).

The orientation of the asymmetry of segregation of the components is thought to be instructed by the positional information of Wnt, acting alongside intrinsic distribution mechanisms. This is demonstrated by the seam cells of animals lacking all Wnt ligands that still perform asymmetric divisions but with randomised polarity (Yamamoto, Takeshita & Sawa, 2011). In contrast if internal components like the frizzled receptors are lost, seam cells fail to divide asymmetrically undergoing symmetric divisions ectopically (Yamamoto, Takeshita & Sawa, 2011).

The five Wnt ligands that act highly redundantly in controlling the polarity of the asymmetric seam cell divisions, function in both pathways and are expressed in partially overlapping domains along the anteroposterior axis of the animal. *lin-44, egl-20* and *cwn-1* are primarily expressed in the posterior, while *mom-2* and *cwn-2* are expressed in more anterior tissues (Harterink *et al.*, 2011; Yamamoto, Takeshita & Sawa, 2011). A Wnt antagonist, the secreted frizzled-related protein *sfrp-1,* is also expressed at the head of the animal (Harterink *et al.*, 2011). The secreted protein of the Wnt ligand EGL-20 has been shown to disperse extracellularly from the expressing cells, forming an anteroposterior gradient, with higher occurrence in the posterior of the animals (Pani & Goldstein, 2018). In combination with the above expression domains it is likely that a Wnt activating potential that is stronger in the posterior of the animal and weaker in the anterior, as it has been documented by Wnt activity reporters, provides the orientation cues in Wnt signalling events (Harterink *et al.*, 2011; Sawa & Korswagen, 2013; Bhambhani *et al.*, 2014).

The above posterior positional cues by Wnt, drive the localisation of the β-catenin WRM-1 to the anterior seam cell cortex along with LIT-1/Nemo-like kinase prior to division (Takeshita & Sawa, 2005). The cortical WRM-1 acts to recruit APR-1 and PRY-1 of the destruction complex to the anterior cortex and APR-1 reciprocally modifies the cytoskeleton to promote export of nuclear WRM-1 from the anterior

nucleus during telophase (Mizumoto & Sawa, 2007a; Sugioka, Mizumoto & Sawa, 2011). In addition, the anterior localisation of components of the destruction complex also causes degradation of SYS-1 in the anterior, resulting to asymmetric localisation to the posterior nucleus. In contrast, Wnt signalling drives positive regulators of the Wβa pathway like *frizzled* receptors (MIG-1, LIN-17) and *dishevelled* (MIG-5, DSH-2, DSH-1) to localise to the posterior cortex (Mizumoto & Sawa, 2007a; Sugioka, Mizumoto & Sawa, 2011; Sawa & Korswagen, 2013; Mizumoto & Sawa, 2007b). By receiving Wnt signals they act to prevent destruction of SYS-1, increasing its concentration in the posterior, while also both facilitating the disengagement of WRM-1/LIT-1 complexes from cortical PRY-1/APR-1 and inhibiting the WRM-1/LIT-1 export from the posterior nucleus (Mizumoto & Sawa, 2007b, 2007a; Sawa & Korswagen, 2013; Lam & Phillips, 2017). The increased nuclear WRM-1/LIT-1 leads to phosphorylation of POP-1 that promotes its export from the nucleus, substantially decreasing its nuclear levels in comparison to the anterior nucleus, while increasing the relative levels of SYS-1 to POP-1 in the posterior (Mizumoto & Sawa, 2007b; Sawa & Korswagen, 2013; Lam & Phillips, 2017; Gleason & Eisenmann, 2010).

Overall, these signalling cascade events have the following outcome: the anterior daughter cell nucleus has low levels of WRM-1 allowing high levels of nuclear POP-1, relative to SYS-1, leading to repression of Wnt target genes. In the posterior, the lower levels of POP-1 in combination with the high levels of SYS-1 mediates the formation of SYS-1/POP-1 complexes that activate Wnt target genes (Lam & Phillips, 2017; Sawa & Korswagen, 2013; Mizumoto & Sawa, 2007b). *egl-18* and *elt-6* are such target genes activated in the posterior daughters of the asymmetric divisions of seam cells as a result of Wnt signalling (Gorrepati, Thompson & Eisenmann, 2013; Gorrepati *et al.*, 2015). Both are major seam cell fate regulators thus dictating the acquisition of the seam cell fate by the daughter that expresses them (even ectopically) whereas the typically anterior daughter lacks the maintenance signal and differentiates (Gorrepati, Thompson & Eisenmann, 2013; Katsanos *et al.*, 2017).

1.4.3 Temporal control of seam cell patterning by the heterochronic pathway

Seam cell divisions occur almost simultaneously and reiterate at every larval stage. The timing of the seam cell divisions is controlled by the heterochronic pathway (Ambros & Horvitz, 1984). The heterochronic pathway is highly dependent on primarily two micro-RNAs (miRNAs), namely *lin-4* and *let-7,* that were the first of the class to be studied and were found in the process of working out the pathway and its role in *C. elegans* epidermal development (Ambros & Horvitz, 1984; Lee, Feinbaum & Ambros, 1993; Liu, Kirch & Ambros, 1995; Reinhart *et al.*, 2000). These miRNAs are subject to developmental control of expression and act to post-transcriptionally supress the expression of protein-coding heterochronic genes (*lin-14, lin-28, lin-41, lin-29, daf-12, hbl-1*), by targeting their 3' UTRs to supress translation or degrade the transcript (Nimmo & Slack, 2009; Slack & Ruvkun, 1997).

The *lin-4/mir-125* and *let-7* miRNA families are highly conserved, acting in similar ways in various systems (Slack & Ruvkun, 1997; Nimmo & Slack, 2009). In controlling the timing of seam cell divisions *lin-4* acts to allow animals to progress to the performance of the L2 and L3 stage divisions, while *let-7* ensures

the termination of the reiterative divisions after the final division at L4 (Slack & Ruvkun, 1997; Nimmo & Slack, 2009; Reinhart *et al.*, 2000; Lee, Feinbaum & Ambros, 1993).

More specifically, animals hatch with high levels of LIN-14 and LIN-28. LIN-14 is required for the asymmetric division of the L1 stage, while it also supresses the L2 symmetric division. Loss of *lin-14* coverts the L1 asymmetric to a symmetric division, whereas overexpression has the opposite effect on the L2 symmetric division (Chalfie, Horvitz & Sulston, 1981; Ambros & Horvitz, 1984). Likely through an onset of feeding signal, *lin-4* expression levels increase towards the end of L1 and supress firstly LIN-14 and then LIN-28 (Slack & Ruvkun, 1997; Lee, Feinbaum & Ambros, 1993). The suppression of LIN-14 permits the execution of the L2 symmetric division which requires LIN-28 to occur, as *lin-28* mutants skip it (Slack & Ruvkun, 1997; Nimmo & Slack, 2009; Ambros & Horvitz, 1984). The suppression of LIN-14 releases *daf-12* expression that further suppresses LIN-28 by the end of L2 to allow progression to L3 fates and to permit the gradually increasing expression of *let-7* by L4 (Slack & Ruvkun, 1997; Nimmo & Slack, 2009).

The *let-7* miRNA at L4 supresses its target LIN-41, which in turn releases the TF *lin-29* from suppression, allowing its expression, which drives terminal differentiation of the seam cells and their fusion to a syncytium, by activating the fusogen *aff-1* (Rougvie & Ambros, 1995; Friedlander-Shani & Podbilewicz, 2011; Reinhart *et al.*, 2000; Nimmo & Slack, 2009). This terminates the division patterns of seam cells that otherwise in the absence of *let-7* continue reiterative divisions into adulthood (Reinhart *et al.*, 2000; Nimmo & Slack, 2009).

Components of the heterochronic pathway have been found to reciprocally genetically interact with the Wnt pathway asymmetry machinery. Interactions with LIT-1/POP-1/APR-1, regulate the stage specific decisions and performance of symmetric or asymmetric divisions, while PRY-1/Axin has been shown to control expression of miRNAs of the heterochronic pathway (Harandi & Ambros, 2015; Mallick, Ranawade & Gupta, 2019).

1.5 Open questions in the seam cell development model system

The postembryonic epidermal development of *C. elegans*, with particular focus on the patterning of the seam cells, is an ideal model to approach questions regarding stem cell behaviour, decisions between asymmetric and symmetric divisions, fate determination, maintenance and differentiation. Such questions broadly encompasse most areas of interest within the field of developmental biology (Chisholm & Hsiao, 2012; Joshi *et al.*, 2010; Brabin & Woollard, 2012). The ease of *C. elegans* cultivation in large numbers and the isogenic nature of its populations, in combination with the naturally highly invariant and fully mapped seam cell lineage, make it ideal for the thorough study of genotype-to-phenotype mechanisms driving developmental programmes like those outlined above. The largely conserved genetic control means findings made utilising this model can have far reaching implications. Therefore, dissecting the mechanisms that govern epidermal development and seam cell patterning has been the focus of our lab and others (cited throughout the previous sections).

In the Barkoulas Lab we have performed forward genetic screens to identify novel actors participating in seam cell development by isolating mutations that perturb the terminal seam cell number (Katsanos *et al.*, 2017; Boukhibar & Barkoulas, 2016). Genetic screens have generally been the methodology of choice for various areas of interest, since Sydney Brenner demonstrated their strength in *C. elegans* (Jorgensen & Mango, 2002; Brenner, 1973). The nature of the screen is such that we expect to find and have found TFs, involved in seam cell development, to add to those previously identified by others (Katsanos *et al.*, 2017).

Until recently, working out how these various components form gene networks and interact in carrying out a developmental programme, was predominantly based on assaying genetic interactions. Studying epistasis, employing reporter analysis and the ease of RNAi by feeding that allows combinations of mutations and knockdowns to be investigated have been the main tools of elucidation of genetic mechanisms. These experimental approaches can be particularly laborious and often rely on serendipity in identifying functional interactions. The advent of CRISPR-Cas9 (Clustered Regularly-Interspaced Short Palindromic Repeats – CRISPR-associated nuclease) and methods like single-molecule fluorescent *in situ* hybridisation (smFISH) have aided substantially in the discovery of molecular interactions between factors controlling seam cell development (examples in (Katsanos *et al.*, 2017)). Based on such methods small networks of largely genetic interactions between TFs and putative target genes, have been proposed in the case of seam cell development (Koh & Rothman, 2001; Thompson *et al.*, 2016; Chisholm & Hsiao, 2012).

Published genetic networks have been somewhat informative, but do not yet capture the full complexity of seam cell patterning. Firstly, currently available networks are rather small. Their expansion with conventional approaches requires either isolation of new mutants, or in the case of known genes with a seam cell mutant phenotype, a series of trial and error experiments. These will have to survey various potential genetic interactions with known factors to potentially suggest a position within the network. The resulting findings can only indicate genetic interactions lacking any resolution to distinguish between direct and indirect regulatory relationships. Thus they are prone to missing multiple components of the network that more precisely describe how a developmental event is regulated.

Resolution is also lacking with most available experimental approaches regarding the spatial specificity of genetic interactions between factors participating in seam cell development. Tissue-specificity is crucial to be able to clarify if a factor is involved in a developmental phenomenon in a cell-autonomous manner. It is also critical when factors with systemic roles are studied. It is required to understand how such a factor acts specifically within a particular tissue of interest to regulate a developmental event.

Current methodologies also present limitations to a more holistic understanding of the regulation of seam cell patterning, as they have only allowed the discovery of limited points of interface between the TF network and signalling pathways (e.g. Wβa and *egl-18*), which are unlikely to fully encompass the full spectrum of interactions between pathways and gene regulatory networks. Moreover, little is currently known about the role of epigenomic regulation in seam cell patterning.

All the layers of complexity described above are currently chanllenging to unravel with conventional methodologies. Thus, addressing questions like: what are the differences between seam cells and hypodermis that are instructive to the specification of these different cell-types of the epidermis, is hard with only the tools available. To be able to approach a lot of these open questions and work towards a more holistic understanding of the regulatory and gene expression state that governs epidermal development, methodologies capable of elucidating tissue-specific protein-DNA interactions, gene expression profiles and chromatin states would have to be employed. With the rise of genomics deep sequencing-based technologies, this information can be acquired *en masse* and in great detail and has already facilitated discoveries in developmental and stem cell biology (St Johnston, 2015; Zeitlinger & Stark, 2010; Cañestro, Yokoi & Postlethwait, 2007). In this study, I present the first application in *C. elegans* of the targeted DamID (TaDa) methodology to begin addressing aspects of all the above questions, regarding the genetic mechanistic underpinnings of postembryonic epidermal development, using this single multi-faceted method.

1.6 Targeted DamID: a versatile tool to study protein-DNA interactions

1.6.1 Basic principle and experimental design

Methylation of DNA is pervasive in life but the types of DNA methylation in eukaryotes and prokaryotes differ (Law & Jacobsen, 2010). The adenine methylation is prevalent in many prokaryotes and a common such modification is the addition of a methyl group on the N6 position of the adenine (6mA) of a GATC sequence by the *Escherichia coli* DNA adenine methyltransferase (Dam) enzyme (Sánchez-Romero, Cota & Casadesús, 2015). Adenine methylation is in general mostly absent from eukaryotes with only a few examples of very low levels of 6mA methylation, which in the case of *C. elegans* occurs predominantly in AGAA or GAGG sequences (Greer *et al.*, 2015; Zhang *et al.*, 2015; Law & Jacobsen, 2010). This distinction between types of endogenous methylation in different domains of life and more specifically the absence of GATC 6mA was taken advantage by van Steensel and Henikoff in an attempt to identify protein-DNA interactions and they established the DNA adenine-methyltransferase identification technique (DamID) (van Steensel & Henikoff, 2000).

DamID is based on the fusion of any protein of interest (POI) that is expected to interact with DNA with Dam and expression of the fusion *in vivo.* The fusion will be recruited at the genomic sites where the POI binds or interacts with DNA, either through sequence affinity, preference for chromatin state or proximity, allowing Dam to add methyl groups to GATC sequences in the vicinity (van Steensel & Henikoff, 2000; van Steensel, Delrow & Henikoff, 2001) (Figure 1.4). GATC 6mA DNA fragments can be isolated *in vitro* from the rest of the genome by restriction digest, catalysed by the enzyme DpnI that only cleaves methylated GATC sequences (van Steensel & Henikoff, 2000). The extracted sequences are usually PCR

amplified and can be identified by protocols that nowadays are mostly dependent on next-generation sequencing (NGS) to create genome-wide maps of the POI-Dam targeted loci (van Steensel, Delrow & Henikoff, 2001; Wu, Olson & Yao, 2016; Aughey & Southall, 2016).
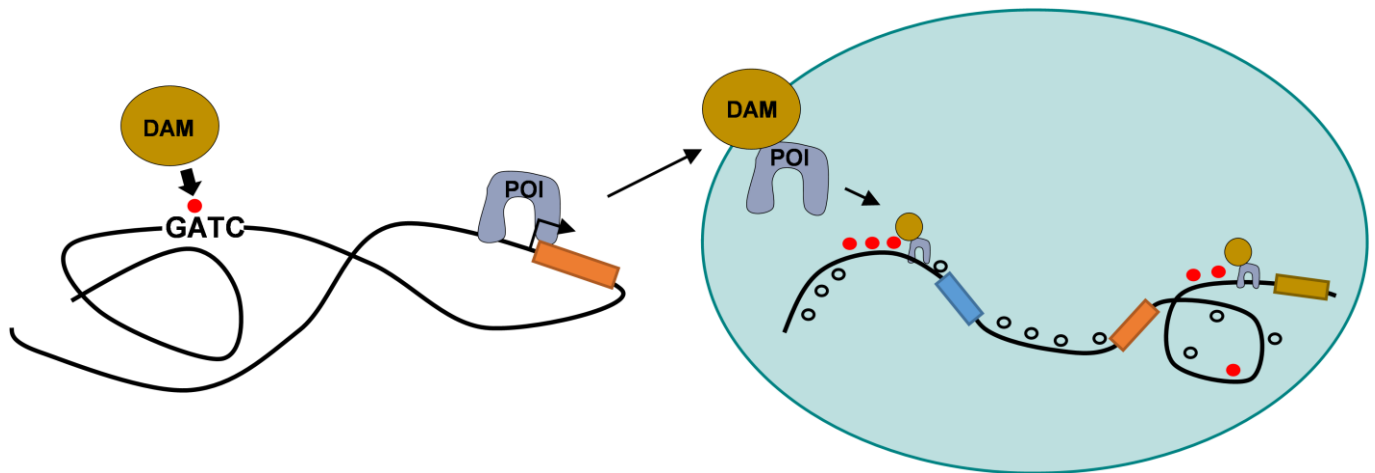
Figure 1. 4 General principle of DamID as a tool to discover protein-DNA interactions Illustration of the key aspects of DamID and the principle of protein-DNA interaction identification by proximity-based methylation. On the left DAM protein adds a methyl group on the adenine of the GATC sequence and a protein of interest (POI) (e.g. a transcription factor) interacts with DNA in a site-specific manner. On the right a fusion between DAM and the POI expressed *in vivo* primarily methylate regions of the genome proximal to where the POI interacts with DNA while others sequences remain relatively un-methylated.

From the original DamID experiments it became apparent that the levels of expression of the Dam-fusions are crucial for the successful identification of POI interactons with DNA (van Steensel & Henikoff, 2000). Constitutive or tissue-specific promoters would produce levels that would saturate DNA with non-specific methylation, preventing identification of genuine targets, while also often causing toxicity associated either with the high levels of the protein or misregulation as a result of the excessive methylation of DNA (van Steensel & Henikoff, 2000; Southall *et al.*, 2013). This obstacle was overcome by the use of low-level basal ("leaky") but ubiquitous expression, driven primarily from uninduced heat-shock promoters (van Steensel & Henikoff, 2000; van Steensel, Delrow & Henikoff, 2001). It should be noted that even at that level, Dam-fusions produce high levels of background methylation, which is thought to reflect chromatin accessibility. Based on that feature, applications employing Dam to probe chromatin structure preceded DamID (Wines *et al.*, 1996). Therefore, a complete DamID experiment utilises a control sample of either untethered or non-targeted Dam-fusion-equivalent, to permit normalisation and removal of the background methylation (Aughey & Southall, 2016).

This conventional form of DamID has been used in multiple organisms but most widely in *Drosophila*, with only one application in *C. elegans* so far for the identification of targets of the TF DAF-16/FoxO using tilling arrays as means of identification (Schuster *et al.*, 2010). The inventive solution of leaky expression of the Dam-fusion from uninduced conditional promoters is also the most significant general limitation of DamID, in that it lacks tissue-specificity as the expression is most often spatially ubiquitous. This can be overcome with recombinase-based systems like FLP/FRT or CRE/lox that require a recombination step to

permit expression from the uninduced promoter within some portion of the tissue of interest, as this approach usually produces mosaics (Aughey, Cheetham & Southall, 2019; Pindyurin *et al.*, 2016; Branda & Dymecki, 2004; Lin & Scott, 2012). The alternative for robust cell-type specific DamID, which is also employed in this study to express Dam-fusions in the *C. elegans* epidermis, is targeted DamID or TaDa (Southall *et al.*, 2013).

TaDa relies on a special design of the Dam-fusion expression transgenes to produce appropriately low-levels of tissue or cell-type-specific expression. It achieves this by utilising a tissue-specific promoter with the desired spatiotemporal expression domain, to drive expression of a bicistronic mRNA made up of two open reading frames (ORFs) interrupted by two stop codons and a frameshift (Southall *et al.*, 2013) (Figure 1.5). The Dam-fusion occupies the secondary ORF which is translated very infrequently due to ribosomal reinitiation, resulting to very low protein levels (Kozak, 2001; Southall *et al.*, 2013). The frequency of the reinitiation is dependent on the size of the primary ORF (Kozak, 1987, 2001), with the length of *mCherry* found to be suitable for the appropriate expression levels (Southall *et al.*, 2013).
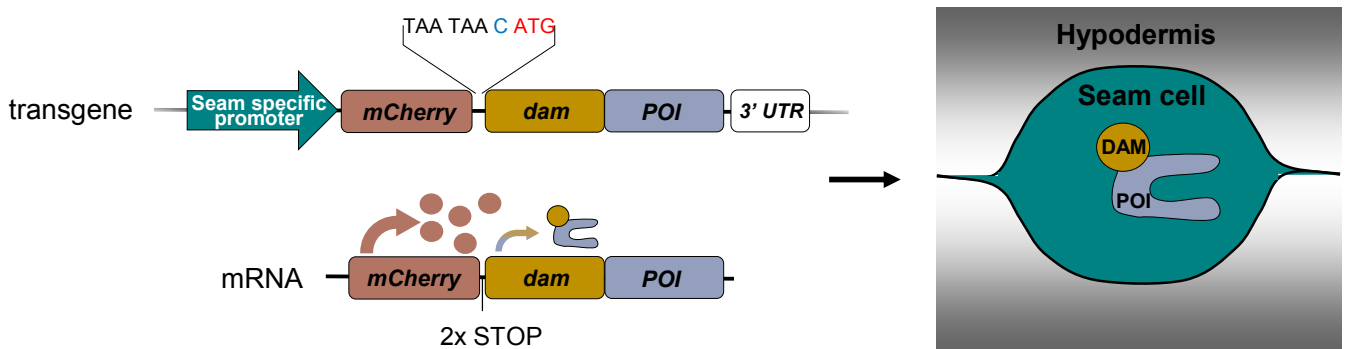
Figure 1. 5 Molecular design and basic mechanism underlying targeted DamID Illustration of the basic features of a system for seam cell-specific DamID. The design of such a system includes a seam cell-specific promoter followed by a primary ORF, in this case *mCherry*, then two STOP codons, a frameshift nucleotide in blue and the coding sequence of the Dam-POI fusion (First ATG codon in red) followed by a 3' UTR. Expression from the transgene produces a bicistronic mRNA which in turn produces very low levels of seam cell-specific DAM-POI protein by rare ribosomal reinitiation of translation at the secondary ORF.

TaDa therefore allows for tissue-specificity, creating the potential for experimental designs that produce information unattainable by conventional DamID or other methods without cell-isolation and can permit comparative assessment of interactions for sets of tissues or cell-types of interest like the seam cells and the hypodermis in our case. These include transcription factor target identification, gene expression profiling, chromatin accessibility profiling, DNA-nuclear lamina interactions identification, chromatin modifiers target identification, transcription factor co-binding or co-factor target identification, long range DNA interaction and even RNA-DNA interaction sites identification (Aughey & Southall, 2016; Aughey, Cheetham & Southall, 2019). In this study, the first 3 of the above applications of TaDa are for the first time performed in *C. elegans*.

1.6.2 TaDa for transcription factor target identification

One of the main applications that DamID has seen is in the identification of TF targets and TaDa can offer the same capability with the added benefit of tissue-specificity. The TF of interest is fused to Dam in an orientation that is unlikely to obstruct its DNA binding domain, while Dam is effective in both N- and C-terminal fusions (Ramialison *et al.*, 2017). The fusion is then transgenically expressed within the tissue of interest to identify the binding of the TF in question, only within that tissue, providing potential key advantages.

However, in *C. elegans* TF target identification has been almost exclusively performed by chromatin immunoprecipitation (ChIP) experiments (Solomon & Varshavsky, 1985), mostly followed by deep sequencing (ChIP-seq), with only a single case of DamID-chip for DAF-16 (Kudron *et al.*, 2018; Schuster *et al.*, 2010). Nevertheless, TaDa can be comparable to the more established ChIP-seq while also conferring some advantages with a brief comparison following (DamID vs ChIP is thoroughly reviewed in (Aughey & Southall, 2016; Aughey, Cheetham & Southall, 2019; Askjaer, Ercan & Meister, 2014)). Firstly, ChIP demands high affinity antibodies which can be hard to produce and in the most optimal case of epitope-knock-in for precipitation, the altered endogenous protein might behave differently. A TaDa advantage is that target identification is performed *in vivo* by the methylation labelling of DNA, while ChIP requires chemical cross-linking and *in vitro* separation of bound DNA that can potentially introduce artefacts (Teytelman *et al.*, 2013). The *in vivo* expression of the fusion in TaDa excludes systems where transgenesis is not possible but *C. elegans* is highly amenable to it. TaDa can capture transient interactions between TFs and DNA because of the covalent nature of the labelling, whereas ChIP requires oversampling to capture rare interactions (Aughey & Southall, 2016). This is also a major difference between the two methods as DamID in general can produce binding profiles from as few as 30 μl of *C. elegans* animal pellet, in stark contrast to the 1-2 ml required for ChIP (Askjaer, Ercan & Meister, 2014).

ChIP on the other hand is thought to have better resolution than TaDa for which resolution is dependent on availability of GATC sites and the size of the fragments. In addition, TaDa requires time for the expression of the fusion before sufficient methylation for detection has occurred. In fast-cycling tissues DNA replication converts methylated GATCs to hemi-methylated that are harder to detect. ChIP is not impeded by these limitations with sufficient sampling (Aughey & Southall, 2016; Askjaer, Ercan & Meister, 2014). Nonetheless, overall the two methods have been shown to produce highly comparable data (Southall *et al.*, 2013; Cheetham *et al.*, 2018).

Most crucial differences relate to the tissue-specificity allowed by TaDa which permits the identification of TF targets only within a tissue-of interest without the cell isolation that ChIP would require. In principle ChIP could also be performed using epitope tagged tissue-specific transgenes but high expression levels of TFs could be fate-changing in developmental systems, an issue not encountered in TaDa due to the low expression levels. This can permit the fragmentation of the endogenous expression domain of a TF to identify cell-type relevant targets and lead to a better understanding of what regulatory

events are potentially important for a tissues functions or identity. This feature of TaDa has been utilised in *Drosophila* neuronal development to identify cell-type-specific targets of the TF Hunchback (Sen *et al.*, 2019). TF target identification by TaDa can aid in the discovery of new genes participating in the seam cell developmental gene network and improve the "wiring" of the network by elucidating direct regulatory interactions that take place specifically within the epidermis.

1.6.3 Assaying gene expression using TaDa

TaDa has also made tissue-specific gene expression profiling possible, by tracking the gene sequences that associate with an RNA polymerase (RNApol) subunit fused with Dam. Therefore, it permits the identification of differences between the batteries of genes expressed in tissues or cell-types of interest, like the seam cells and the hypodermis in our case. Applications in *Drosophila* neuronal cell types have identified tissue-specifically expressed genes (Southall *et al.*, 2013).

Similar to above the novelty in this application relates to the tissue-specificity that is readily achieved in TaDa. Cell isolation is required to obtain tissue-specific transcriptomes using more traditional RNA-sequencing (RNA-seq) approaches (Celniker *et al.*, 2009; Spencer *et al.*, 2011; Kaletsky *et al.*, 2018). Cell isolation is particularly difficult in *C. elegans* especially for cuticle-associated tissues like the epidermis (Zhang & Kuhn, 2013). In combination with cell-type selection by methods like fluorescent activated cell-sorting (FACS), the total recovery of tissue is extremely low and requires very large amounts of starting material for sufficient mRNA extraction (Spencer *et al.*, 2014). In contrast, as stated above TaDa is expected to generate expression profiles from vastly fewer animals as previously reported (Aughey & Southall, 2016; Southall *et al.*, 2013). Other approaches like INTACT have focused on performing tissue-specific nuclei isolations prior to mRNA extraction that are substantially easier than tissue, particularly in *C. elegans,* but still require more material than reported for TaDa and are prone, like cell-sorting, to selection biases (Deal & Henikoff, 2011; Steiner *et al.*, 2012). In addition, loss of cytoplasmic mRNA could potentially capture an inaccurate transcriptome. Tissue-specific extraction of mRNA using transgenically expressed poly(A)-binding protein (PABPC) and sequencing (PAT-seq) is another approach that has been used in *C. elegans* but with potential drawbacks relating to poly(A)-tail length biases and toxicity due to high levels of PABPC (Blazie *et al.*, 2015, 2017; Yang, Edenberg & Davis, 2005). Lastly, cell type-specific single cell transcriptomes have been identified by single cell combinatorial indexing coupled with RNA sequencing (sci-RNA-seq), which is a powerful method that does not however possess the versatility to also provide simultaneously other information like chromatin accessibility or TF binding like TaDa (Cao *et al.*, 2017; Southall *et al.*, 2013; Aughey *et al.*, 2018).

Gene expression profiling by TaDa is not outright quantitative as the above are thought to be and its comparability with other methods remains to be seen. In this study, I perform the first TaDa gene expression profiling in the *C. elegans* epidermis in an attempt to capture differences in the transcriptional state of seam cells and hypodermis that determine their identity, while also identifying novel factors involved in seam cell patterning.

1.6.4 Probing genome-wide chromatin accessibility using TaDa

An evident advantage of TaDa as a methodology is its versatility as a tool to generate multiple types of genome-wide information. The most prominent such case is the identification of genome-wide chromatin accessibility which can be acquired from the Dam control fusions used in any TaDa experiment (Aughey *et al.*, 2018). As previously mentioned, untethered Dam or control fusions of Dam with proteins like GFP, are used as background models to remove non-specific methylation from the binding profiles of targeted Dam-fusions. That is because to a large extent the background has been found to reflect accessibility of chromatin (Wines *et al.*, 1996). Compact chromatin structure with DNA highly bound by nucleosomes is less accessible to freely diffusing Dam and more unlikely to be methylated by chance than open chromatin state regions. This effect is captured by the control Dam samples and the permutation of the method is called chromatin accessibility TaDa or CATaDa (Aughey *et al.*, 2018).

Chromatin accessibility is thought to reflect the epigenomic state of a cell that is consequential to its transcriptional state. Accessible chromatin regions harbour *cis*-regulatory elements that drive spatiotemporal patterns of gene expression (Tsompana & Buck, 2014; Klemm, Shipony & Greenleaf, 2019). A major level of control of cell fate and identity, as well as differentiation, relies on chromatin state and permissibility that dictates which parts of the genome can be deployed and TaDa can be used to resolve this in a tissue-specific manner (Klemm, Shipony & Greenleaf, 2019). Tissue-specific assaying of chromatin accessibility can provide the resolution required to understand how cell-types differ at that level and how these differences relate to gene expression profiles.

Other methods that can capture chromatin accessibility include: traditional DNA-nucleases approaches (usually DNase), that determine accessibility based on sensitivity to cleavage (Gross & Garrard, 1988), Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) which is based on cross-linking of chromatin and shearing by sonication to extract accessible regions (Giresi *et al.*, 2007) and the assay for transposase-accessible chromatin (ATAC) that utilises the hyperactive Tn5 transposase for *in vitro* tagmentation of accessible chromatin (Buenrostro *et al.*, 2013). Sequencing is mostly used for identification of the accessible regions and ATAC-seq is considered to be the most sensitive and powerful of the three, although their results have been found to be comparable both between them and with CATaDa (Buenrostro *et al.*, 2013; Aughey *et al.*, 2018).

Nonetheless, all these alternative methods by default lack the tissue-specificity of CATaDa and require laborious or challenging cell or nuclei isolation and sorting to achieve it (McClure & Southall, 2015). Therefore, so far in *C. elegans* only whole-animal ATAC-seq and DNase-seq experiments have been performed that reveal global open chromatin (Ho, Quintero-Cadena & Sternberg, 2017; Daugherty *et al.*, 2017; Jänes *et al.*, 2018). A previous attempt has also been made to acquire tissue-specific accessible chromatin profiles by expressing DAM:GFP using muscle, hypodermis and gut promoters but was not successful in generating informative accessibility maps with regions more accessible than the average (Sha *et al.*, 2010). This is most likely due to methylation saturation resulting from the very high levels of

expression driven by the unhindered tissue-specific promoters. In this study, I use CATaDa to attempt the first tissue-specific genome-wide chromatin accessibility mapping in *C. elegans*, focusing on the seam cells and the hypodermis.

1.7 Aims of this research

The fundamental pursuit of my doctoral research was to expand our understanding on how the genome is regulated and deployed in carrying out key developmental events in tissue morphogenesis. I have used the postembryonic development of the *C. elegans* epidermis and in particular the stem cell-like model system of the seam cells, to begin addressing questions on how gene regulatory networks of transcription factors act to determine seam cell and hypodermal fate, how batteries of expressed genes dictate cell-identity and how chromatin state participates in those decisions. To that end, I have adapted targeted DamID (TaDa) for use in *C. elegans,* as a single methodology to approach the questions above. This thesis presents the first application in this model organism, and in doing so presents findings that enrich our understanding and knowledge on seam cell patterning, while creating a framework for new experimentation and analysis in the future.

In the 3rd chapter of this dissertation I present the first application of tissue-specific transcription factor target identification in the *C. elegans* epidermis using TaDa. Specifically, I utilise the transcription factors LIN-22 and NHR-25 that possess seam cell developmental roles, to set up the method and validate its effectiveness and reproducibility. I study the resulting genome-wide binding profiles for their genomic localisation preferences and regulatory potential. Comparisons with available datasets are made to assess the comparability of TaDa with the ChIP-seq methodology and to test the biological meaningfulness of the discovered targets, which is aided by gene-set enrichment analysis. Utilising smFISH, novel TaDa-identified direct targets are validated for the two TFs, while new potential mechanisms of their functions in the epidermis are discussed. This chapter illustrates the feasibility of TF target identification by TaDa in the epidermis and provides findings that expand the known seam cell gene network.

In chapter 4, I present seam cell and hypodermis-specific gene expression profiling, performed by assaying RNApol occupancy by TaDa. As this was also a first, to ensure appropriate cell type-specificity, I perform *de novo* identification of suitable promoters with sufficient specificity to resolve gene expression profiles between related cell-types. I study methylation profiles and use gene-set enrichment analyses on TaDa-identified expressed genes to evaluate tissue-specificity and biological relevance to the corresponding cell-type. Comparisons of TaDa transcriptomes with available equivalent datasets are performed for method assessment and show extensive overlap. Lastly, small scale RNAi screens of TFs and chromatin factors from the identified seam cell-specific genes, as well as overexpression experiments of TaDa-identified seam cell-specific miRNAs, validate the method and reveal previously unknown seam cell development regulators. The results in this chapter demonstrate the potential of TaDa for gene

expression profiling in *C. elegans* and highlight how seam and hypodermal-specificity can be utilised to identify genes with developmental roles.

Chapter 5 of this thesis employs the CATaDa methodology to create the first tissue-specific chromatin accessibility maps of *C. elegans,* focusing on the seam cells and hypodermis. The genome-wide localisation of accessible chromatin is assessed in its positional preferences and co-occurrence with marks of active regulatory roles, chromatin states and TF binding sites. They are also compared with whole-animal ATAC-seq and DNase-seq datasets to assess the agreement across methodologies. By assigning open chromatin to genes, comparisons with gene expression profiles are made and tissue-relevance is evaluated. As validation, a selection of accessible chromatin regions are tested for their harbouring of enhancer or promoter *cis*-regulatory elements that drive expression in the epidermis. This chapter demonstrates how control samples from other TaDa experiments can be readily used to uncover genuine tissue-specific chromatin accessibility, pinpointing locations of *cis*-regulatory elements with similar tissue-specificity.

In the general discussion of chapter 6, I summarise and discuss the results from the scope of the advantages that have been brought forward from the application of TaDa and how they broadly compare to other methodologies. I then bring together findings from all three chapters and incorporate them along with literature information, to propose an expanded gene regulatory network that describes our understanding of how epidermal fates are being determined in *C. elegans*. Finally, I discuss how the experimental framework of this study can be further utilised for the elucidation of a more detailed quantitative regulatory mechanism underlying seam cell patterning.

Chapter 2

Materials and Methods

2.1 General *C. elegans* methods

2.1.1 Maintenance

The *C. elegans* strains used in this study were maintained on Nematode Growth Medium (NGM) (0.05 M NaCl, 0.25% w/v bacto-peptone, 1.7 % w/v Agar, 5 µg/ml cholesterol, 1 mM $MgSO_4$, 1 mM $CaCl_2$, 25 mM KPO_4 in de-ionised H_2O) in 55 or 90 mm polystyrene petri dishes/plates (Corning) grown monoxenically on a lawn of the uracil auxotroph *E. coli* strain OP50 as a food source (Brenner, 1974) unless otherwise stated. The N2 strain is used as the reference wild-type strain. Plates with animals used in experiments presented here were kept and grown in a free-standing cooled incubator (LMS™ series 4) at 20 °C unless otherwise stated. For all routine observation of animals a Nikon SMZ745 dissecting scope was used. For strain population maintenance two approaches have been used. Animals were either picked using a platinum wire affixed to the tip of a glass pipette (VWR), with the use of bacterial growth as adhesive to transfer individuals or groups of animals onto new plates, or a scalpel was used to cut and transfer a piece of NGM or "chunk" with animals on it onto a new plate (Corsi, 2006; Stiernagle, 2006). Both the scalpel and the platinum pick were sterilised before and after each transfer by flaming to prevent plate contamination and strain cross-contamination. For short-term storage plates of animals were kept at 15 °C (LMS™ series 2 incubator) to slow down population propagation. A complete list of all strains used in this study can be found in Appendix A.1. Some strains were provided by the *Caenorhabditis* Genetics Centre (CGC), a *C. elegans* strain repository at the University of Minnesota, USA, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440).

2.1.2 Strain decontamination and synchronisation

To decontaminate *C. elegans* strains and start new clean cultures, spot bleaching was performed (Stiernagle, 2006). A drop of ~50 µl of bleaching solution (28.5% v/v commercial bleach, 0.7 N NaOH in de-ionised H_2O) was placed on a new NGM plate outside of the lawn area. A number of gravid adult animals depending on the number of progeny required were transferred by picking from the contaminated plate and releasing into the drop. If large amounts of contamination were transferred another drop was added on top. To acquire large amounts of decontaminated animals or to achieve post-embryonic developmental synchronisation of strains large scale bleaching (egg-prep) was performed (Corsi, 2006; Stiernagle, 2006). To improve egg recovery the following steps were followed. Plates of animals grown preferably to contain multiple gravid adults were washed with 2 ml of the isotonic M9 buffer (42 mM Na_2HPO_4, 22 mM KH_2PO_4, 85 mM NaCl, 1 mM $MgSO_4$ in de-ionised H_2O) and were transferred in a 15 ml centrifuge caped tube (STARLAB). The tubes were centrifuged horizontally at 1200 g for 2 min in an Allegra® X-12 centrifuge (Beckman Coulter) using the SX4750 rotor. After supernatant removal the animals were treated with 2-3 ml of bleaching solution with frequent agitation until 70-80% of adult bodies

had cracked or had started dissociating and eggs were released. The tubes were centrifuged again for 3 min at 1200g, the bleaching solution was removed and the egg pellet was washed with 15 ml of M9. After a final centrifugation for 3 min at 1200 g the egg pellet was re-suspended in 200 µl of M9 and the eggs transferred around the bacterial lawn of an NGM plate. When large numbers of eggs were required, washed adult animals from multiple plates could be batched in a single tube, adjusting the volume of bleaching solution and M9 used for resuspension.

2.1.3 RNA interference (RNAi) by feeding

Gene expression knockdown through RNAi was performed by feeding *C. elegans E.coli* expressing double-stranded RNA (dsRNA) corresponding to a target gene of interest (Corsi, 2006; Timmons & Fire, 1998). All the RNAi bacterial clones used in this study are from the Chromatin and Transcription Factor sub-libraries of the commercially available Ahringer RNAi Library (Kamath & Ahringer, 2003) (Source Bioscience). A complete list of the RNAi clones used here is available in Appendix A.2.

RNAi bacteria were initially streaked from the deep-frozen library plates onto a Lysogeny broth (LB) agar (MILLER, Merck) plate with 50 µg/ml ampicillin (Sigma-Aldrich) and 12.5 µg/ml tetracyclin (Sigma-Aldrich) for selection and were incubated at 37 °C overnight for single colony isolation. Single colonies were seeded in 8 ml LB broth (Merck) containing 50 µg/ml ampicillin and 12.5 µg/ml tetracyclin and were grown at 37 °C overnight. 3 ml of the culture were used for small scale plasmid preparation (mini-prep) using the PureYield™ Plasmid Miniprep kit (Promega) and confirmation of the correct target gene sequence in the dsRNA expression unit was performed by Sanger sequencing (Eurofins Genomics) using the standard M13 uni (-21) primer, followed by BLASTn against the *C. elegans* PRJNA13758 genome assembly on www.wormbase.org/tools/blast_blat . The remaining 5 ml of RNAi culture were seeded as 300 µl lawns on NGM plates supplemented with 25 µg/ml ampicillin, 12.5 µg/ml tetracyclin and 1 mM Isopropyl β -D-1-thiogalactopyranoside (IPTG). RNAi plates were allowed to dry and lawns to form in the dark at room temperature for 48 hours, then stored at 4 °C and used within one month.

RNAi treatment of strains was performed by transferring 5 L4 animals on RNAi plates and were allowed to lay progeny that were observed for gene knockdown phenotypic effects at the stage of interest. For post-embryonic RNAi treatment, strains were transferred onto RNAi plates either by spot bleaching gravid adults or by seeding eggs from an egg-prep. Control treatments were performed for all RNAi experiments in parallel with treatments for targets genes, with precisely the same experimental conditions, by feeding animals on lawns of the same strain of HT115 bacteria that do not however express dsRNA targeting a gene. To test the effectiveness of the RNAi plates a *pop-1* RNAi clone culture was seeded in parallel with other target gene clones. 5 L4 N2 animals were transferred on the *pop-1* plates and were observed 2 days later for absence of first-generation (F1) progeny and high numbers of unhatched eggs due to embryonic lethality (Lin, Hill & Priess, 1998).

2.1.4 Genetics

Where combinations of genotypes were required, genetic crosses were performed by placing a ratio of 3:1 males to hermaphrodite animals (usually 9:3) at L4 on an NGM plate with a small (50 or 100 µl) OP50 lawn to increase frequency of interactions (Brenner, 1974). F1-Fn progeny were selected for the genotypes of interest based on either associated phenotypic traits or molecular genotyping (see section 2.3.2). In the absence of male producing genotypes (e.g. *him-5(-)*) or naturally occurring males, induction of males was performed by placing 30 L4 hermaphrodites at 33 °C for 3 hours moving them to 37 °C for 30 min and allowing them to recover, grow and lay progeny at 20 °C. Infrequent males in the brood were used to set up further mating crosses to increase the number of available males if F1 numbers were not sufficient.

2.1.5 Transient and stable transgenesis by microinjection

Transient transgenesis by formation of multi-copy extra-chromosomal arrays through microinjection was achieved following established protocols (Corsi, 2006; Mello *et al.*, 1992; Evans, 2006). In brief, an injection mix was prepared in a microcentrifuge tube (StarLab®) containing: 5-50 ng/µl of each plasmid carrying a genetic construct of interest, 5-20 ng/µl of a co-injection marker plasmid and a "carrier DNA" plasmid (pBJ36) up to a final concentration of at least 100 ng/µl of all DNA in a final volume of 10 µl in UltraPure™ Distilled Water (Life Technologies). The injection mix was centrifuged at full speed in an Eppendorf™ Benchtop 5424 or miniSpin 5804 microcentrifuge (Eppendorf) for 5 min. 2 µl of the mix were loaded in a 1 mm Kwik-Fil™ Borosilicate Glass Capillary (World Precision Instruments) needle pulled using a PC10 (NARISHIGE) needle puller and was fixed on the injection tube of a FemtoJet® 4x (Eppendorf) injector set at 1000 Pi. The injection tube was placed on an Eppendorf hydraulic controller system fitted to an inverted Ti-eclipse Microscope (Nikon) capable of DIC optics. Day-one adult animals to be injected were picked in a drop of halocarbon oil 700 (Sigma-Aldrich) and immobilised on a dried agarose pad prepared on 24 x 60 mm glass coverslip (VWR). After injections animals were recovered from the oil using M9 buffer and were singled-out on individual NGM plates and grown at 20 °C. The plates were screened 2-3 days post-injection for F1 progeny showing the phenotype driven by the co-injection marker using an AXIO Zoom V16 fluorescent dissecting scope (Zeiss). Independent transgenic lines were established using the transgenic F2 progeny for each injected parental (P0) animal.

Stable transgenic lines with single-copy locus-specific inserted transgenes were produced for the purposes of this study employing the *Mos1*-mediated single-copy insertion (MosSCI) method (Frokjaer-Jensen *et al.*, 2014). Standard protocols available at www.wormbuilder.org with adaptations were followed (Nance & Frøkjær-Jensen, 2019). In more detail, ~30 day-one adult animals of the EG6699 strain with a *Mos1* transposon insertion on chromosome II (*ttTi5605* locus) showing the uncoordinated (unc) phenotype were used for each transgene insertion by microinjection. All the MosSCI injection mixes used comprised of 50 ng/µl of a universal MosSCI vector carrying the transgene of interest flanked by the *ttTi5605* left and

right recombination arms along with plasmids harbouring the *Mos1* transposase (pCFJ601), a heat-shock inducible *peel-1* toxin (pMA122) and co-injection markers (pGH8, *myo-2::dsRed, myo-3::mCherry*) at the concentrations described in www.wormbuilder.org. Post-injection, animals were kept at 25 °C until plates were completely starved and adults had perished. The heat-shock treatment that follows was performed at 34 °C for 3.5 hours, after which the plates were allowed to recover for 3 hours at room temperature before "reverse chunking" was performed, where NGM chunks from the lawn of a new plate were placed on top of the starved, treated, plate with the OP50 lawn facing upwards (O'Connell, 2010). The next day the top of lawns were screened for normally roaming animals (non-unc) with absence of co-injection markers, which were transferred on a new NGM plate per injected P0. After homozygosity was achieved, putative lines were confirmed molecularly for single-copy insertion of the transgene of interest by Polymerase Chain Reaction (PCR) genotyping (see section 2.3.2) using oligos NM3880 and NM3884 (Appendix A.3). An independent transgenic line was established for each injected P0 that produced progeny with homozygous molecularly confirmed single copy insertions. A complete list of the transgenes produced for this study along with injection mix make-up information is available in Appendix A.4.

2.1.6 Cryopreservation

New strains or strains that required long-term storage were deep-frozen through the following process. A strain to be frozen was grown in a clean NGM 55 mm plate until the plate was fully populated. It was then separated in 6 equally sized chunks that were transferred onto 6 new 55 mm plates. Within the next 3-5 days the plates reached starvation and produced a large population of L1 larvae that were collected in a 15 ml centrifuge caped tube (STARLAB) by washing each plate with 2 ml of M9 buffer. The tube was centrifuged at 1200 g for 3 min and washed with 10 ml of M9. The washes were repeated until contaminants were removed. The clean animal pellet was resuspended in 3 ml of M9 to which 3 ml, of approximately body temperature, freezing solution (0.1 M NaCl, 50 mM KH_2PO_4, 5.6 mMNaOH, 0.6% w/v Agar, 24% v/v glycerol, 0.5 mM $MgSO_4$, 0.33 mM $CaCl_2$) was added and mixed thoroughly. The mixture was in turn aliquoted equally into 3 cryotubes (STARLAB) that were placed in a Styrofoam box and put at -80 °C for at least 3 days to freeze. Some amount from one of the aliquots was scraped and thawed onto a new NGM plate to test the quality of the freezing 72 h later. If some animals recovered within 24 hours two of the aliquots were transferred and were permanently kept in liquid nitrogen storage.

2.2 Microscopy and image analysis

2.2.1 Microscopic observation and image acquisition

To observe and image phenotypes visible on free roaming animals on a plate at a low magnification an AXIO Zoom V16 (Zeiss) dissecting scope with metal halide UV source was use, fitted with an Axiocam 305 mono camera (Zeiss) controlled via the ZEN imaging software (Zeiss). When immobilisation of animals

was required a drop of M9 containing 100 µM of the anaesthetic sodium azide (NaN$_3$) was placed on the plate and the animals were soaked in it before imaging.

For higher resolution and higher magnification observations and imaging, live animals were mounted on fresh 2% agarose pads, containing 100 µM NaN$_3$ for immobilisation, on a glass slide (VWR) covered by a glass 18 x 18 mm coverslip (VWR). The slides were then observed and imaged using either an AxioScope A1 (Zeiss) upright epifluorescence microscope with a Light Emitting Diode (LED) light source fitted with a RETIGA R6™ camera (Q IMAGING) controlled via the Ocular software (Q IMAGING) or on an inverted Ti-eclipse fully motorised epifluorescence microscope (Nikon) with a metal halide light source fitted with an iKon M DU-934, 1024 x 1024 CCD-17291 camera (Andor) controlled via the NIS-Elements software (Nikon). Where comparisons of fluorescence intensity between strains or treatments are made in this study, image acquisition had always been performed during the same session for all compared samples using the same microscopy set-up, magnification and exposure time.

2.2.2 Counting seam cell and postdeirid neuron numbers

Scoring of the terminal seam cell number phenotype as a proxy of seam cell developmental differences between strains or treatments was performed by mounting a sufficient sample size per condition, of late-L4 up to early adult animals, on glass slides as described above. The seam cell number (scn) of the lateral side most proximal to the objective was counted for every animal using the microscopy set-ups described above. For each particular experiment where comparisons need to be made all samples were grown in parallel and scored on the same day. When scoring had to be performed over multiple days, the equivalent control sample was scored in parallel. For postdeirid neuron counting, late-L4 up to early adult animals carrying the *dat-1p::GFP* marker were mounted on slides and the dopaminergic neuron bodies visible on the lateral side proximal to the objective lens and posterior to the pharynx were counted and compared across samples.

2.2.3 Single molecule fluorescent *in situ* hybridisation (smFISH)

In this study, smFISH results are presented for animals at the late-L1 and during the asymmetric seam cell division of L2 and L3 stages. To achieve this, synchronisation of large populations of animals was performed by egg-prep (as in section 2.1.2) and subsequent growth of animals at 20 °C for 18 hours for late L1, 25 hours for the L2 and 35 hours for the L3 animals was required. Animals were collected off plates in 1.5 ml microcentrifuge tubes (STARLAB) with 2ml M9 and washed to remove bacteria by consecutive rounds of centrifugation at 1200 g for 3 min and addition of 1.5 ml clean M9. They were subsequently fixed with 1.5 ml 4% formaldehyde (Sigma-Aldrich) in 1x PBS (Ambion) for 45 min on a vertical Stuart™ Rotating disk (Cole-Palmer), rotating at 15 revolutions per minute (RPM). They were washed with 1.5 ml 1x PBS twice before being stored at 4 °C in 1.5 ml 70% ethanol for at least 24 hours.

For hybridization, the ethanol was removed, the animals were washed and incubated for 5 min at room temperature in 1.5 ml Wash buffer (2 x saline-sodium citrate (SSC) (Ambion), 10% formamide (Ambion)) and were then resuspended in 100 µl of Hybridisation buffer (100 mg/ml dextran sulphate (Sigma-Aldrich), 10% formamide in 2x SSC). 1 µl of a diluted in water between 1:5 and 1:50 custom-made mixture of 21-48 Quasar 670 labelled oligonucleotide probes (Biosearch Technologies) targeting the gene of interest was added to the suspension and the samples were incubated at 30 °C for 16 hours. A complete list of probes and their dilution used in this study along with their sequences is available in Appendix A.5.

Post-hybridisation the samples were washed with 1.5 ml Wash buffer once before being incubated in 1.5 ml Wash Buffer at 30 °C for 30 min, followed by another incubation in 1 ml Wash buffer containing 5 ng/ml of 4′,6-diamidino-2-phenylindole (DAPI) at 30 °C for 30 min. The Wash buffer/DAPI solution was removed and the animals were kept in 1.5 ml 2x SCC at 4 °C for at least 2 hours and were imaged within 2 days.

For imaging the animals were resuspended in 100 µl GLOX buffer (0.4% glucose, 10 mMTris-HCl in 2x SSC) supplemented with 1 µl of 3.7 mg/ml glucose oxidase (Sigma-Aldrich) and 1 µl of 5 mg/ml Catalase (Sigma-Aldrich). A volume of 4 µl of animal suspension was transferred to a round 10 mm glass coverslip (VWR) using a pipette tip that had been treated with bacterial culture to coat and prevent animal adhesion to plastic. A 24 x 24 mm coverslip was placed on top and once excess buffer was removed and the two coverslips were tightly pressing on the animals they were placed with the round coverslip facing down onto a silicone seal siting on a glass microscopy slide (VWR) and were carefully sealed. Imaging was performed using the Nikon set-up described in section 2.2.1 using the seam cells closest to the objective lens as homing coordinates to acquire 17 Z-stack slices with a step of 0.8 µm for each of the DAPI, Cy5 and GFP channels (Semrock). Acquisition was performed using a 100x oil immersion objective with exposure set at 100 ms for DAPI with intensity of excitation light reduce to 1/32, 3 s for Cy5 at full intensity and 300 ms for GFP at full intensity. Data were exported as .TIFF files and were analysed, as described in Katsanos *et al*., 2017, using a custom MATLAB® (MathWorks) pipeline (Barkoulas *et al.*, 2013). In brief, selected animal DAPI and GFP images were used to annotate seam cells and draw regions of interest (ROIs) around the nuclei for at least 5 slices within which smFISH spots would be counted. An animal specific threshold for spot detection was set by manually sampling spots based on which automated counting was carried out.

2.2.4 Microscopy image-processing for presentation

Representative microscopy images that are shown in the results sections of this study, after relevant image analysis had been performed, were processed using the Fiji software (NIH) for presentation purposes. Processing included cropping, file conversion, pseudo-colouring, adjustment of brightness and contrast in those cases were intensity of signal is not the informative variable and straightening. Straightening was performed when required using a Macro developed by an undergraduate student Fu

Xiang Quah. Specifically for smFISH photos the probe channel (Cy5) was inverted and sharpened once for clarity and to improve the resolution of the spots.

2.3 Molecular Methods

2.3.1 Small-scale genomic DNA (gDNA) extraction

To perform molecular genotyping of strains or to clone sequences from the *C. elegans* genome, small scale genomic DNA extraction had to be performed to be used as template for amplification. The method followed allowed for sufficient amounts of gDNA to be extracted even from single animals which could allow genotyping of single mothers of clonal populations. Between a single and up to 20 animals, depending on the application, were transferred in a 0.2 ml capped PCR tube (STARLAB) containing 7-10 µl of Worm Lysis Buffer (WLB) (1 mM $MgCl_2$, 0.45% Tween in 1x Colorless GoTaq® Buffer (Promega)) supplemented with 0.2 µg/ml Proteinase K (QIAGEN). The tube was incubated at 65 °C for 1 hour for lysis of animals to occur, followed by 15 min at 95 °C to inactivate the Proteinase K in an Eppendorf Mastercycler® nexus X2 thermocycler. The lysate could be stored at -20 °C and 0.5 to 2 µl was used for downstream application.

2.3.2 Sequence amplification by Polymerase Chain Reaction (PCR) and molecular genotyping

Amplification of DNA sequences required for genotyping of strains, screening of bacterial colonies or cloning, was performed by PCR. The majority of strain genotyping and colony screening was performed by amplification of sequences smaller than 3 kilobase-pairs (kb) and was done using the GoTaq® G2 DNA polymerase system. For *C. elegans* strains genotyping, 0.5 to 2 µl of lysate produced by the above process was added to a 0.2 ml PCR tube (STARLAB) along with 5 µl 5x Colorless GoTaq® Buffer, 0.5 µl of 10 mM deoxynucleotide triphosphates (dNTPs) (Promega), 0.025 units (or 0.125 µl) of GoTag® G2 DNA polymerase (Promega), 0.5 µl of a 10 µM dilution of the forward primer DNA oligonucleotide (oligo) and 0.5 µl of a 10 µM dilution of the reverse primer DNA oligo, targeting the sequence of interest, toped up to 25 µl with UltraPure™ Distilled Water (Life Technologies). For bacterial single colony screens, the mix was prepared similarly but instead of lysate as template a pipette tip was used to transfer some of the colony into the mix. The reactions were performed either in an Eppendorf Mastercycler® nexus X2 thermocycler or a PCRmax™ Alpha Cycler 1 with the following basic PCR program: 1) 2 min at 95 °C, 2) 30 sec at 95 °C for denaturation, 3) 30 sec at 50-65 °C for primer annealing depending on the sequence, 4) 1 min/kb at 72 °C depending on the length of the amplification, 5) repeat steps 2-4 35 times, 6) 10 min at 72 °C for final extension.

For the genotyping of MosSCI single-copy insertions, where amplification of sequences larger than 10 kb was required, the Expand™ Long Template PCR system (Roche) was used. Buffer 1 was used for

sequences up to 9 kb, buffer 2 for 9 to 12 kb and buffer 3 for larger than 12 kb. The reactions were performed in a volume of 25 µl of 1x of the appropriate buffer in UltraPure™ Distilled Water with 1.25 µl of 10 mM dNTPs, 0.75 µl of each 10 µM primers dilution for amplification and 0.375 µl (or 1.875 units) of Expand Long Template Enzyme mix. For the cycling program the denaturation temperature was set at 94 °C and the elongation at 68 °C and cycling was split in two parts: the first 10 cycles and the next 25. The length of the elongation step was set according to the manufacturer's guidelines, increasing in increments of 20 sec after every cycle in the second part of the program.

Where amplification fidelity was essential and sequence accuracy was required, either for cloning or transgenesis, the Phusion® High-Fidelity DNA polymerase system (New England Biolabs) was used. The reaction mix was in a total volume of 50 µl of 1x Phusion® HiFi Buffer in UltraPure™ Distilled Water containing, 1 µl of 10 mM dNTPs, 2.5 µl of each primer at 10 µM and 0.5 µl (or 1 unit) of Phusion® Polymerase. The amount of template varied from 0.5 ng of DNA in the case of plasmid and fosmid template, to 2 µl when lysate was used. The cycling program used was similar to the basic program described for GoTaq® with initial hot-start and denaturation temperature increased to 98 °C.

To assess the results of PCR, a volume between 2 µl and the total reaction volume, was mixed with 6x Loading Dye (Thermo Scientific) and water to 1x and was used to perform DNA electrophoresis using a 1% agarose (Sigma-Aldrich) in 1x Tris-Borate Ethylene-diamine-tetraacetic Acid (EDTA) (TBE buffer) (Sigma-Aldrich) gel, stained with 0.1x SYBR® Safe DNA gel stain (Invitrogen) and using 5 µl of MassRuler Express Forward DNA Ladder Mix (Thermo Scientific) as a size marker. A PowerPac™ power supply (BioRad) was used to perform electrophoreses at a voltage of 85-150 V for 40 min to 1.5 hours. Electrophoresis results were viewed using a Safe Imager™ (Invitrogen) transilluminator and captured using the InGenius™ gel documentation system (Syngene). All the oligos used in this study are listed in Appendix A.3.

2.3.3 General cloning practices

The design of all cloning pipelines, reactions and oligos was performed using the Benchling Software (www.benchling.com). All DNA restriction digestions in this study were performed using the FastDigest™ system (Thermo Scientific) unless otherwise stated. Reactions were performed according to the manufacturer's guidelines, with the digestion time extended to 1.5 -2 hours when plasmid backbones were prepared for cloning. Prior to downstream cloning reactions, digestion fragments and PCR products were run on 1% agarose gels as described above to resolve DNA bands needed and were excised using a scalpel and gel extracted using the Wizard® SV Gel and PCR Clean-Up System (Promega). Insertion of sequences in plasmid vectors by sticky-end cloning was performed using the T4 DNA ligase (Promega) with a molar ratio of insert to vector of 3:1 in a total volume of 10 µl of 1x T4 DNA Ligase buffer or 1x Rapid ligation buffer (Promega) in UltraPure™ Distilled Water. The amount of vector DNA used was 30 to 100 ng per reaction with 1-3 units of T4 DNA ligase and the reactions were incubated at room temperature for 2 to 16 hours. Cloning by isothermic Gibson assembly (Gibson *et al.*, 2009) was performed in a 0.2 ml

PCR tube using a 5 µl "homemade" Gibson mix containing 0.2 M Tris-HCl (Merck), 20 mM MgCl$_2$, 1.6 mMdNTPs (Promega), 20 mMDithiothreitol (DTT) (Sigma-Aldrich), 1.04% w/v Polyethylene glycol-800 (PEG-800) (Merck), 2 mM β-Nicotinamide adenine dinucleotide hydrate (NAD) (Merck), 0.008 U/µl T5 exonuclease (Epicentre), 0.05 U/µl Phusion™ polymerase (New England Biolabs), 4.5 U/µl Taq Ligase (New England Biolabs). To the mix, 50 ng of the digested vector DNA and an equimolar amount of each of the fragments to be inserted were added and toped-up to 10 µl with UltraPure™ Distilled Water. The reaction was incubated in a thermocycler for 1 hour at 50 °C. Gateway cloning reactions were performed following the MultiSite Gateway® Technology protocols from Invitrogen using the Gateway® BP Clonase™ II Enzyme mix (Invitrogen) and the Gateway® LR Clonase™ II Plus Enzyme Mix (Invitrogen). Where the adapted version of Golden gate assembly (Engler, Kandzia & Marillonnet, 2008) is used in this study, a custom mix was prepared in a 0.2 ml PCR tube containing 0.5 µl (or 2.5 units) of T4 DNA Ligase (Thermo Scientific), 0.5 µl of the BpiI and 0.5 µl of the Esp3I FastDigest™ enzymes (Thermo Scientific), 50 ng of the vector plasmid and an amount of insert for a 2:1 insert:vector molar ratio in 1x T4 Ligase buffer (Thermo Scientific) in a total volume of 10 µl. Reactions were incubated at 37 °C for 30 min, 5 min at 50 °C and 5 min at 80 °C in a thermocycler.

For all the above cloning reactions a volume of 5 µl was transformed in Lab prepared Dh5α competent bacteria by mixing 100 µl of the bacteria with an 100 µl mixture containing 0.1 M KCl, 0.03 M CaCl$_2$, 0.05 MMgCl$_2$ and the 5 µl of the reaction in UltraPure™ Distilled Water. The resulting bacterial suspension was incubated on ice for 30 min followed by a heat-shock at 42 °C for 45 sec and a 1 hour incubation at 37 °C shaking at 400 RPM on a Labnet (AccuTherm) shaking heat-block after the addition of 800 µl LB broth. Transformed bacteria, concentrated in 200 µl of LB by centrifugation at 2500 g for 3 min and supernatant removal, were resuspended and spread on LB agar plates with the correct antibiotic selection (one or a combination of: 100 µg/ml ampicillin, 12.5 µg/ml tetracyclin, 30 µg/ml Chloramphenicol, 50 µg/ml Kanamycin) and were incubated at 37 °C overnight.

Single colonies for plasmid preparation were grown overnight in 4 ml of LB with the correct antibiotic selection (see above) and small scale plasmid extraction (mini-prep) was performed using the PureYield™ Plasmid Miniprep System (Promega) and plasmids were eluted in UltraPure™ Distilled Water. All concentrations of DNA solutions in this study were measured using either a NanoDrop Lite (Thermo Scientific) or a DS-11 Spectrophotometer (DeNovix). Confirmation of the correctness of cloning products was achieved by Sanger sequencing (Eurofins Genomics) using the appropriate oligos to acquire informative reads that were aligned to the expected plasmid assemblies using the MAFFT algorithm on www.benchling.com.

2.3.4 Cloning of seam cell and hypodermis specific promoters

For the assembly of the hypodermis reporters to test the specificity of promoters the following cloning was performed. The pre-existing pIR6*(pdpy-7::unc-54 3'UTR)* plasmid was digested with EcoRI and SmiI to remove the *pdpy*-7 promoter and linearise the plasmid. Using oligos MBA270 and MBA271 the promoter

of *dpy-7* was amplified while at the same time altering 2 GATA sites on the 5' and the 3' of the sequence to form the *dpy-7syn1* promoter, which was inserted by Gibson assembly in the digested pIR6 to form the pIR16(*dpy-7syn1::unc-54 3'UTR)* plasmid (pIR16 cloning was performed by a lab technician Iqrah Razzaq). pIR16 was linearised with SmiI digestion and the sequence of *mCherry-H2B* was amplified using oligos DK46 and DK47 from a pre-existing pENTR *mCherry-H2B* plasmid and inserted by Gibson in pIR16 to form pDK18(*dpy-7syn1::mCherry-H2B::unc-54 3'UTR*).

To study the expression pattern using reporters of the putative seam cell specific promoter of the *srf-3* gene, 3 versions of the promoter were amplified from N2 lysate. The isoform a promoter *srf-3ap* was amplified using DK33 and DK34 oligos and was inserted in a Gibson assembly along with pCFJ151 backbone, *C. elegans* optimized GFP (GFPo) amplified from JH01 (Heppert *et al.*, 2016) with DK35 and DK36 oligos, H2B amplified from the pENTR *mCherry-H2B* plasmid mentioned above using oligos DK37 and DK38 and *unc-54* 3'UTR amplified with DK39 and DK40 oligos from pIR6. The resulting construct was pDK16(*srf-3ap::GFPo-H2B::unc-54 3'UTR + cb-unc-119*) which was digested with NheI/XmaJI to remove the promoter and to be used for Gibson assembly of the other versions of the *srf-3* promoter. The isoform b promoter *srf-3bp* was amplified using oligos DK33 and DK59, the *srf-3* intron 1 was amplified with oligos DK64 and DK65 and was fused by fusion PCR to the *pes-10* minimal promoter amplified with DK66 and DK67 from L3135 (Fire Lab vector Kit). Both were inserted in NheI/XmaJI digested pDK16 to create pDK26(*srf-3bp::GFPo-H2B::unc-54 3'UTR*) and pDK32(*srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR*).

2.3.5 Cloning of transcription factor and *mCherry*-lacking TaDa constructs

To construct a TaDa ready backbone plasmid for epidermis specific expression of TFs fused upstream of Dam the pCFJ151 universal MosSCI vector (Frokjaer-Jensen *et al.*, 2014) was digested with BcuI/BspTI enzymes, the promoter of *wrt-2* was amplified from N2 lysate with oligos PB16 and PB7, the *C. elegans* optimised *wormCherry* was amplified from the pAA64 (Barkoulas *et al.*, 2016) plasmid using oligos PB8 and PB17, the *dam* sequence was amplified from the pUAST attB LT3 Dam plasmid (kindly donated by Tony Southall) using oligos PB18 and PB13 and the *unc-54 3' UTR* was amplified from N2 lysate using oligos PB14 and PB15. All 4 fragments and the digested backbone were inserted in a multi-fragment Gibson assembly reaction to produce the pPB7(*wrt-2p::wormCherry::dam::unc-54 3'UTR + cb-unc-119*) plasmid. The XmaJI site between *wormCherry* and *Dam* was digested to linearise the vector and allow the in-frame to *dam* insertion of the *nhr-25* coding sequence amplified using oligos PB19 and PB20 from N2 cDNA to produce the pPB10(*wrt-2p::wormCherry::nhr-25:dam::unc-54 3'UTR + cb-unc-119*) plasmid. The cloning for pPB7 and pPB10 was performed by a Master's student in the lab Patrick Brehm.

To construct the seam cell driven *lin-22:dam* fusion and the *NLS-GFP:dam* control for the TF TaDa experiments, the *lin-22* gene was amplified with oligos DK11 and DK12 from fosmid WRM0627dG07 while *NLS-GFP* was amplified from plasmid pPD93_65 (Fire Lab vector Kit) using oligos DK15 and DK16. Both amplicons were inserted upstream and in-frame with *dam* by Gibson assembly in an XmaJI linearised pPB7 vector like above. The resulting plasmids produced were pDK4(*wrt-2p::wormCherry::lin-*

22:dam::unc-54 3'UTR + cb-unc-119) and pDK8(wrt-2p::wormCherry::NLS-GFP:dam::unc-54 3'UTR + cb-unc-119).

To test the importance of the *mCherry* primary ORF to the viability of animals and methylation levels, versions of the *lin-22:dam* and *NLS-GFP:dam* TaDa constructs without *wormCherry* were produced. In more detail, the pPB7 plasmid was digested with BcuI/MunI, the 4085 bp and 6141 bp fragments were excised, extracted and kept. *lin-22* was amplified from pDK4 using DK102 and DK11. The 2 digestion fragments, *the lin-22* amplicon and the repair oligo DK103 were all inserted into a Gibson reaction to produce pDK49(*wrt-2p::lin-22:Dam::unc-54 3'UTR + cb-unc-119*) which was then digested with BcuI/XmaJI to remove *lin-22* and insert via Gibson assembly a DK108 and DK15 amplified fragment of *NLS-GFP* from pDK8 to generate pDK50(*wrt-2p::NLS-GFP:dam::unc-54 3'UTR + cb-unc-119*).

2.3.6 Cloning of the RNA polymerase TaDa constructs

For ease of future applications a versatile TaDa vector called pDK7 was constructed and used for the RNApol TaDa plasmids presented here. The *att* recombination cassette of pDest R4-R3 (Invitrogen) was amplified including the *attR4* site, *ccdb* and *CamR* genes but excluding the *attR3* site using the oligos DK17 and DK18 including half of the *attL1* site sequence on the 3' DK18 primer. *wormCherry* was amplified from pPB7 using oligos DK19 and DK20 carrying the other half of the *attL1* site on the 5' of DK19. *dam* was amplified from pPB7 with oligos DK21 and DK22. The *unc-54 3'UTR* was amplified from pPB7 with oligos DK23 and DK24. All 4 fragments were inserted in a Gibson assembly reaction along with BcuI/BspTI doubly digested pCFJ151 vector to generate pDK7(*attR4-L1::wormCherry::dam-myc::unc-54 3'UTR + cb-unc-119*).

The *srf-3i1::pes-10* promoter was amplified with DK89 and DK90 from pDK32 and the *dpy-7syn1* with DK113 and DK114 from pDK18 and donor vectors were produced via a BP reaction (pDK44 and pDK61 respectively). pDK7 digested with PaeI was used to insert *rpb-6* amplified with DK27 and DK28 from N2 lysate and *NLS-GFP* amplified form pPD93_65 with DK43 and DK44, downstream and in-frame with *dam*. The two intermediate plasmids were inserted in parallel LR reactions with pDK44 and pDK61 to finally produce 4 different plasmids, pDK54(*srf-3i1::pes-10::wormCherry::dam-myc:NLS-GFP::unc-54 3'UTR + cb-unc-119*), pDK55(*srf-3i1::pes-10::wormCherry::dam-myc:rpb-6::unc-54 3'UTR + cb-unc-119*), pDK64(*dpy-7syn1::wormCherry::dam-myc:NLS-GFP::unc-54 3'UTR + cb-unc-119*) and pDK65(*dpy-7syn1::wormCherry::dam-myc:rpb-6::unc-54 3'UTR + cb-unc-119*).

To test *ama-1*, the major subunit of RNA polymerase, for TaDa the p304(*cb-unc-1119 + phsp-16::ama-1:dam::unc-54 3'UTR*) plasmid (kindly donated by Peter Meister) was converted into a TaDa versatile vector with a gateway docking site for easy promoter insertion by digesting with XmaJI/PteI to remove the existing promoter along with a part of *dam*. From pDK7 using the primers DK41 and DK42 a compatible Gibson amplicon containing *attR4-L1::mcherry::dam(part)* was amplified and inserted in a Gibson assembly with the above vector to produce pDK20(*cb-119 + attR4-L1::wcherry::dam-myc::ama-*

1::unc-54 3'UTR). pDK20 was inserted in LR reactions with pDK44 and pDK61, as above, to produce pDK46(*cb-unc-119 + srf-3i1::pes-10::wormCherry::dam-myc:ama-1::unc-54 3'UTR*) and pDK62(*cb-unc-119 + dpy-7syn1::wormCherry::dam-myc:ama-1::unc-54 3'UTR*).

2.3.7 Cloning the *lin-17* conserved promoter regions reporters

For the *lin-17 CRE1* and *CRE2* transcriptional reporters the oligos DK115 and DK116 along with DK118 and DK119 were used to amplify each of the respective regions from N2 lysate. The *Δpes-10* core promoter was amplified from L3135 using either the *CRE1* or the *CRE2* compatible forward primers DK117 and DK120 along with the DK107 reverse and was cloned along with the respective CRE amplicon in a NheI/XmaJI digested pDK16 to create pDK59(*lin-17CRE1:: Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119*) and pDK60(*lin-17CRE2:: Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119*).

2.3.8 Cloning the hairpin RNAi constructs

To achieve stable and heritable epidermis specific RNAi gene knockdown the following constructs were assembled, to allow expression of hairpin RNA gene fragments that were easily cloned by golden gate assembly. The plasmid pDK102 carrying a *dpy-7syn1* promoter and the *p10 3'UTR* (Pfeiffer, Truman & Rubin, 2012) in a pCFJ151 backbone was digested with XmaJI/PacI to linearise and allow for cloning between the promoter and the 3'UTR. A gene fragment called GoldenGateHairpin carrying a compatibility arm to the *dpy-7syn1* promoter, an outron, 2 inverted repeats of the BpiI enzyme the 5th intron from the *srf-3* gene, two inverted repeats of the Esp3I enzyme and compatibility arm to the *p10 3'UTR* was synthesised (GENEWIZ) and was inserted in the digested pDK102 to form the intermediate plasmid pDK109(*dpy-7syn1::GoldenGateHairpin::p10 3'UTR + cb-unc-119*). Using oligos DK186 and DK179 the sequence from the promoter to the 3'UTR was amplified and inserted by Gibson assembly in a KpnI/NotI digested pBluescript vector to form pDK110(*dpy-7syn1::outron::GGBpiI::srf-3a intron5::GGEsp3I::p10 3UTR*). To modify the golden gate (GG) enzyme sites such that they leave non-palindromic scars to allow for specific directional cloning the *srf-3* intron 5 was amplified from pDK110 using oligos DK212 and DK214. The resulting amplicon was amplified again and extended with oligos DK213 and DK215 and was inserted in a BpiI/Esp3I digested pDK110 backbone by Gibson assembly to produce pDK127(*dpy-7syn1::outron::non-palGGBpiI::srf-3a intron5::non-palGGEsp3I::p10 3'UTR*).

To initially test the effectiveness and specificity of the system a fragment from GFP not containing sites for BpiI and Esp3I was amplified from L3135 using oligos DK203 and DK204 and was inserted by Golden gate assembly in pDK127 in two inverted repeats upstream and downstream of the *srf-3* intron 5 to create pDK130(*dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR*). To create a seam cell expressing version the *srf-3i1::Δpes-10* promoter was amplified from pDK126 using oligos DK234 and DK244 and was inserted in a Gibson assembly reaction with SalI digested pDK130 to remove

the *dpy-7syn1* promoter and create pDK134(*srf-3i1::Δpes-10::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR*).

To produce a seam cell expressing construct for an *hda-1* hairpin the empty Golden gate cassette was digested out of pDK127 using XmaJI and PacI and was inserted by sticky end cloning in XmaJI/PacI digested pDK134 to create pDK144(*srf-3i1::Δpes-10::outron::non-palGGBpiI::srf-3a intron5::non-palGGEsp3I::p10 3'UTR*). Because the *srf-3i1* sequence contains a BpiI recognition sequence site directed mutagenesis was performed on pDK144 to add an adenine nucleotide (GAAGAC to GAAAGAC) and disturb the sequence using oligos DK249 and DK250 and produce pDK157(*srf-3i1-mut::Δpes-10::outron::non-palGGBpiI::srf-3a intron5::non-palGGEsp3I::p10 3'UTR*). A fragment from the *hda-1* gene overlapping the 2nd and 3rd exons that doesn't contain a BpiI or Esp3I was amplified using oligos DK247 and DK248 to produce pDK158(*srf-3i1-mut::Δpes-10::outron::>hda-1 fragment>::srf-3a intron::<hda-1 fragment<::p10 3'UTR*).

2.3.9 Cloning of miRNA overexpression constructs

The vectors pDK127 carrying a *dpy-7syn1* promoter and *p10 3'UTR* and pDK82 carrying a *srf-3i1::Δpes-10* promoter and *p10 3'UTR* were both digested with XmaJI and PacI to remove sequences between the promoter and 3'UTR and prepare them to be used as backbones in Gibson assemblies. The miRNAs *mir-42, mir-43* and *mir-44* that form a compact complex on chromosome II were amplified on the same fragment to ensure proper post-transcriptional processing, using oligos DK217 and DK218 as well as DK219 and DK218. The two amplicons were inserted in the above digested pDK82 and pDK127 respectively to form pDK133(*srf-3i1::Δpes-10::mir-42-44::p10 3'UTR*) and pDK147(*dpy-7syn1::mir-42-44::p10 3'UTR*). Similarly *mir-47* was amplified using the pairs of oligos DK220 and DK221 as well as DK222 and DK221 to produce a pDK127 and a pDK82 compatible product that were inserted by Gibson assembly to create pDK139(*srf-3i1::Δpes-10::mir-47::p10 3'UTR*) and pDK148(*dpy-7syn1::mir-47::p10 3'UTR*). For all miRNAs the oligos incuded part of the endogenous 3'UTR to ensure sequences required for post-transcriptional processing are present. The cloning of these constructs was performed by a Master's student in the Lab Mar Ferrando-Marco.

2.3.10 Cloning of reporters from CATaDa-identified regulatory sequences

To assess the spatial expression regulation capacity of CATaDa-identified epidermal open chromatin sequences, oligos were designed within 100 bp of the start and end coordinates of each region to be tested. For ease regions were named according to the closest gene they could be regulating. Thus, N2 lysate was used to amplify the sequences outlined below with the indicated pair of oligos:

1. *F22B7.3* upstream element with DK237 and DK238
2. *rps-25* upstream element with DK239 and DK240
3. *K02A2.5* upstream element with DK223 and DK224

4. *nhr-4* distal upstream element with DK225 and DK226
5. *nhr-4* proximal upstream element with DK227 and DK228
6. *YF38F1A.8* upstream element with DK231 and DK232
7. *nhr-25* upstream element with DK233 and DK234
8. *nhr-25* downstream element with DK235 and DK236

All the oligos used were compatible with an XbaI/NheI digested L3135 vector to which they were inserted via a Gibson assembly upstream of a core *pes-10* promoter to create pDK146(*CATaDa F22B7.3 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*), pDK145(*CATaDa rps-25 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*), pDK140(*CATaDa K02A2.5 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*), pDK154(*CATaDa nhr-4 distal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*), pDK149(*CATaDa nhr-4 proximal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*), pDK152(*CATaDa Y38F1A.8 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*), pDK141(*CATaDa nhr-25 -upstream element::pes-10::GFP:LacZ::unc-54 3'UTR*) and pDK150(*CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR*). The cloning for the above plasmids was performed by a Master's student in the Lab Mar Ferrando-Marco.

2.4 Targeted DamID (TaDa) lab protocol

2.4.1 Strain cultivation, population expansion and collection

Plates to culture *C. elegans* for TaDa experiments where prepared by growing a single colony of dam^-/dcm^- *E. coli* (New England Biolabs) in LB containing 30 µg/ml Chloramphenicol and seeded as 300 µl lawns on empty NGM plates. Strains for which TaDa was performed were transferred onto TaDa plates by spot bleaching 10 adults on two separate plates per strain that were grown independently throughout but processed simultaneously and constituted the biological replicates of the experiment for each strain. Great care was taken throughout to prevent contamination of TaDa plates that could be detrimental to the experiment due to growth of dam^+ bacteria. By spot bleaching the single plate per replicate, per strain, was expanded to two plates. Once these plates became fully populated with large numbers of gravid adults, animals were collected and were large-scale bleached and eggs split in 5 new TaDa plates. The washed plates that were covered with laid eggs were chunked in a total of 4 TaDa plates. Taking advantage of the already laid eggs allowed for quicker expansion of the population and more animal material in a shorter amount of time reducing the probability of persistent contamination to arise by reducing the number of generations on TaDa plates. The total of 9 plates per repeat per strain were grown until animals reached gravid adulthood and were large-scale bleached and eggs split in 5 TaDa plates.

In this study TaDa was performed in L2 and L4 synchronised populations. Therefore, 24 hours after bleaching the L2 staged animals (also assessed by microscopy) from the 3 most populated plates per repeat per strain were washed thoroughly using 2 ml M9 buffer per plate (serially washed twice with 1 ml)

and collected in a 15 ml centrifuge tube, such that almost all individuals were transferred. Similarly, 48 hours after bleaching the L4 animals from the remaining 2 plates were collected. For both stages and for all samples, as soon as animals were collected they were placed on ice to halt further development and reduce movement. Extensive washing followed (as in section 2.1.2) by centrifuging animals at 1200 g for 3 min, removing the supernatant and washing the pellet (~100 µl) with 10 ml M9. After each addition of M9 the tube was briefly vortexed or shaken to ensure complete resuspension of any bacterial pellet. In total 3 washes with 10 ml of M9 and 2 washes with 5 ml of M9 were performed per sample, keeping the tubes on ice in between. The final animal pellet (~100 µl) was clear from bacteria to the eye. This served to reduce the amount of bacterial DNA in the samples after the extraction that could interfere with downstream processing. Pellets were frozen at -20 °C before gDNA extraction.

2.4.2 Large-scale gDNA extraction for TaDa

Tubes with animal pellets were thawed and 750 µl of Cell Lysis Solution (QIAGEN) was added to each tube and animals were resuspended and transferred to a clean 1.5 ml microcentrifuge tube (STARLAB). To each tube 4 µl of 20 mg/ml Proteinase K (QIAGEN) was added and mixed. The samples were then placed on a heat-block at 55 °C shaking at 500 RPM for 16 hours overnight. Lysates were treated with 4 µl of 20 mg/ml RNase A (QIAGEN) at 37 °C shaking at 500 RPM for 3 hours. In turn, 250 µl of Protein Precipitation Solution (QIAGEN) was added to each sample and were incubated on ice for 5 min followed by vigorous vortexing of each sample for at least 30 sec and another incubation on ice for 5 min. The samples were then centrifuged at 6000 g for 10 min at 4 °C in an Eppendorf 5804R refrigerated centrifuge (Eppendorf). The supernatant (~900 µl) was transferred into a new 1.5 ml microcentrifuge tube for each sample already containing 750 µl of Isopropanol (the real capacity of the 1.5 ml tubes is closer to ~1.7 ml), they were caped and mixed very well by inverting at least 200 times. The samples were centrifuged again as before and a very small white DNA pellet was formed in most. After careful removal of the isopropanol the DNA pellets were washed with 750 µl of 70% ethanol by inverting the tubes again at least 200 times, centrifuging as before and removing most of the ethanol carefully to leave the pellet in the smallest volume possible. The tubes were left to air-dry for 1 hour with the caps open in a chemicals fume-hood with care not to dry pellets completely. The tubes were then transferred on a heat-block at 50 °C, the DNA pellets were hydrated with 55 µl of heated at 50 °C UltraPure™ Distilled Water and were incubated for 10 min to allow the pellet to dissolve. To ensure complete dissolution of the DNA pellet the tubes were left for 48 hours at 4 °C. To confirm successful extraction of intact gDNA from all samples, 5 µl of each were run on a 1% agarose gel and for all samples in this study a single clear band larger than 10 kb band was observed. The concentration of all samples was measured using a DS-11 Spectrophotometer (DeNovix).

2.4.3 TaDa methylated DNA isolation and amplification

The protocol followed here is an adapted version of the one presented in (Marshall *et al.*, 2016) for TaDa in *Drosophila* with minimal alterations. Of the extracted gDNA samples above, a total amount of up to 5 μg was transferred into a new 1.5 ml tube and was brought to 43 μl with the addition of UltraPure™ Distilled Water. For those samples where 5 μg were not available, 43 μl of the original sample were transferred. In this study the starting amount of gDNA processed was not found to correlate with sequencing, mapping or signal quality. To each of those tubes 5 μl of 10x CutSmart Buffer (New England Biolabs) and 2 μl of DpnI restriction enzyme (New England Biolabs) were added and mixed by gentle flicking instead of pipetting to prevent shearing of gDNA. The samples were incubated at 37 °C for 16 hours overnight and were in turn cleaned-up using the QIAquick PCR Purification kit (QIAGEN) and eluted with 40 μl of 50 °C water (~35 μl are recovered). 30 μl of each clean digestion product were split equally in two 0.2 ml PCR tubes (15 μl in each) and 4 μl of *Adaptor ligation buffer* (5x T4 DNA Ligase Buffer (New England Biolabs) and 10 μM of the dsAdR adaptor) along with 1 μl (400 U) T4 DNA Ligase (New England Biolabs) were added in each. The samples were then incubated at 16 °C for 2 hours, followed by 10 min at 65 °C in a thermocycler. The double-stranded adaptor dsAdR was initially prepared by mixing equal volumes of 100 μM of the single stranded oligos AdRT (5'- CTAATACGACTCACTATAGGGCA GCGTGGTCGCGGCCGAGGA-3') and AdRb (5'- TCCTCGGCCG-3') in a 1.5 ml tube and immersing in a boiling-hot water-bath, letting it to cool down to room temperature to allow for gradual annealing.

Following the adaptor ligation each sample was mixed with 20 μl of a 2x DpnII Digestion Buffer and 10 units (1 μl) of DpnII restriction enzyme (New England Biolabs) mastermix and were incubated for 3 hours at 37 °C. At this stage for each original sample two 40 μl digestion reaction products were available. For methylated DNA amplification by PCR each one of these products was mixed with 118 μl of *DamID PCR buffer* and 2 μl (10 units) of MyTaq™ DNA polymerase (Bioline) and were aliquoted at 40 μl in 4 different 0.2 ml PCR tubes. The *DamID PCR buffer* consisted of 1.36x MyTaq™ Buffer (Bioline) and 1.06 μM of the DamID PCR primer (Adr_PCR: 5'- GGTCGCGGCCGAGGATC-3') that anneals on the adaptor sequence. In total for each original gDNA sample 8 PCR reactions were performed using the following cycling program:

Single cycle of steps 1-4:

1. 72 °C for 10 min
2. 94 °C for 30 sec
3. 65 °C for 5 min
4. 72 °C for 15 min

3 cycles of steps 5-7:

5. 94 °C for 30 sec

6. 65 °C for 1 min
7. 72 °C for 10 min

21 cycles of steps 8-10:

8. 94 °C for 30 sec
9. 65 °C for 1 min
10. 72 °C for 2 min

Final extension step:

11. 72°C 5 min
12. Slow cool down to room temperature and storing at 10 °C.

In the version of the protocol used in this study the number of cycles for steps 8-10 was increased from 17 described in (Marshall *et al.*, 2016) to 21 as is more commonly used in conventional DamID experiments in *C. elegans* (Askjaer, Ercan & Meister, 2014). 3 µl of a PCR reaction each representing a respective gDNA sample was run on a 1% agarose gel to assess the success of the isolation of amplified sequences from methylated gDNA by the presence of a smear between 2 kb and 200 bp.

Following the amplification the 8 reactions per sample were pooled in a 2 ml microcentrifuge tube (Eppendorf) and were cleaned-up using QIAquick PCR Purification kit (QIAGEN). The column washing step was repeated 3 times to better clean the column from impurity build-up due to the large volume of sample that was passed through it. The amplicons were eluted using 50 µl of 50 °C heated UltraPure™ Distilled Water (~43 µl of eluent) and their concentration was measured using a DS-11 Spectrophotometer (DeNovix). To remove the adaptor sequences from the resulting PCR products up to 2.5 µg of product was transferred to a new 1.5 ml tube and was brought up to a volume of 44 µl with water. To each, 5 µl of 10x Cutsmart Buffer and 1 µl of AlwI restriction enzyme (New England Biolabs) was added, mixed and the samples were incubated at 37 °C for 16 hours. They were cleaned-up using the QIAquick PCR Purification kit, eluted with 50 µl of 50 °C heated UltraPure™ Distilled Water and the concentration was measured as before. From these final products an amount between the total product and up to 2 µg was sent to GENEWIZ® for library preparation and Next-Generation Sequencing (NGS) using the Illumina® HiSeq 4000 paired-end 2x150 bp platform. In this study no clear correlation was found between the amount of material sent and quality of signal acquired. NGS results were returned as FASTQ files for downstream processing.

2.5 Bioinformatics and statistical analysis

2.5.1 Calculation of TaDa signal profiles and initial analysis

FASTQ files representing single-end reads for each sample and replicate, were initially assessed using the fastq-stats perl script (available at https://github.com/owenjm/damid_misc/blob/master/fastq-stats) for uncut adaptors, primer dimer and internal GATC content as a post-sequencing quality control step for the wet lab executed protocol (Appendix B.1). For transcription factor and RNA polymerase TaDa the perl script damidseq_pipeline v1.4.5 (Marshall & Brand, 2015) (available at https://github.com/owenjm/damidseq_pipeline) was used whereas for analysis of Chromatin accessibility TaDa the damidseq_pipeline_1.4.2_output_Dam_only script (available at https://github.com/tonysouthall/damidseq_pipeline_output_Dam-only_data) was used. Both pipelines were run calling, Bowtie 2 v2.3.4 (Langmead & Salzberg, 2012) for alignment to *C. elegans* bowtie indices from genome assembly WBcel235 (available from illumina® iGenomes page), Samtools v1.9 (Li *et al.*, 2009) for alignment manipulations and a GATC fragment file with the coordinates of all GATC fragments across the *C. elegans* genome in gff format, built from a WBCel235 FASTA file (available at https://www.ensembl.org/Caenorhabditis_elegans/Info/Index) using the gatc.track.marker.pl script (available at https://github.com/owenjm/damidseq_pipeline). The pipelines require a pair of FASTQ files corresponding to a Dam-fusion sample and an appropriate Dam-control as input to map to the *C. elegans* genome, calculate normalised alignment read count maps (in BAM format) and calculate $\log_2$(Dam-fusion/Dam-control) ratio scores per GATC bin, in GFF format, which constitutes the informative TaDa signal profile as described in (Marshall & Brand, 2015).

In this study 2 biological replicates have been performed for every sample and control at every stage that is investigated. Therefore, for each stage and Dam-fusion 4 relevant FASTQ files had to be processed, namely 2 replicates for the Dam-fusion of interest at the given stage and 2 replicates of the appropriate control of the same stage. Pairs of FASTQ files based on the above rationale were initially run through the pipeline. The mapping/alignment information, number of uniquely mappable reads and depth of sequencing as coverage for each sample used in this study is available in Appendix B.1. The number of uniquely mappable reads that have been acquired by the experiments presented here and that align to the genome only once, varied from over 6 million up to ~30 million reads per sample with the genome coverage being between ~9x and 45x (times) which surpasses the genome coverage reported as the threshold in previous studies (Aughey *et al.*, 2018). Samples with fewer reads were not processed. By running the pipeline in pairs of samples two ratio signal files were produced per Dam-fusion sample for L2 or L4 stages and four normalised aligned read count BAM files, one for each original FASTQ file.

Correlation between samples and reproducibility of replicates was assessed using the deeptools3 (Ramírez *et al.*, 2016) multiBamSummary (--binSize 300) and plotCorrelation (--corMethod pearson, --whatToPlot heatmap, --skipZeros, --removeOutliers) tools (on https://usegalaxy.eu/). Principal component

analysis presented here was performed using the deeptools3 plotPCA tool on the multiBamSummary-calculated read count density summary matrices. Bins with zero or very large counts were excluded from PCA and correlation analysis to prevent artificially inflating correlation.

Because the pairing of Dam-fusion and Dam-control FASTQ files between available replicates is arbitrary the BAM files were in turn used as input for the pipeline (--bamfiles option) to perform all pairwise comparisons between Dam-fusion replicates and equivalent Dam-control replicates to finally acquire four ratio files per Dam-fusion per developmental stage. The $\log_2$(Dam-fusion/Dam-control) score per GATC bin of the genome was averaged (arithmetic average) across the four files to produce a final averaged TaDa signal ratio file in GFF format for every Dam-fusion and stage that was used for all downstream processing. In the case of CATaDa the pipeline produces a Dam only GFF formatted signal file with reads per million (rpm) scores across the GATC bins of the genome for every Dam-control replicate processed.

Genomic coordinate files produced and used throughout this study were converted between formats (BED, GFF, Bedgraph, BigWIg, Wig, GTF) using Excel, the <u>Convert between GTrack/BED/WIG/bedGraph/GFF/FASTA files</u> tool of the Galaxy powered GSuite Hyperbrowser (elixir) (at <u>https://hyperbrowser.uio.no/hb/!mode=advanced</u>) and the UCSC browser binaries bedGraphToBigWig, BigWigToBedGraph, bigWigToWig. BED and GFF signal and feature track files were visualised and captured using the SignalMap NimbleGen software (Roche).

Aggregation plots and heatmaps of signal localisation preference around given genomic features were generated for the above files before or after statistical analysis using the SeqPlots GUI application (Stempor & Ahringer, 2016) with specific settings mentioned individually for each presented result. Aggregation plots represent signal averages for 10 bp bins in regions of varying but specified length around positional features of the genome. For genes all of their start and end coordinates, based on the largest transcript and used here as the transcriptional start site (TSS) and transcriptional end site (TES) of genes, are anchored to two positions of the X-axis and their genic sequence is pushed or stretched to a pseudo-length of usually 2 kb. For other features the midpoint coordinate is used to align all to the same position on the X-axis which then extents upstream and/or downstream of that region. For each position around the feature an average is calculated across all the features to generate the aggregation plot line with a shaded area representing the 95% confidence interval. When z-scores are presented on the Y-axes, those have been calculated as deviations from the mean signal seen across the plotted region.

2.5.2 Gene-calling, peak-calling and annotation of peaks

To identify transcribed genes based on the signal enrichment over gene bodies in RNApol TaDa experiments the Rscript polii.gene.call (Marshall & Brand, 2015) (available at <u>https://github.com/owenjm/polii.gene.call</u>) was used. The averaged GFF files calculated above were used as input along with a GFF file listing genes and coordinates from the WBcel235.36 assembly annotation. A false discovery rate (FDR) lower than 0.05 is used here as threshold to call expressed genes. The outputted gene lists produced here were updated to the most recent WBCel235.99 annotation (<u>https://www.ensembl.org/Caenorhabditis</u>

elegans/Info/Index) post-processing. To call expressed miRNAs using the RNA pol TaDa signal profiles the miRNA genomic coordinates used here were extended 500 bp upstream and 500 bp downstream to expand their size for better FDR assignment.

Identification of significantly enriched peaks across the genome for TF TaDa and CATaDa signal profiles was performed using the perl script find_peaks (available at https://github.com/owenjm /find_peaks) with an FDR<0.05 (--fdr=0.05) and default settings. The input for TF TaDa were the averaged GFF files and for CATaDa the per replicate GFF files. The output was a list of genomic interval coordinates for statistically significant peaks in GFF format. After identification of significant peaks for each of the CATaDa replicates the resulting files were merged with averaging of overlapping peaks into a single GFF or BED file for each tissue and stage, using bedtools merge (-o mean) on BEDtools (Quinlan & Hall, 2010) v2.25. The merged profiles were used for CATaDa downstream processing.

TF TaDa and CATaDa significant peaks in BED format were assigned to nearby genes that they may regulate using UROPA (Kondili *et al.*, 2017) as a web tool (available at http://loosolab.mpi-bn.mpg.de/UROPA_GUI/) with Caenorhabditis_elegans.WBcel235.99.gtf (from http://www.ensembl.org/ Caenorhabditis_elegans/Info/Index) as the genome annotation file. Peaks were assigned to genes on any strand when their centre coordinate was positioned up to 6 kb upstream of a gene start site or 1 kb downstream of the end site and the location of the peak relative to the gene was assigned based on the full length of the peak and the strand of the gene (options: feature gene, distance c(6000,1000), feature.anchor any_pos, direction any_directions). To avoid discarding valid regulatory relationships which is exacerbated by the compactness of the *C. elegans* genome, no prioritisation was set for peaks when multiple assignments were available and all were reported.

2.5.3 Assessment of overlaps between sets of genomic intervals or gene-sets

To identify overlapping peaks between samples or other genomic interval or features the bedtools intersect tool was used with settings dependent on the prospected outcome of the processing. To graph these overlaps as venn diagrams the venn module of the Intervene package (Khan & Mathelier, 2017) was used.

To test if sets of genomic coordinates representing various features (e.g. TaDa peaks, ATAC-seq peaks, ChIP-seq peaks, genomic features) show statistically significant overlaps across the genome, Monte Carlo simulations have been performed using the python pipeline OLOGRAM, part of the gtftk package (Ferré *et al.*, 2019). *p*-values are calculated based on the occurrence of intersections between intervals and overall length of overlap (in bp) across the genome. BED files were used as inputs with default settings for comparisons between peak profiles. BED files for input and a GTF formatted genomic annotation file were used for overlaps with genomic features.

For statistical assessment of the level of association between patterns of peak (TaDa or ChIP-seq peaks) localisation across the genome for different TFs, the IntervalStats tool (Chikina & Troyanskaya,

2012) as part of the coloc-stats webserver (https://hyperbrowser.uio.no/coloc-stats/) was used. In brief, for the TFs used in this study ChIP-seq optimal IDR-thresholded peak coordinate files from L2 animals were downloaded from the modERN (Kudron *et al.*, 2018) and modENCODE (Celniker *et al.*, 2009) databases and were combined along with the L2 TaDa TF samples into a GSuite of genomic tracks on coloc-stats. Each peak file was then used as query against the GSuite of reference sequences to calculate the IntervalStats statistic for co-localisation for all pairwise comparisons of peak coordinates. As described in (Araya *et al.*, 2014) the values in the resulting comparison matrix representing comparisons between the same two TFs with different directionality (query-reference) were averaged to symmetrise the matrix and calculate the final co-association values that were plotted as a heatmap using the R package heatmap3 and hierarchical clustering.

To assess the statistical significance of overlaps between sets of genes found in different samples, hypergeometric distribution tests were performed either on http://nemates.org/MA/progs /overlap_stats.html or using the R software package *SuperExactTest.* For both tests when sets of coding genes are compared the size of the sampling pool was set to 20191, the number of annotated coding genes in the most recent WBcel235 assembly. Representation of overlaps is either in the form of Venn diagrams generated using http://bioinformatics.psb.ugent.be/webtools/Venn/ or in the form of the output of the *SuperExactTest* package.

2.5.4 Genomic interval conservation assessment

In this study, conservation of *C. elegans* sequences was assessed where required either with a Vista analysis (Frazer *et al.*, 2004) (at http://genome.lbl.gov/vista/index.shtml) or taking advantage of the pre-calculated PhastCons7way scores (Spieth, Hillier & Wilson, 2005) on UCSC browser for the ce10 (WBcel215) assembly of the *C. elegans* genome (ce10.phastCons7way.bw). Vista analysis was performed using homologous sequences from *C. elegans*, *C. briggsae* and *C. brenneri* (from https://wormbase.org) and the rankVISTA (or rVista) tool (70% Cons Identity and 100 bp Calc Window). To calculate conservation scores for genomic intervals of interest the ce10.phastCons7way.bw track containing per base conservation scores from multiples alignments to the *C. elegans* genome of 6 *Caenorhabditis* species (*C. briggsae*, *C. brenneri, C. remanei, C. sp. 11, C. japonica, C. angaria*) was shifted to the ce11 (WBcel235) assembly coordinates, using Crossmap.py (Zhao *et al.*, 2014) (bigwig) (ce11.phastCons7way). Average PhastCons7way scores were calculated for genomic intervals of interest in BED format using the UCSC binary bigWigAverageOverBed.

2.5.5 Motif identification from TF TaDa peaks and in promoters of TaDa-identified expressed genes

To identify motifs associated with promoters of genes expressed in a certain tissue, gene sets that were found by RNA pol TaDa experiments and that are specified in each case where such results are

presented, were converted to NCBI Refseq ID names using SimpleMine (at https://wormbase.org/tools/mine/simplemine.cgi). The Refseq IDs list was used as input for the perl script findMotifs.pl of the HOMER v4.11.1 platform (Heinz *et al.*, 2010) using a prefabricated *Caenorhabditis elegans* promoter set and looking for motifs 6,8 or 10 bp long (options: worm -len 6,8,10). The size of the interrogated promoter is specified where results are presented.

Identifying motifs from transcription factor TaDa peaks was done here using the MEME-suite of tools (Bailey *et al.*, 2009) and HOMER. To use MEME, peak intervals (selected based on criteria that are specified where such results are presented) in BED format were used to extract FASTA sequences from the WBcel235 assembly using the bedtools getfasta tool. Repetitive sequences from the FASTA file were masked using RepeatMasker rmblast (at http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) and the output FASTA file was used as input for MEME v5.1.1 (at http://meme-suite.org/tools/meme) Classic mode, to identify 5 motifs with a width between 6 and 10 bp using zoops (zero or one occurrence per sequence) scoring and a 0-order background model. To use HOMER those same peak interval files were used as input for the findMotifsGenome.pl script using the ce11 genome assembly, masking of the sequences and the option to analyse the size of sequences provided by the interval file (options: ce11 -size given -mask).

The logos presented here for motifs identified using homer were generated after converting the homer positional weight matrix into a transfac matrix using the RSAT (Nguyen *et al.*, 2018) Metazoa convert matrix tool (http://rsat.sb-roscoff.fr/convert-matrix_form.cgi) and importing to Weblogo3 (Crooks *et al.*, 2004) (http://weblogo.threeplusone.com/) for logo drawing. Identification of similar known motifs to the *de novo* identified motifs was performed using the TOMTOM tool of MEME-suite. The default parameters were used and the interrogated motif matrices were compared against the JASPAR core 2018 non-redundant database.

2.5.6 Gene-set enrichment analysis

Gene-sets identified in this study were assessed for enriched gene ontology (GO) terms or association with tissue specific expression using the worbase.org Enrichment Analysis tool (Angeles-Albores *et al.*, 2016) (https://wormbase.org/tools/enrichment/tea/tea.cgi) with a *q*-value threshold of <0.1. For significant GO terms presented here the $-\log_{10}q$ value is plotted. Association of gene-sets with biological pathways was evaluated using the gProfiler gOst tool (https://biit.cs.ut.ee/gprofiler/gost) and a g:SCS calculated significance threshold of <0.05.

2.5.7 Statistical analysis

Statistical analysis for comparisons between datasets that is not covered in the above paragraphs was performed using GrapPad prims 7 (www.graphpad.com). To test differences in the mean between seam cell scoring or smFISH counting datasets, an unpaired two-tailed t-test was performed when the

comparison was between two datasets and a one-way Analysis of Variance (ANOVA) was performed when multiple datasets were compared. One-way ANOVA was followed by a Dunnet's post hoc test when the mean of multiple datasets was compared to that of a control or a Tukey's test when all pairwise comparisons between datasets were calculated. Differences in the variance between datasets were tested with a Levene's median test. Differences in proportions in binary phenotypes were tested with a Fisher's exact test. Correlation of peak intensity values between datasets of overlapping genomic features were performed using Pearson's correlation test. The significance level used throughout is $p<0.05$.

Chapter 3

Identification of LIN-22 and NHR-25 targets in the *C. elegans* epidermis using targeted DamID

3.1 Introduction

Development is a tightly regulated process guiding the reproducible formation of complex multicellular organisms from single-cell zygotes. This requires the genetic information to be decoded, in a manner that allows for specific spatiotemporal patterns of gene expression to occur, driving the development of specialised, differentiated tissues. Such regulation of gene expression is to a large extent achieved at the level of transcriptional control by activation or repression of genes by transcription factors that bind *cis*-regulatory elements that control them (Spitz & Furlong, 2012). Multiple transcription factors participate in finely tuned gene regulatory networks (GRNs) that receive inputs from signalling pathways and the environment, to govern various developmental events across systems (Davidson, 2010). Elucidating the exact structure and dynamics of these networks, by identifying participating factors, their direct targets and the type of regulation they exert, is central to understanding how genotype is transformed to phenotype during development.

The *C. elegans* epidermal seam cells are post-embryonically developmentally active. They follow a stem cell-like pattern of symmetric and asymmetric divisions throughout larval development allowing for maintenance of the tissue and production of differentiated hypodermal or neuronal cells (Sulston & Horvitz, 1977). Discovering the underlying network of factors controlling the cell fate maintenance and differentiation decisions, in the form of symmetric and asymmetric divisions, is central to stem cell research (Morrison & Kimble, 2006). Due to the highly tractable nature of *C. elegans,* the epidermis and seam cells have been proposed and used as a model system to approach such questions (Joshi *et al.*, 2010; Brabin & Woollard, 2012; Chisholm & Hsiao, 2012).

In the seam cells these events are instructed by the Wnt signalling and a regulatory network with only a few known transcription factors, the majority of which are conserved (Joshi *et al.*, 2010; Chisholm & Hsiao, 2012). Two such factors that have been shown to regulate aspects of seam cell development and are studied in this chapter are the Hairy/Enhancer of split (*Hes*)-related bHLH transcription factor LIN-22 and the nuclear hormone receptor NHR-25, orthologue of the *Drosophila ftz-f1*. *lin-22* is expressed in the H0-V4 seam cells throughout post embryonic development. It has been found to act on seam cell development, playing a role in the establishment of the correct division symmetry or asymmetry partly by antagonising Wnt signalling and suppressing ectopic neurogenesis (Wrischnik & Kenyon, 1997; Katsanos *et al.*, 2017; Waring, Wrischnik & Kenyon, 1992). NHR-25 is a major epidermal factor and has been implicated in correct vulva development, molting, seam cell development by establishment of cell contacts and neurogenesis in the T lineage (Hayes, Frand & Ruvkun, 2006; Chen, Eastburn & Han, 2004; Gissendanner & Sluder, 2000; Šilhánková, Jindra & Asahina, 2005; Shao *et al.*, 2013; Hajduskova *et al.*, 2009). It is expressed in the epidermis throughout development, however its relationship to other epidermal factors remains largely unknown. Expanding the epidermal gene network to include more participating factors and resolving the nature of the interactions between them and with their targets, can provide valuable information on stem-cell behaviour across systems.

So far the seam cell developmental regulatory network has mainly been interrogated at the level of genetic interactions between its factors and/or suspected target genes. Despite being informative, dissecting the make-up of a network based on phenotypic outcomes, lacks the resolution required to identify direct regulatory interactions and expand the network with links to previously unknown participating factors. For example, there are currently no confirmed direct targets of LIN-22 and its position in the gene network in relation to other factors is still unclear (Katsanos *et al.*, 2017). Deciphering how it brings about specific developmental outcomes requires the identification of a broad spectrum of targets that it regulates in the tissue of interest.

Identification of such targets for transcription factors has been predominantly pursued by ChIP experiments in *C. elegans* (Araya *et al.*, 2014) with only one example using DamID-chip to identify targets of DAF-16 (Schuster *et al.*, 2010). These applications have been extremely valuable but as explained in chapter 1, they have limitations such as: 1. they often lack tissue specificity or interrogate the complete spatial domain of a factor's expression, 2. they require overexpression of potentially fate changing factors or utilise artefact-prone chemical crosslinking protocols and 3. they can miss dynamic transient interactions (Aughey, Cheetham & Southall, 2019; Askjaer, Ercan & Meister, 2014). The establishment of targeted DamID (TaDa) has allowed for determination of the tissue where target identification will be performed (Southall *et al.*, 2013). It has been successfully used in *Drosophila* to dissect mechanisms of neuronal fate determination (Vissers *et al.*, 2018; Sen *et al.*, 2019) and in mammalian stem cell lines to reveal binding of pluripotency factors (Cheetham *et al.*, 2018).

In this chapter, I present the first application of the TaDa method in *C. elegans*, performing the first example of DamID sequencing-based identification of transcription factor targets in this model organism. I used two transcription factors as examples: LIN-22, which we recently recovered from a genetic screen (Katsanos *et al.*, 2017) and NHR-25, for which ChIP-seq datasets are available, to facilitate method validation. I took steps to validate the system's functionality in *C. elegans* and investigated key aspects of the protocol. The technique was found to have good reproducibility and the acquired profiles exhibited expected TF binding-related characteristics. Comparisons between TaDa and ChIP-seq datasets were performed for method assessment and binding profile evaluation based on function. The feasibility of DNA-binding motif identification using TaDa peaks was also shown. Lastly, sets of genes identified as targets were analysed for enrichment of related gene ontologies and compared with available datasets for overlaps. They were mined for specific candidate genes, related to epidermal development, for the target confirmation experiments presented here. Overall, this chapter demonstrates the feasibility of TF target identification by TaDa in the *C. elegans* epidermis. It highlights the plethora of information that can be acquired in this approach and refines the positions of LIN-22 and NHR-25 within the regulatory network, through the confirmation of novel TaDa-predicted direct links with factors that participate in epidermal development.

3.2 Results

3.2.1 The TaDa transgene configuration prevents Dam associated toxicity and saturated methylation

One of the main limitations of conventional DamID since its establishment, has been that constitutive or tissue-specific expression of Dam-fusions were described to cause saturated, non-targeted methylation and toxicity (van Steensel & Henikoff, 2000; van Steensel, Delrow & Henikoff, 2001). This limitation that precluded tissue-specific applications was addressed by TaDa with the introduction of a primary ORF of *mCherry* followed by two STOP codons and a frameshift, preceding the Dam-fusion (Figure 1.5). This configuration allowed the use of tissue-specific promoters and has been shown to prevent toxicity while generating tissue-specific methylation patterns in *Drosophila* and mammalian cell lines (Southall *et al.*, 2013; Cheetham *et al.*, 2018). In *C. elegans,* induced ubiquitous expression by a heat-shock promoter has been shown to produce saturated methylation (Schuster *et al.*, 2010) but the effects of tissue-specific expression of Dam-fusions on methylation levels and toxicity has not been investigated.

To assess whether the TaDa transgene configuration is better tolerated over direct Dam expression and would be preferred for applications in the *C. elegans* epidermis, constructs including or lacking a primary ORF were produced (Figure 3.1A). In more detail, for all the epidermis-specific TF target identification presented here, the promoter of *wrt-2* (*wrt-2p*) was used, which drives expression predominantly in the seam cells and to a lesser extent in the hypodermis and rectum (Aspock *et al.*, 1999; Cao *et al.*, 2017). LIN-22, with an N-proximal predicted DNA binding domain, was fused upstream to *dam* and a control with equivalent fusion between *NLS-GFP* and *dam* was produced (Figure 3.1A). For each fusion a version with and without a *C. elegans* optimised *mCherry* (*wormCherry*) as primary ORF was generated (Figure 3.1A) and inserted in the genome as a single-copy transgene. The resulting transgenic lines lacking *wormCherry* expressed Dam-fusions at *wrt-2* native levels and showed a moderate dumpy phenotype and significantly reduced population growth speeds. In contrast, the lines with the TaDa configuration transgenes were generally comparable to wild-type (WT) in growth and morphology.

To further dissect the slow population growth phenotype, developmental speed was compared between the above lines and wild-type animals. WT *C. elegans* larvae reach adulthood approximately 46 hours after hatching in standard lab conditions at 20 °C (Corsi, 2006). To assess deviations from that benchmark, eggs were seeded on plates for each line and the proportion of animals that had reached adulthood 48 hours later was scored. No significant difference was found in the developmental speed compared to the WT for any of the transgenic lines (Figure 3.1B). However, when the brood size for each of these lines was assayed, those carrying fusions lacking *wormCherry* showed highly significant reductions in the number of progeny they produced (*p*<0.0001) compared to WT. The mean brood size for *wrt-2p::lin-22:dam* and *wrt-2p::NLS-GFP:dam* transgenics was 83.6 and 21.8 respectively, compared to

233.8, 229.5 and 176.8 for WT, *wrt-2p::wormCherry::lin-22:dam* and *wrt-2p::wormCherry::NLS-GFP:dam* respectively.

In the course of these assays, it was also observed that both lines lacking the primary ORF, displayed a penetrant exploding vulva phenotype (Figure 3.1D). This phenotype was found to occur in 54.5% of the *wrt-2p::lin-22:dam* animals and 75% of the *wrt-2p::NLS-GFP:dam* animals but not the WT or animals with TaDa-design transgenes and was often accompanied by intestinal expulsion that obstructed egg-laying, leading to premature lethality. By the 4th day of adulthood 73% and 100% of animals from the above lines were dead respectively. The substantial toxicity observed by both Dam-fusions when *wormCherry* is lacking, that is in the presence or absence of the TF, rules-out overexpression of LIN-22 as the reason for the observed phenotypes. Instead, it points to the effects of increased Dam presence possibly leading to over-methylation as the culprit. The above results indicate toxicity of Dam expression under tissue-specific promoters.

To identify DNA sequences interacting with a Dam-fusion, methylated gDNA is extracted and amplified by PCR prior to identification by NGS. The product should appear on a standard agarose gel as a smear between 200 bp and 2 kb when methylated DNA has been effectively detected and can be used as a proxy of methylation levels and protocol success (Marshall *et al.*, 2016). Besides toxicity, saturated methylation that hinders target identification, has been shown to result from tissue-specific expression levels of Dam-fusions (van Steensel & Henikoff, 2000; Southall *et al.*, 2013). Therefore, presence and intensity of the smear were used as measures to assess if the protocol used in this study could successfully detect methylated DNA from the epidermis and also to determine whether saturated methylation was occurring when expression was driven in the absence of a primary ORF. Smears were detected both for transgenic lines with a primary *wormCherry* ORF and lacking it, whereas a control strain that does not harbour a dam transgene exhibited minimal amplification (Figure 3.1E) when grown on *dam⁻ E. coli*. Moreover, when the control and *wrt-2p::wormCherry::lin-22:dam* carrying strain were grown on the common *dam⁺* OP50 *E. coli* as food, strong smears were recorded (Figure 3.1E) due to amplification of bacterial GATC-methylated DNA contamination, further verifying that the amplification products found in the absence of *dam⁺* bacteria represent Dam-fusion generated DNA methylation. Notably, the smears produced when *wormCherry* is lacking are considerably more intense signifying increased methylation, as previously reported. Due to the increased smear intensity and associated toxicity the feasibility of target identification using the *wormCherry* lacking strains was not further investigated.

3.2.2 Constructs for LIN-22 and NHR-25 TaDa drive expression in the seam cells, show fusion functionality and produce methylation

In this study, TaDa is used to identify direct targets of two transcription factors, LIN-22 and NHR-25. These two factors were chosen for the following reasons. First, LIN-22 is one of the factors we previously recovered from a variable seam cell number screen and characterised its function in the epidermis.
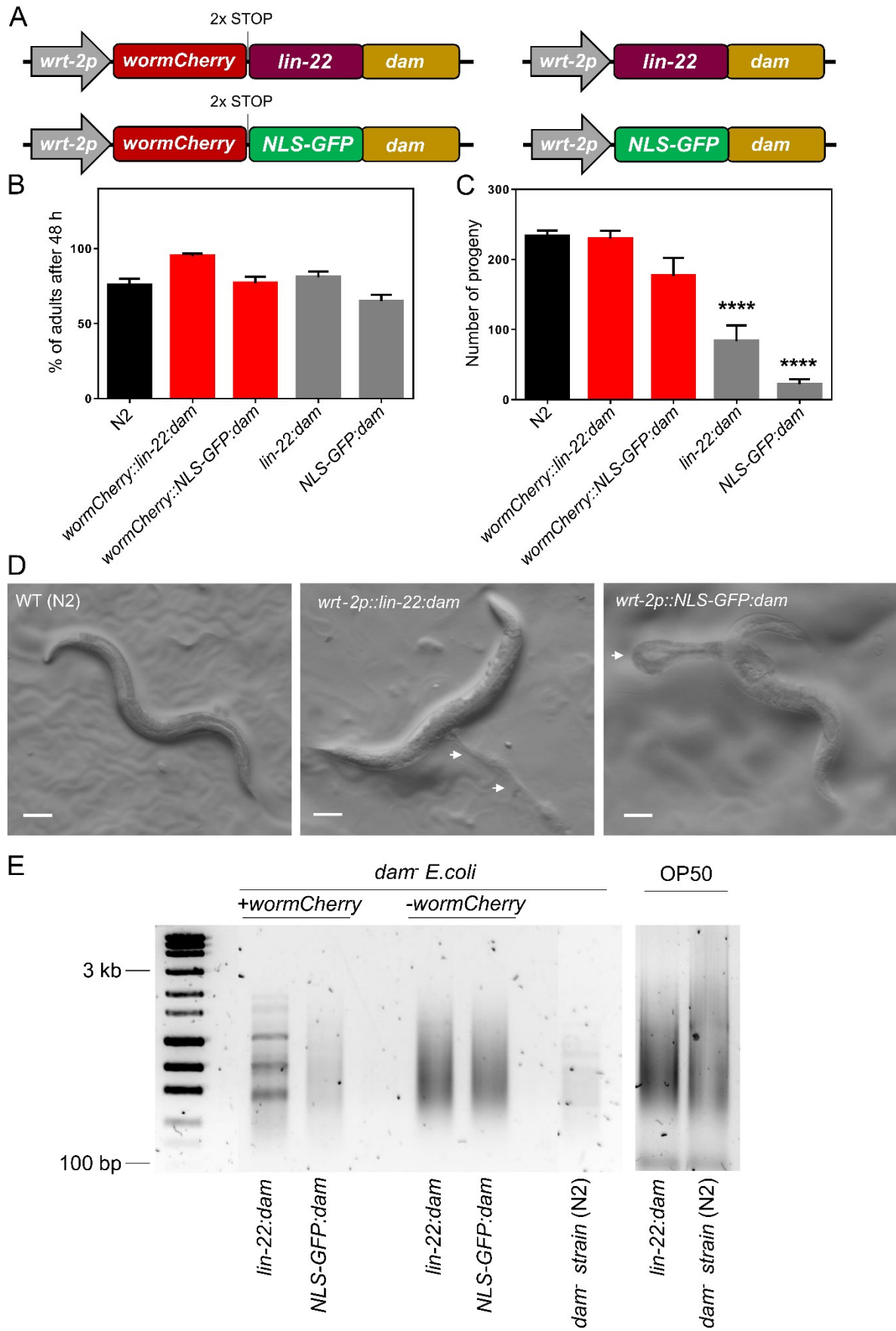
Figure 3. 1 The TaDa construct configuration prevents toxicity and saturated methylation. **(A)** Illustration of the key features of single-copy transgenes used to assess the effect of a *wormCherry* primary ORF (TaDa configuration) in preventing toxicity and high levels of non-specific methylation. Expression of the *lin-22:dam* and the respective control *NLS-GFP:dam* fusions is driven mainly in seam cells using the *wrt-2* promoter. **(B)** Proportions of animals from strains carrying the transgenes outlined in (A) that have reached adulthood 48 hours after synchronization by egg preparation as a measure of developmental speed. No significant reduction in developmental speed was observed between WT (N2) (n=107) animals carrying TaDa transgenes with the *lin-22:dam* (n=161) or *NLS-GFP:dam* (n=99) fusions as secondary ORFs and transgenic animals carrying the same fusions as primary ORFs (*lin-22::dam* n=112, *NLS-GFP:dam* n=128). Error bars indicate standard error of the proportion. **(C)** Quantification of brood size in the above strains. In the absence of the *wormCherry* primary ORF, significant reduction in brood size was found for both fusions. The "n" number of tested hermaphrodites were 15, 15, 11, 13, 12 in the order they appear on the graph. Error bars indicate SEM and black stars indicate statistically significant changes of the mean determined by a one-way ANOVA test (**** $p<0.0001$). **(D)** Representative brightfield images of adult WT and transgenic animals with constructs lacking the primary ORF, showing the exploding vulva phenotype. White arrows indicate tissue outside the body. Scale bars in D are 100 µm. **(E)** Amplification products from methylated gDNA extracted from the strains described above and a WT dam⁻ strain fed on dam⁻ *E. coli.* All samples except for the *dam⁻* strain show a pronounced 2 kb to 200 bp smear. Wells for all samples were loaded with the same volume of reaction product. Note the marked increase in methylation observed in the absence of *wormCherry.* Strains grown on *dam⁺* OP50 show extensive amplification due to bacterial GATC-methylated DNA.

Second, NHR-25, which is also implicated in seam cell development, was selected due to the availability of existing ChIP-seq datasets that would allow for comparisons and assessment of the results obtained.

TaDa constructs were assembled, as previously described (Southall *et al.*, 2013), with the promoter of *wrt-2* driving expression of a *wormCherry* primary ORF followed by two STOP codons, a nucleotide for frame-shift and the TF-*dam* fusions followed by the 3' UTR from *unc-54*. Both TFs were inserted in-frame upstream of *dam* to minimise the potential obstruction of the DNA binding domains that are N-proximal in both proteins. They were predicted by InterProScan (Jones *et al.*, 2014) to be between the 19th to 84th amino-acid out of 173 for LIN-22 and between the 15th to 90th out of 572 for NHR-25. As a common control for both TFs, *NLS-GFP* was also fused upstream of *dam* and all 3 constructs were inserted as single-copy transgenes, on the same genomic locus of chromosome II (*ttTi5605*). Key features of the transgenes are illustrated in Figure 3.2A and in the rest of this chapter the fusion names (*lin-22:dam*, *nhr-25:dam*, *NLS-GFP:dam*) are used as shorthand to indicate results associated with each fusion. All transgenes were assessed for expression, Dam-fusion functionality and methylation capacity before further use.

To confirm that all transgenes drove expression in the expected tissue, which is dictated by *wrt-2p* and is expected to be primarily in the seam cells, *wormCherry* expression was used as proxy to investigate. Microscopy for all three transgenic lines at the L4 stage showed expression in the seam cells in all observed animals (Figure 3.2B). Absence of GFP signal in *NLS-GFP:dam* transgenics (not shown) advocated to the greatly reduced, almost undetectable, expression levels of the secondary ORF, achieved by the TaDa construct configuration as previously reported (Southall *et al.*, 2013).

Another important question to address is whether the TFs fused to Dam have retained their functionality and DNA-binding characteristics. This would strongly indicate that the identified methylation profiles would represent the native binding of the investigated TFs. The *lin-22:dam* construct was injected
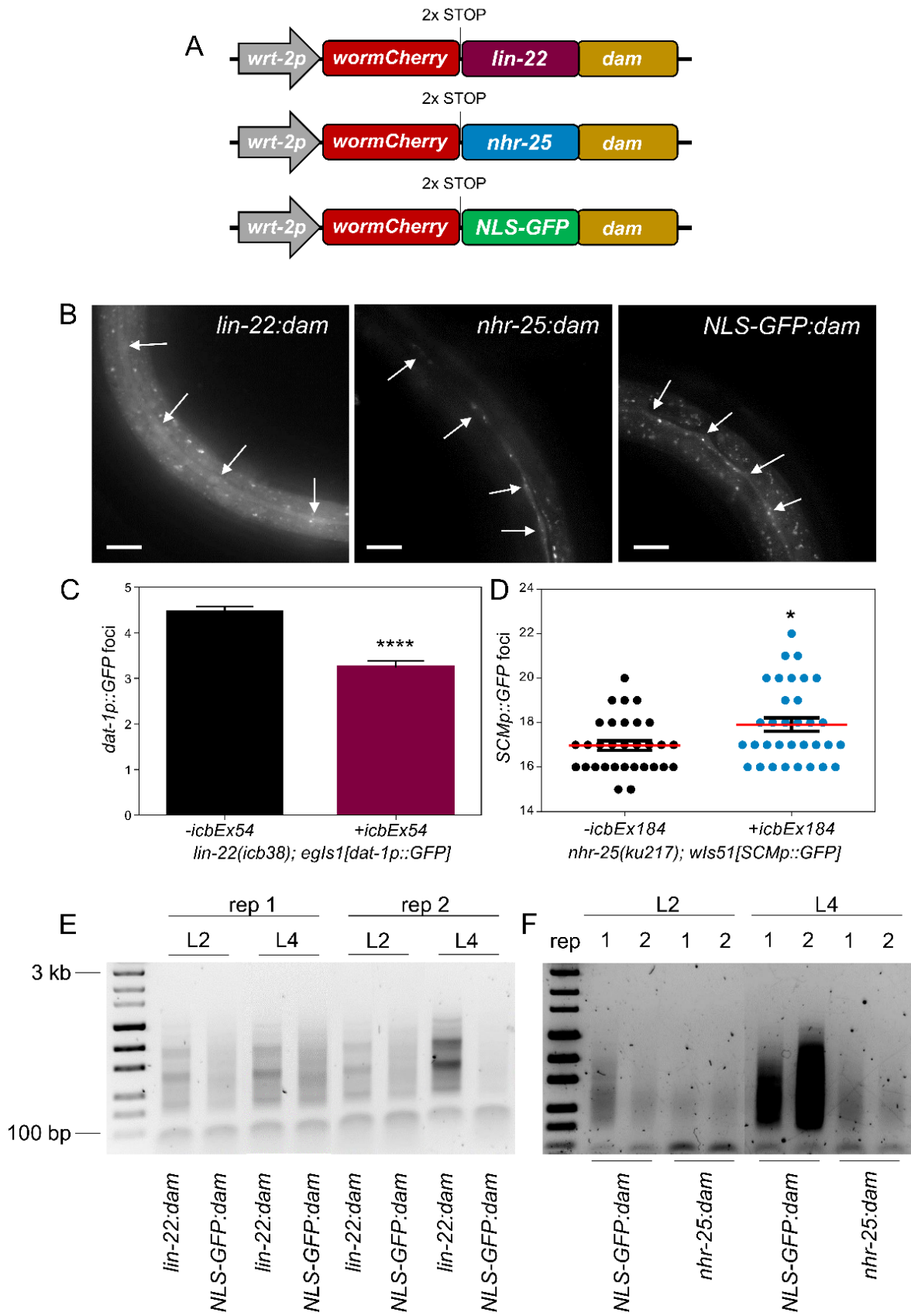
Figure 3. 2 Assessment of the *lin-22* and *nhr-25 dam*-fusion transgenes confirms tissue specific expression, fusion functionality and methylation capacity. (A) Illustration of the key features of single-copy transgenes used in this study for LIN-22 and NHR-25 target identification by TaDa. **(B)** Confirmation of single-copy transgene expression in the tissue of interest using *wormCherry* expression as a proxy. Animals were imaged at the L4 stage. White arrows indicate expression in the seam. Scale bars are 20 µm. **(C)** Overexpression of the *lin-22:dam* TaDa construct as a multi-copy transgene (*icbEx54*) in the *lin-22(icb38)* mutant significantly decreases the number of PDE neurons labelled by *dat-1p::GFP* (n=31 for –*icbEx54*, n=34 for +*icbEx54*). **(D)** Overexpression of the *nhr-25:dam* fusion as a multi-copy transgene (*icbEx184*) in the *nhr-25(ku217)* mutant significantly increased the mean seam cell number (n=31 for –*icbEx184*, n=33 for +*icbEx184*). **(E-F)** Amplification products from methylated gDNA showing 200-2000 bp smears for the LIN-22 (E) and NHR-25 target (F) identification TaDa experiments. Extractions were performed at L2 and L4 stages, with each combination of Dam-fusion and stage being represented by two biological replicates. The same volume of amplification product was loaded for each sample in each gel. Note that band patterns mostly visible in samples from the *lin-22:dam* strain are somewhat reproducible. In C and D error bars indicate the SEM and black stars indicate statistically significant differences in the mean with a t-test (* $p<0.05$, **** $p<0.0001$).

in the putative loss-of-function *lin-22(icb38)* mutant to generate a multi-copy extrachromosomal array transgene (*icbEx54*). *lin-22* mutants exhibit ectopic postdeirid (PDE) neurons, because V1-V4 seam cells undergo a transformation to the V5 fate at the L2 stage (Wrischnik & Kenyon, 1997). This results in an increase in the PDE numbers from 1 per lateral side in wild-type to a mean of approximately 4.5 in *lin-22* mutants. The reasoning was that due to the multiple copies of the transgene, the collective expression of *lin-22:dam* from the TaDa construct might suffice to rescue the phenotype if the *lin-22:dam* fusion was functional. Indeed, *lin-22(icb38)* mutants carrying the *icbEx54* transgene showed a significant (*p*= 1.15101e-08) reduction in the mean number of PDEs from 4.45 to 3.24 (Figure 3.2C), as visualised using the *dat-1p:GFP* marker, partially rescuing the phenotype and suggesting that LIN-22 in fusion with dam has retained its functionality.

Similarly, the *nhr-25:dam* construct was injected in the reduction-of-function *nhr-25(ku217)* mutant to generate a multi-copy transgene (*icbEx184*). The *nhr-25(ku217)* mutant has been found to show an increase in the mean seam cell number, compared to wild-type, from approximately 16 to 17.5 (Chen, Eastburn & Han, 2004). Here, mutants carrying the transgene showed enhancement of the phenotype with a significant (*p*=0.0155) increase in the mean from 16.96 to 17.9 (Figure 3.2D). The effect on the phenotype suggests that the capacity of NHR-25 to control seam cell development is maintained when in fusion with Dam. The fact that the overexpression further amplifies the *nhr-25(ku217)* phenotype might result from the expression by the *wrt-2* promoter in a domain different to the native or due to previously reported self-regulation (Shao *et al.*, 2013).

The target identification by TaDa for LIN-22 and NHR-25 was executed in two different experiments (Figure 3.2E, F). The DamID protocol for the *nhr-25:dam* strains and experiment was executed by Maximilien Biguet a Master's student I supervised. Material was collected at the L2 and L4 stages and each combination of fusion and developmental stage was represented by 2 biological replicates. As confirmation that all fusions at all stages were generating methylation that would allow target identification, the amplification products from isolated GATC-methylated gDNA were run on standard agarose gels. The

presence of smears between 200 bp and 2 kb for all samples (Figure 3.2E, F) indicated successful methylation. Interestingly, the *lin-22:dam* samples showed some reproducible distinct bands within their smears (Figure 3.2E) that were not observed in *NLS-GFP:dam* or *nhr-25:dam* samples and could indicate amplified fragments corresponding to regions of increased binding. Reproducible sample-specific smearing patterns are hard to interpret but they were considered encouraging signs of fusion-specific methylation.

3.2.3 Sequencing results reveal replicate reproducibility and distinct TF signatures

The amplification products for the two experiments underwent next-generation sequencing to identify targets for the transcription factors. The sequencing results were processed using the damidseq-pipeline (Marshall & Brand, 2015) to align reads onto the *C. elegans* genome and generate normalised aligned read count maps for every sample. These maps were then used to calculate the final DamID normalised $\log_2$(*TF:dam/NLS-GFP:dam*) ratio score, per GATC fragment of the genome, between pairs of *TF:dam* and *NLS-GFP:dam* sequencing results of the same experiment and stage. The sequencing yielded between 6 and 30 million, single-end, 150 bp-long reads per sample with a sequencing depth ranging between 9x and 44x times genomic coverage (Appendix B.1), which is comparable or above what was used in previous studies (Aughey *et al.*, 2018; Marshall *et al.*, 2016; Askjaer, Ercan & Meister, 2014).

The sample-specific normalised sequence alignment read count maps were used to evaluate replicate reproducibility and assess correlation between samples by calculating Pearson's correlation coefficient matrices. High levels of reproducibility between biological replicates for all fusions was found in both experiments, with Pearson's correlation coefficient values between 0.94 and ~1 for the LIN-22 experiment (Figure 3.3A) and 0.88 and 0.98 for the NHR-25 experiment (Figure 3.3B). Differences in the range of values occupied by Pearson's coefficients amonst the two experiments for the TFs (Figure 3.3A, B) most likely reflect differences in technical manipulations. The two experiments were executed at different times but included the same control strains that show somewhat different levels of correlation in each experiment, thus advocating to the technical source of some of the observed discrepancy.

Importantly, all Dam-fusions showed high correlation coefficients across samples from L2 and L4 stages but low correlation coefficients between *TF:dam* and *NLS-GFP:dam* samples in both experiments. As expected, the *TF:dam* samples likely represent largely stage-independent methylation patterns that correspond to genome-wide TF binding preferences that differ notably to those of the non-targeted *NLS-GFP:dam* controls. When Pearson's correlation calculations were made across all samples from the two experiments, high correlation coefficients, between 0.82 and 0.95, were observed between control *NLS-GFP:dam* samples of the same stage (Figure 3.3C). This further supports the across-experiment reproducibility of the protocol. In addition, this high degree of similarity likely advocates to the fact that although methylation from control-Dam fusions occurs serendipitously, by the freely diffusing fusion, it represents accessible chromatin thus producing reproducible methylation patterns.

When analysed together, control samples from both experiments cluster together but separately from TF samples, both by Pearson's correlation and principal component analysis (PCA) (Figure 3.3C, D), further supporting that TFs show methylation preferences different from the non-targeted control samples. Tight grouping of control samples even across experiments likely rules out inter-experimental variation as the major contributor to differences observed between the studied TFs. Moreover, the PCA indicates higher similarity between the samples from the two TFs than with those of the controls. Based on their distance on the PCA space when analysed along with thecontrol samples, the proximity of read-count maps from replicates of the two TFs potentially point to some degree of similarity in binding preference. However sample comparisons across experiments revealed separate clustering of *lin-22:dam* and *nhr-25:dam* samples, with low correlation values between them (Figure 3.3C) indicating capture of different TF-specific preferences. PCA clearly illustrates this point when control samples are omitted by distinct grouping of *lin-22:dam* and *nhr-25:dam* samples (Figure 3.3E).

The meaningful TaDa signal enrichment profiles that represent a more accurate image of the expected binding, as described above, corresponds to tracks of normalised $\log_2(TF:dam/NLS\text{-}GFP:dam)$ ratio scores for each GATC fragment of the genome (Marshall & Brand, 2015). Considering that two biological replicates are available for each sample, there are 4 potential pairwise calculations between TF-fusion and control-fusion samples that can be executed for each TF at each stage. Due to the increased reproducibility between replicates described above, the 4 profiles where found to be very similar qualitatively (Appendix C.1). These 4 profiles were arithmetically averaged per GATC fragment to generate a single representative profile per TF and per developmental stage (Figure 3.3F) that was used for downstream processing (complete genome-wide signal profiles available in Appendix C.2, C.3). These profiles showed signal enrichment, represented as peaks scattered across the genome that are expected to correspond to frequently methylated sites due to TF binding. This approach prevented loss of genuine or gain of false signal that could result from small differences in some but not all combinations of TF and control samples used for the calculation.

3.2.4 Identified TaDa signal enrichment profiles and peaks associate with putative regulatory regions of the genome

As a first line of evidence to further corroborate that the observed signal enrichment is biologically meaningful and specifically represents each TF's binding, the signal enrichment across previously predicted or confirmed targets was examined. Such examples of close-up inspections are presented in Figure 3.4A, B. For *lin-22:dam,* signal enrichment forming significant peaks (FDR<0.05) was observed upstream of the *frizzled* receptor *lin-17*, the Hox gene *mab-5* and throughout the *C. elegans achaete-scute* homologue *lin-32* (Figure 3.4A). We had previously shown that LIN-22 inhibits *lin-17* and *mab-5* expression in anterior to V5 lineages (Katsanos *et al.*, 2017), although it was not clear whether this relationship is direct or not. Moreover, both *mab-5* and *lin-32* had been predicted through genetics to interact with *lin-22,*
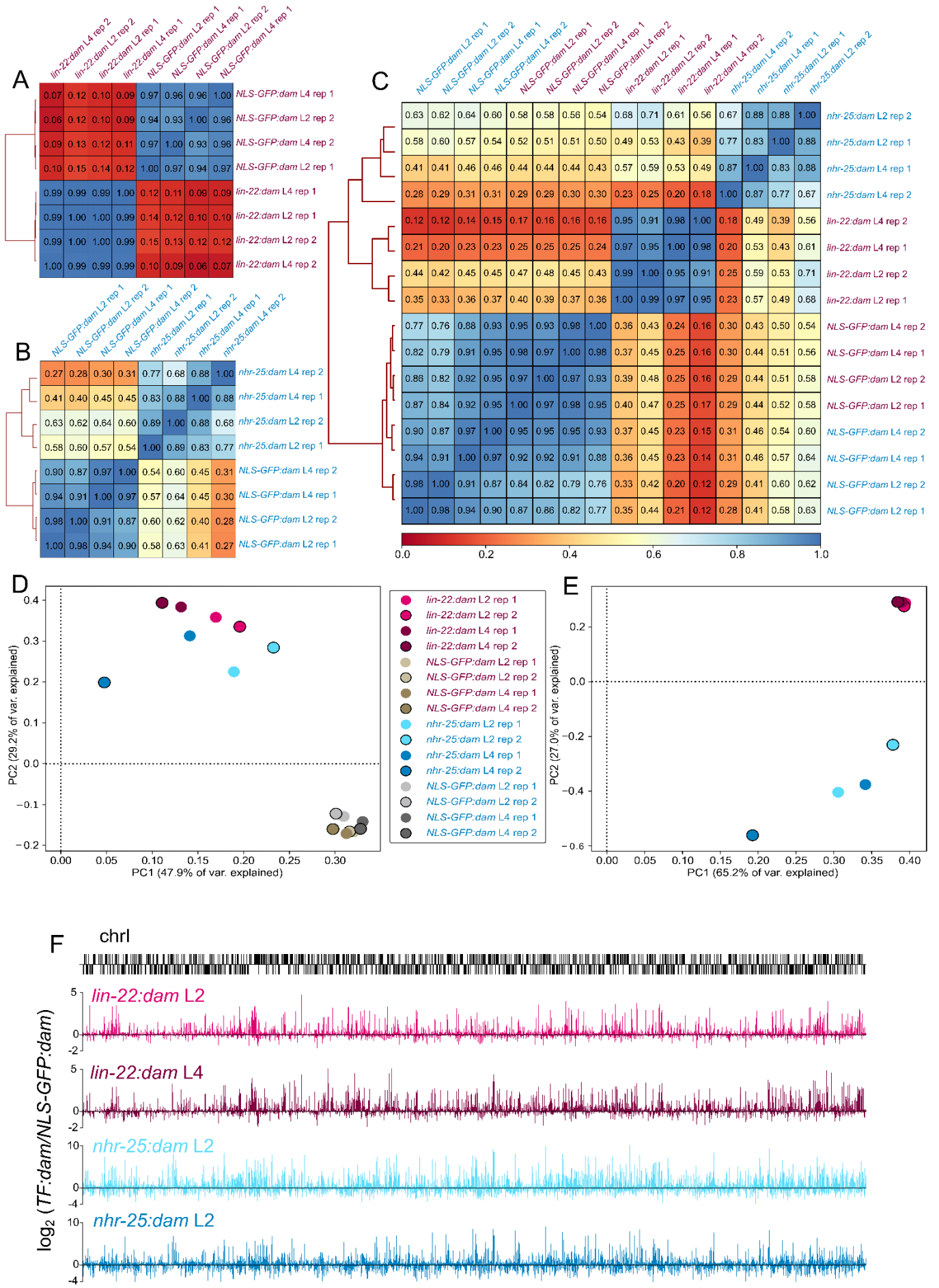
Figure 3. 3 Analysis of sequencing results reveals replicate reproducibility and fusion-dependent methylation. **(A-B)** Pearson correlation heatmaps based on normalised aligned read count maps for the LIN-22 (A) and NHR-25 (B) TaDa experiments. The correlation coefficient for each pairwise comparison is printed in each cell of the heatmaps. All Dam-fusions show strong replicate reproducibility and high within-fusion correlation, indicated by high correlation coefficients. TF and control samples show low correlation between them and cluster separately. **(C)** Summary heatmap of Pearson correlations between all samples for the TF TaDa performed in this study. Note the high correlation between control samples from different experiments and low correlation with distinct separate clustering between the *lin-22:dam* and *nhr-25:dam* samples. Samples belonging to the LIN-22 or NHR-25 experiment are indicated with purple or blue writing respectively. **(D)** Principal component analysis on normalised aligned read count maps for all samples shows distinct grouping between TF and control fusions. **(E)** Principal component analysis only on *lin-22:dam* and *nhr-25:dam* replicates shows distinct grouping between the two TF samples. Key is shared between D and E. **(F)** Example of averaged signal enrichment profiles across proein coding genes of chromosome I (shown as black bars) for *lin-22:dam* and *nhr-25:dam* fusions in L2 and L4 stages. The Y-axes represent normalised $log_2(TF:dam/NLS-GFP:dam)$ scores. Scale bar length is 2 Mb.

with *achaete-scute* having been described as a target of LIN-22 orthologues (*Hes*-related factors) in other systems (Ohsako *et al.*, 1994; Sasai *et al.*, 1992). For *nhr-25:dam* profiles, signal forming significantly enriched peaks, is shown in the promoters of previously identified by ChIP-seq confirmed target genes *idh-1* and *rpl-3* (Shao *et al.*, 2013)*,* constituting preliminary evidence of overlap between the two methods (Figure 3.4B left, middle). Additionally, evidence of previously reported self-regulation of *nhr-25* (Shao *et al.*, 2013) is also identified by TaDa with signal enrichment in its vicinity (Figure 3.4B right).

The preference for increased signal enrichment in the upstream region of genes was also captured at the genome-wide level. All profiles showed increased average enrichment scores in regions upstream to the transcriptional start site (TSS) of protein coding genes, with the highest scores recorded for proximal to the TSS upstream sequences (Figure 3.4C). Genic regions showed comparably much lower average scores confirming the mostly intergenic localisation of the signal, in keeping with expected TF binding characteristics.

Statistical processing of the signal profiles, to determine regions with statistically significant enrichment (threshold of FDR<0.05), identified a selection of scored peaks (or peak profiles) for each TF at each stage, that represent potential target sites. For the rest of the chapter "peaks" refer to statistically significant regions. 1965 and 1972 peaks were identified for *lin-22:dam* at the L2 and L4 stage respectively, while 2044 and 2169 peaks were found for *nhr-25:dam* at the L2 and L4 stage respectively. Hierarchical clustering of the localisation and score of those peaks that at least partially lie within 5 kb upstream to 1kb downstream of genes, demonstrated broadly similar localisation preferences to those indicated by the aggregate genome-wide signal profiles over the same regions (Figure 3.4D). This substantiated the status of the peaks as potential binding sites for LIN-22 and NHR-25.

The localisation of the peaks in relationship to genes was further dissected by assigning each peak only to the closest gene when the centre of the peak was positioned within 6 kb upstream to 1 kb downstream of the TSS or TES of a gene respectively. The peak was then annotated relative to that gene's genomic location. For all peak profiles, between 94% to 96% of the peaks were assigned to genes (exact
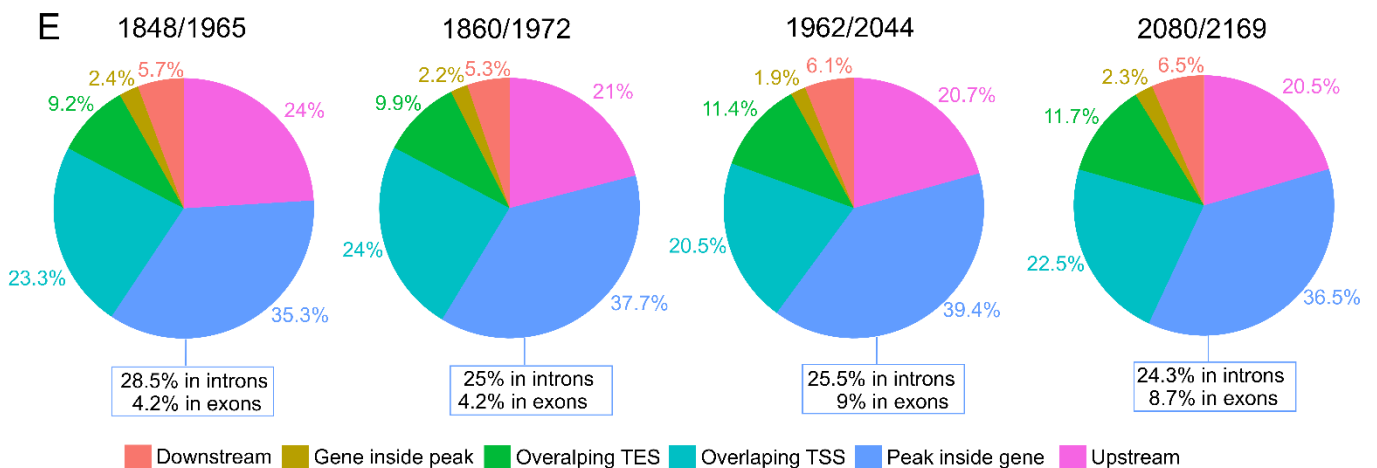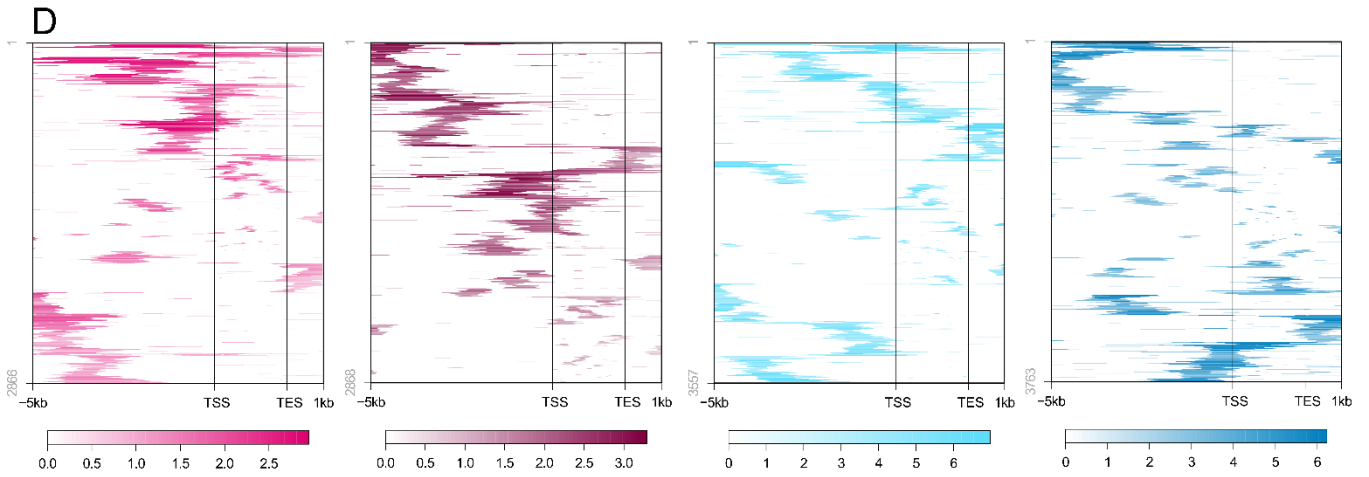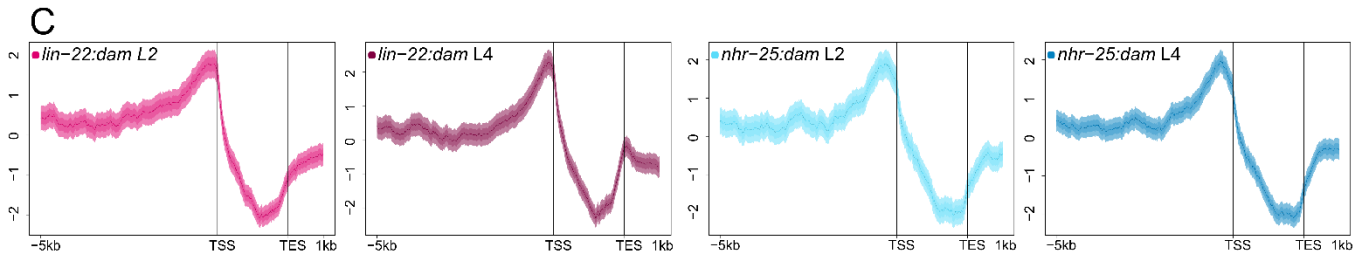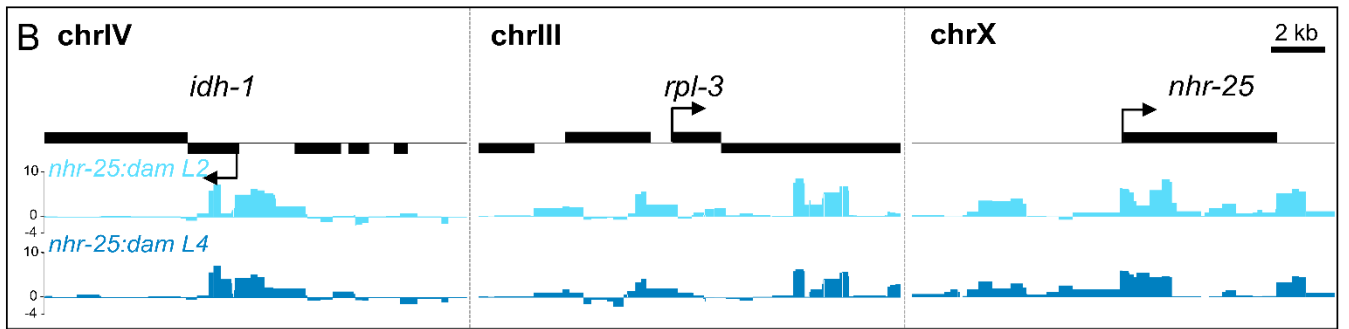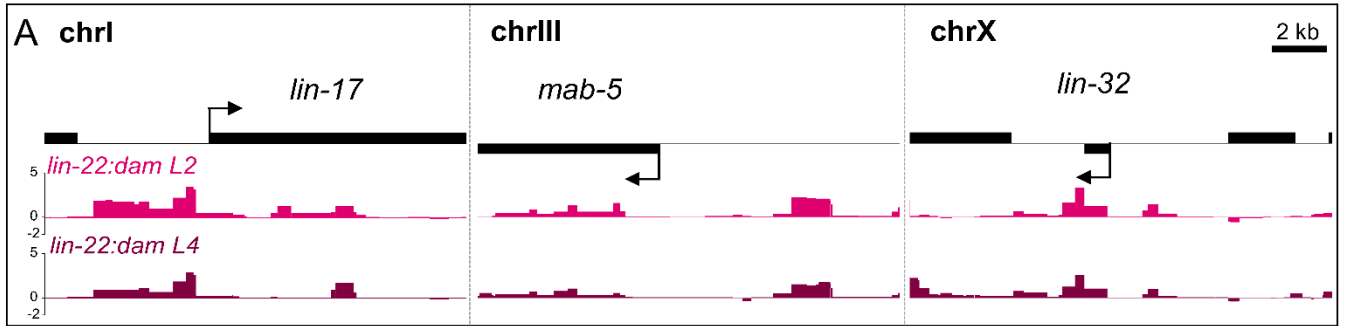
Figure 3. 4 LIN-22 and NHR-25 signal is enriched upstream of genes in putative regulatory regions. **(A-B)** Examples of normalised signal profiles in regions with statistically significant enrichment peaks. For *lin-22:dam* (A) enrichment is shown in promoters of the previously predicted LIN-22 targets *lin-17*, *mab-5* and *lin-32,* while for *nhr-25:dam* (B) enrichment is shown around *nhr-25* and in the promoters of previously confirmed targets *idh-1* and *rpl-3*. The Y-axes represent normalised log$_2$*(TF:dam/NLS-GFP:dam)* scores*. Scale bar length is 2 kb as indicated. **(C)** Aggregation plots for all indicated TF-Dam fusions in all stages, showing average enrichment scores in 10 bp bins for regions of the same length across all of the specified features on the X-axis of the genome. Strong enrichment preference is seen for upstream to genes regions. Plots show 5 kb upstream of the TSS of genes to 1 kb downstream of the TES, with gene bodies pushed into a 2 kb pseudo-length. Y-axes are z-scores for the plotted sequence length and shaded areas represent 95% confidence intervals. **(D)** Heatmaps representing the hierarchically clustered localisation and enrichment score of all statistically significant peaks (FDR<0.05) within 5 kb upstream and 1 kb downstream of a gene. Note that the majority of significant peaks reside in proximal upstream to genes regions. **(E)** Proportions of single-gene assigned peaks residing in different genomic locations, relative to the genes, indicate localisation preference for upstream to the TSS regions for both TFs and stages. For proportion calculations peaks were assigned only to the closest gene, if any, when their centre coordinate localised within 6 kb upstream and 1 kb downstream of the TSS and TES respectively. Ratios above pie charts are the number of assigned peaks to the total number of significant peaks found. The proportions of peaks within genes with exclusive intron or exon localisation is indicated under the pie charts.

numbers shown as ratios on Figure 3.4E) illustrating that almost all peaks associate with at least one gene. The largest proportion of peaks, between 41.2% and 47.3%, for both TFs and all developmental stages, were upstream to their assigned gene, with over half of them overlapping the TSS (Figure 3.4E). Of the peaks fully positioned upstream of genes, approximately half were localised within the first 2 kb upstream (46%, 47%, 52%, 54% of the upstream peaks for *lin-22:dam* L2 and L4 and *nhr-25:dam* L2 and L4 respectively), a region reported to harbour the majority of binding sites for multiple TFs in *C. elegans* promoters (Araya *et al.*, 2014). In addition, a significant proportion of peaks were found to be within genes but around a quarter of those peaks (between 24.3% and 28.5%) were exclusively residing within introns, which are known to contain regulatory elements in *C. elegans* (Fuxman Bass *et al.*, 2014). In stark contrast only between 4.2% and 9% of peaks inside genes where exclusively exonic, despite exons being more frequent in the genome and their median size being double that of introns (Spieth *et al.*, 2014). Overall, these findings highlight the regulatory potential of the identified peak regions, which could be exerted on the associated target genes.

To further investigate the relationship of the identified peaks to regulatory elements, open chromatin signal tracks from ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) experiments on whole animal L2 and L4 *C. elegans* (Jänes *et al.*, 2018) were mapped onto the LIN-22 and NHR-25 L2 and L4 peaks respectively. TaDa peak sites for all factors showed strikingly increased average chromatin openness compared to neighbouring regions (Figure 3.5A). Moreover, the majority of the peaks for all factors showed overlaps with increased chromatin accessibility regions (Figure 5B). More specifically, only between 11% and 27.6% of the total identified peaks for each TF at each developmental stage did not show an overlap with an open chromatin sequence element (Figure 3.5C) and the overlaps for all profiles were highly statistically significant by Monte Carlo simulations (p<5.3e-279).

A

Open chromatin ATAC-seq L2

Open chromatin ATAC-seq L4

B

LIN-22 L2
TaDa peaks

NHR-25 L2
TaDa peaks

LIN-22 L4
TaDa peaks

NHR-25 L4
TaDa peaks

C
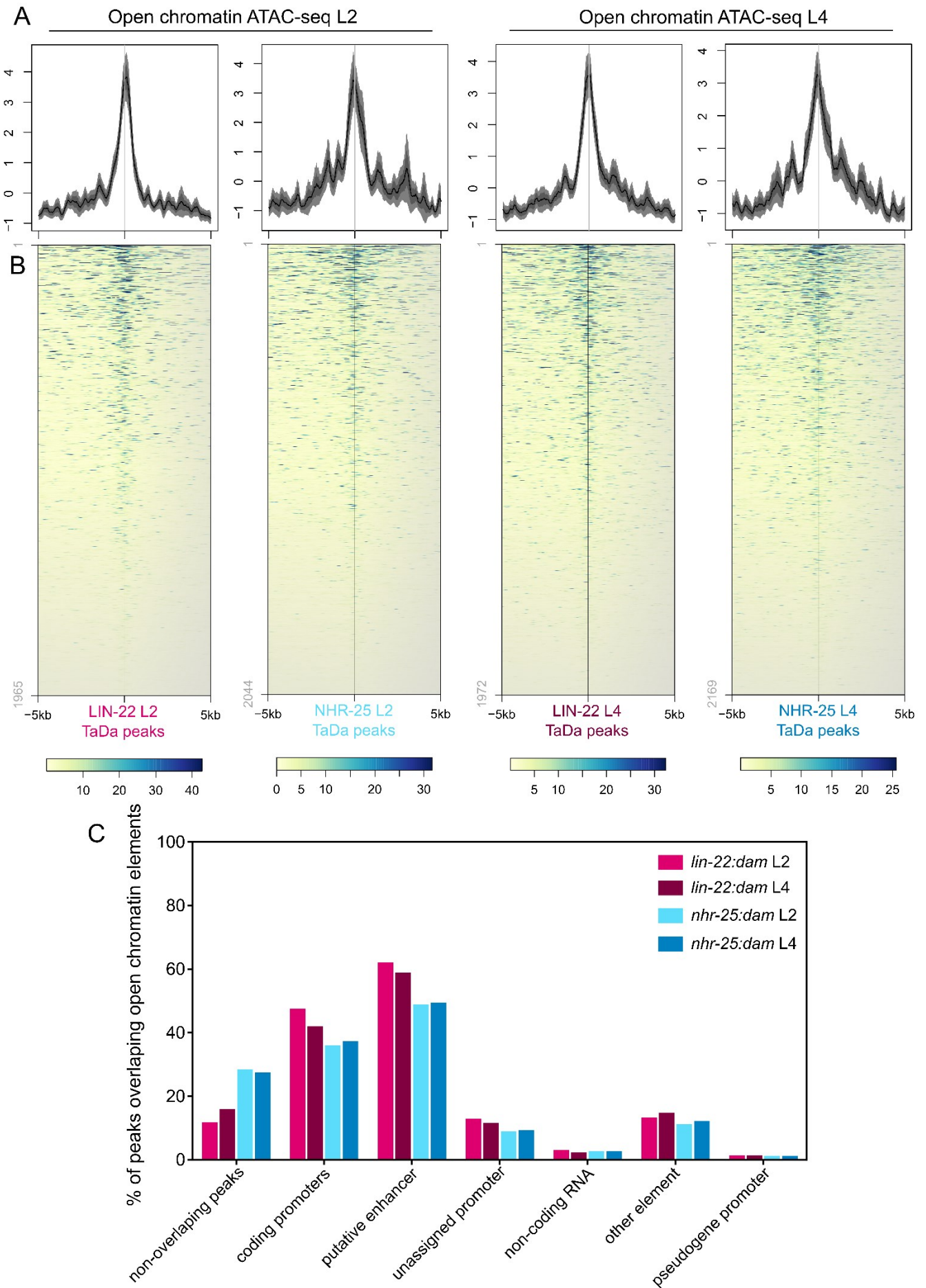
Figure 3. 5 The majority of LIN-22 and NHR-25 peaks overlap with open chromatin sequences. (A) Aggregation plots of open chromatin signal from ATAC-seq L2 and L4 data (Jänes *et al.*, 2018) over LIN-22 and NHR-25 TaDa L2 and L4 peaks respectively, indicate increased chromatin openness at the sites of peaks. **(B)** Heatmaps of open chromatin signal centred on TaDa TF peaks show that the majority of peaks overlap with open chromatin regions. The key is previously calculated ATAC-seq genome-wide normalised coverage scores (Jänes *et al.*, 2018). **(C)** Proportions of total peaks for each indicated TF at each stage overlapping with different categories of regulatory annotated open chromatin elements. A TF peak can overlap more than one element. In A and B ±5 kb around the peak centres have been plotted. In A, Y-axes are z-scores for the plotted sequence length and shaded areas represent 95% confidence intervals.

These open chromatin elements, identified in Jänes *et al.*, 2018, had been annotated for their regulatory type and categorised as: coding promoters, putative enhancers, unassigned promoters, non-coding RNAs, pseudogene promoters or other elements. Taking advantage of that information, the number of peaks, for each TF at each stage, that overlapped with each type of element was determined and is presented as proportions in Figure 3.5C. Due to the broad nature of TaDa peaks, some overlapped more than one accessible chromatin elements, which could be of different types and were thus assigned to multiple categories. Interestingly, the majority of peaks for all factors at all stages were overlapping either coding promoters (between 35.2% and 46.7%) or putative enhancers (between 48% and 61.2%); regions within which TF binding is expected to happen. This evidence supports the hypothesis that the peaks identified by TaDa very likely correspond to *bona fide* TF binding sites based on the combination of their location on the genome and relationship to previously identified regulatory sequences.

3.2.5 Comparison of peak localisation profiles between methods and transcription factors

Considering that this is the first application of TaDa in *C. elegans* I decided to assess the TF target identification capacity of the method in comparison to the more established ChIP-seq approach. The *nhr-25:dam* fusion was included in this study, both to explore its epidermal development role and due to the availability of ChIP-seq datasets that could be used for method validation. Two such datasets utilised here are from L1 and L2 staged animals from Shao *et al.*, 2013 and Araya *et al.*, 2014 respectively and for the rest of the chapter are referred to as NHR-25 ChIP-seq L1 or L2. Initial qualitative assessment by comparison of the signal tracks showed good agreement between TaDa and ChIP-seq data. The locations of the TaDa signal enrichment around *nhr-25* and *idh-1*, presented in the examples of Figure 3.4B, are employed in this instance to represent the agreement between the profiles (Figure 3.6A left, middle). All samples exhibit statistically significant peaks for the areas of signal enrichment that is displayed in Figure 3.6A and the precise positions of those peaks overlap between samples.

It is interesting to note that the ChIP-seq signal peak summits (average size 400 bp), that usually represent the significant peak regions with greater resolution than the broader TaDa peaks (average size for *nhr-25:dam* L2 1115 bp), were largely observed to overlap with the GATC fragment (average size 368 bp) with the highest score within a region of TaDa signal enrichment. The signal over the promoter of the

seam cell fate regulator *egl-18* (Figure 3.6A right) demonstrates relatively higher enrichment levels in TaDa than ChIP-seq samples in comparison to the other examples. This could suggest that the tissue-specificity of TaDa improves identification of targets expected to be regulated in a tissue-specific manner.

On a genome-wide level, aggregate *nhr-25:dam* L2 and L4 peak signal over all ChIP-seq L1 and L2 peaks, exhibited strong preference for overlapping localisation in comparison to localisation in adjacent regions (Figure 3.6B), further supporting the similarity of peak profiles across methods. More specifically, both the L2 and L4 *nhr-25:dam* peaks showed significant localisation overlaps across the genome, assessed by Monte Carlo simulations, with the 683 peaks of the NHR-25 ChIP-seq L1 (identified *de novo* from pile-up data) and the 5980 peaks of the ChIP-seq L2 datasets (Figure 3.6C). A much larger overlap was found with the L2 ChIP-seq dataset, with approximately 37% (726 peaks) and 38.5% (835 peaks) of the total L2 and L4 TaDa peaks overlapping respectively. As an added note, of the 726 *nhr-25:dam* L2 peaks overlapping the ChIP-seq L2 peaks, 543 had overlaps occurring with the highest enriched GATC fragment of the peak. This, in combination with the qualitative observation of ChIP-seq summits aligning with the most enriched GATC fragments of TaDa peaks, suggests that some crude positional information for the precise location of the TF binding within a peak could be acquired from TaDa signal. In contrast to ChIP-seq L2, only 7.3% and 7.2% of TaDa peaks from L2 and L4 were overlapping with ChIP-seq L1 peaks. The disparity between overlap sizes could be at least partly attributed to the earlier stage of the ChIP-seq L1 dataset, even though this is unlikely as the L4 TaDa dataset overlaps well with the ChIP-seq L2 dataset. A more likely explanation could be the lower quality of the raw data which is evident in the comparison of signal profiles between ChIP-seq samples (Figure 3.6A). Increased signal noise in the L1 dataset could potentially hinder peak-calling.

The reverse overlap of ChIP-seq L2 peaks with TaDa L2 peaks identified 971 overlapping peaks, which is just 16% of the ChIP-seq L2 dataset. Non-overlapping peaks between the two datasets could reflect differences stemming from the different expression domains within which identification of targets occurs. The native *nhr-25* domain is interrogated in ChIP-seq, while the *wrt-2* expression domain, including mostly the seam cells, is studied in TaDa. Therefore, it is conceivable that a number of ChIP-seq peaks reflect NHR-25 targets outside of the seam cells.

The overlapping peaks between the above samples were used to assess whether peak intensity as determined by TaDa or ChIP-seq scores correlated. Since peak intensity is generally accepted to reflect TF binding "strength", scores between common peaks were expected to somewhat correlate. Here, no correlation was found between TaDa and ChIP-seq peaks with a Pearson's correlation test (R^2=3.421e-5) (Figure 3.6D), similarly to what has been previously reported in comparisons for mammalian pluripotency factors (Cheetham *et al.*, 2018). In contrast, common peaks between L2 and L4 stage for NHR-25 TaDa showed correlation with R^2=0.5752. Since TaDa peaks are expected to reflect binding within a specified tissue, features like peak intensity could differ to that reported by ChIP-seq as it is averaged across the complete expression domain, providing a potential explanation for the discrepancy.

Next, the overlaps between the TaDa peak profiles identified here for the two TFs were examined. Interestingly, *lin-22:dam* L2 and L4 peaks showed an increased average preference to map onto both L2 and L4 NHR-25 TaDa peaks, in comparison to adjacent regions (Figure 3.6E top). This aggregate preference was mirrored by *nhr-25:dam* L2 and L4 signal over L2 and L4 TaDa LIN-22 peaks (Figure 3.6E bottom). LIN-22 and NHR-25 both participate in epidermal development and it could be reasonably expected that they share target genes. Therefore, this phenomenon could be readily attributed to genuine proximity or overlap of binding sites for LIN-22 and NHR-25. The specificity of TaDa could be amplifying this effect by restricting the pool of targets for both to those regulated in the *wrt-2* expression domain.

Another potential explanation could be if profiles shared a range of peaks that correspond to genomic sites where promiscuous binding of multiple transcription factors is observed. Such regions, referred to as High-Occupancy Target (HOT) regions, have been determined so far only from ChIP-seq experiments and have been reported to mostly represent biologically relevant sites of true binding, rather than ChIP-seq artefacts (Araya *et al.*, 2014). TaDa peaks from L2 and L4 showed significant (by Monte Carlo simulations) but small overlaps with whole-animal HOT regions from L2 and L4 animals respectively. 19% of *lin-22:dam* L2 (p=3.5e-116), 12% of *lin-22:dam* L4 (p=1.6e-80), 13% of *nhr-25:dam* L2 (p=1.3e-74) and 8% of *nhr-25:dam* L4 (p=1.4e-37) peaks were overlapping with HOT regions compared to 34% of NHR-25 ChIP-seq L2 peaks. Moreover, out of the 2167 HOT regions found in L2 (Araya *et al.*, 2014) only 292 (13%) were overlapping with TaDa NHR-25 L2 peaks, in stark contrast to the 1807 (83%) that were overlapping with ChIP-seq L2 peaks. These data suggest that TaDa peak profiles are somewhat less representative of HOT regions which could be down to method-dependent differences or the tissue-specificity that again restricts possible targets.

In order to further investigate the overlaps between the TaDa profiles for the two factors, all the pairwise intersections between them were performed and the number of overlapping peaks were measured. At the genome-wide level all the studied overlaps were found to be non-random by Monte Carlo simulations (Figure 3.6F). Around 60% of peaks of each profile overlapped across the two developmental stages for the same TF, whereas across TFs overlapping peaks were always <33% of the total (Figure 3.6F left). Overlaps and their significance were re-calculated by restricting sampling to promoter regions (defined here as 5 kb upstream to 500 bp downstream of TSS), to avoid artificially inflating p-values by interrogating the entire genome even though TF binding sites are expected to localise on promoters. Again, for these calculations all overlaps were still found to be highly significant (Figure 3.6F right) indicating that even within promoters, NHR-25 and LIN-22 TaDa peaks showed propensity to overlap. Notably, the majority of the measured genome-wide overlaps (approximately between 67% and 72%) for all pairwise intersections were found to be occurring within the promoter sequences (Figure 3.6E right) signifying that these are more likely regulatory in character rather than overlaps due to TaDa artefacts. In addition, less than a third of the across-TF overlaps in promoters occurred in HOT regions (32.5% of NHR-25 L2 and LIN-22 L2, 31% of NHR-25 L2 and LIN-22 L4, 25.5% of NHR-25 L4 and LIN-22 L2, 24.3% of NHR-25 L4 and LIN-22 L4), indicating that the vast majority of overlaps are more likely to be in LIN-22 and NHR-25-
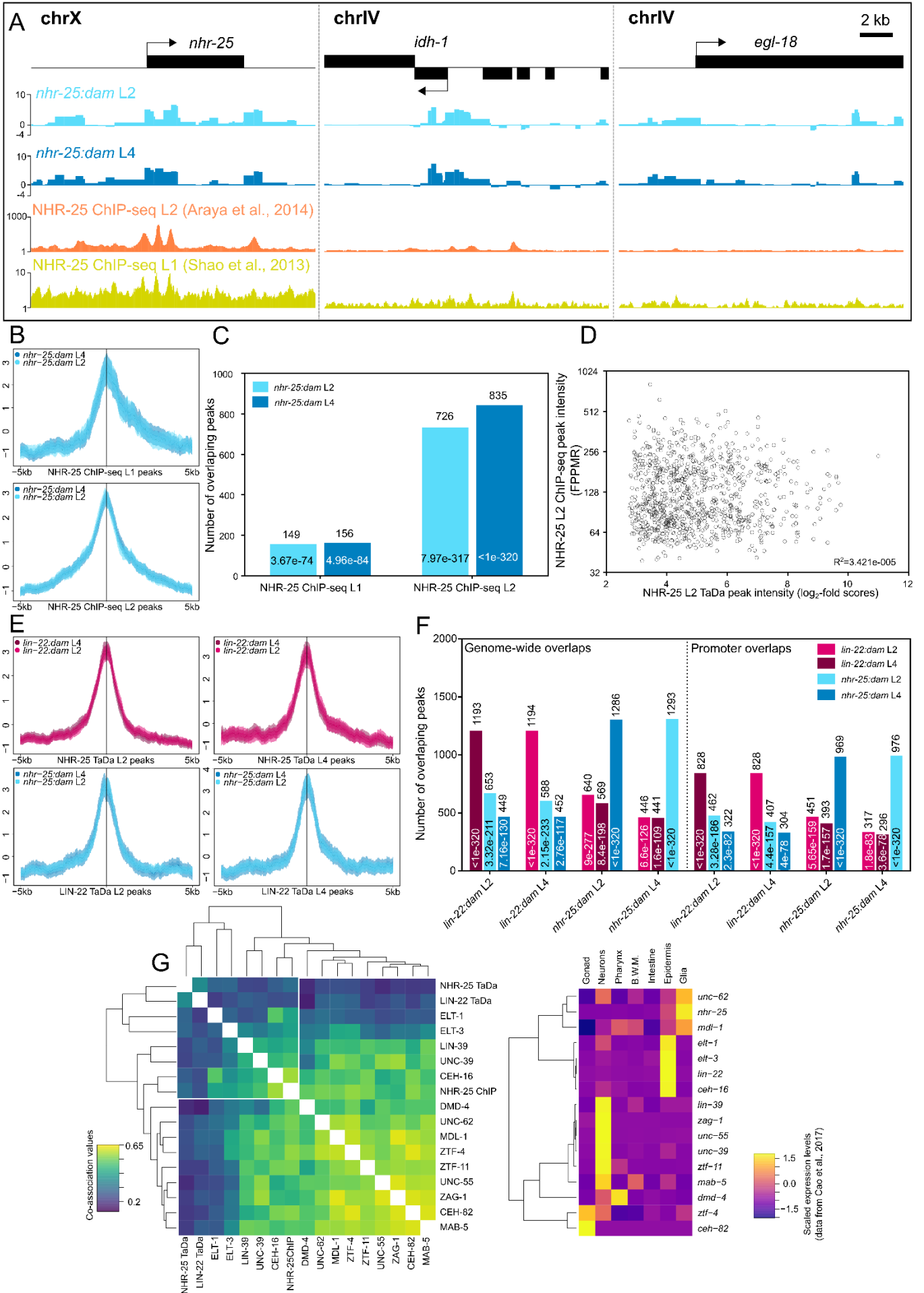
Figure 3. 6 Comparisons of peak localisation profiles show agreement with ChIP-seq data and overlaps between the two transcription factors. (A) Representative snapshots showing good agreement between *nhr-25:dam* TaDa data and two ChIP-seq signal profiles from L1 (Shao *et al.*, 2013) and L2 (Araya *et al.*, 2014) staged animals over *nhr-25*, *idh-1* and *egl-18*. Note the relatively higher signal enrichment in TaDa samples over the promoter of the seam cell fate regulator *egl-18*. For all samples signal shown here forms statistically significant peaks that overlap. The Y-axes for TaDa samples represent normalised $\log_2(nhr\text{-}25{:}dam/NLS\text{-}GFP{:}dam)$ scores while for ChIP-seq the fragment pile-up per million reads score. Scale bar is 2 kb as indicated **(B)** Aggregation plots showing enrichment of *nhr-25:dam* L2 and L4 signal over centre of NHR-25 ChIP-seq L1 (top) and L2 (bottom) peaks. **(C)** Graph indicating the number of peaks from *nhr-25:dam* L2 and L4 samples that overlap with the L1 or with the L2 NHR-25 ChIP-seq peaks datasets. **(D)** Scatterplot of peak intensity for peaks common between *nhr-25:dam* L2 TaDa and NHR-25 L2 ChIP-seq datasets. TaDa peak intensity is normalised $\log_2(nhr\text{-}25{:}dam/NLS\text{-}GFP{:}dam)$ scores and ChIP-seq is fragment pile-up per million reads (FPPMR). The R^2 value is indicated. **(E)** Aggregation plots of *lin-22:dam* L2 and L4 signal over NHR-25 L2 and L4 TaDa peaks (top) and *nhr-25:dam* L2 and L4 signal over LIN-22 L2 and L4 TaDa peaks. All show enrichments for the peak location. **(F)** Numbers of peaks from each sample, indicated by colour, that overlap the peaks of the sample indicated on the X-axis. The graph is separated in two parts representing numbers of overlaps and *p*-values for their statistical significance across the whole genome or restricted to promoter regions 5 kb upstream to 500 bp downstream of the TSS. **(G)** Heatmap of symmetrized co-association values (left), hierarchically clustered, for all pairwise comparisons between peak localisation patterns of the indicated TFs from L2 animals. All peak profiles for the tested TFs (except of the TaDa produced here) derive from ChIP-seq experiments (Kudron *et al.*, 2018). Note the separate clustering of epidermal and neuronal factors. (Right) Per gene scaled heatmap of tissue specific expression levels (from(Cao *et al.*, 2017)) in the gonad, neurons, pharynx, body wall, muscle, intestine, epidermis and glia for the TFs tested for peak pattern co-association. In B and E, ±5 kb around the peak centres have been plotted and the Y-axes represent z-scores for the plotted sequence length and shaded areas represent 95% confidence intervals. In C and F, the exact number of peaks with overlaps is written above the bars and *p*-values from Monte Carlo simulations for the statistical significance of overlaps is printed inside the bar.

specific binding regions. Overall, these results strongly support that the observed overlaps are due to genuine binding of the same promoters, likely of genes that both factors regulate because of their shared ontology.

As a means to test this hypothesis, peak profiles for various TFs from ChIP-seq experiments conducted at L2 were acquired from the modERN database (Kudron *et al.*, 2018) to perform comparisons between the peak localisation patterns they exhibit. Such profile-wide comparisons of peak localisation between factors, based on overlap and proximity statistics, provide a measure of co-association that can statistically rank factors based on the similarity of their binding (Chikina & Troyanskaya, 2012). Previous calculations of co-association matrices have shown that TFs clustering together and separate of others, due to similarity of their binding patterns, often regulate the same targets, belong to the same ontology or act in a specific tissue (Araya *et al.*, 2014; Kudron *et al.*, 2018). Here, to assess the nature of the overlaps between LIN-22 and NHR-25 TaDa peaks in promoters, a co-association matrix was calculated.

To validate that the observed promoter binding by both factors is not due to promiscuous promoter methylation but because of genuine co-localisation of binding sites, peak profiles from factors expected to have different or similar regulatory targets were utilised. More specifically, factors were selected from the modERN database, based on their gene ontology. The profiles of the epidermis regulators ELT-1, ELT-3, CEH-16 and the NHR-25 L2, along with profiles for factors that relate exclusively or to a large extent to

neurogenesis or neural function, such as LIN-39, UNC-39, DMD-4, UNC-62, MDL-1, ZTF-4, ZTF-11, UNC-55, ZAG-1, CEH-82, MAB-5 were selected.

Encouragingly, the epidermal factors along with the NHR-25 and LIN-22 TaDa clustered separately from the majority of neuronal factors (Figure 3.6G). Pairwise co-association values were generally higher between neuronal factors which most likely reflects higher target sharing compared to epidermal factors that even amongst ChIP-seq data exhibit lower values, suggesting that lower co-association values seen for TaDa are not due to loss of targets. However, an interesting observation is that the ChIP-seq and TaDa profiles for NHR-25 do not cluster as closely as expected. This is likely to be biological since in TaDa the sampling pool is expected to be restricted to those targets regulated within the epidermis (*wrt-2* expression domain) excluding those that NHR-25 might have in glia, where it is also expressed based on published single-cell RNA-seq data (Figure 3.6G right)(Cao *et al.*, 2017). These targets in the neuron related glial tissue might increase the co-association to neural factors resulting in the observed positioning of the NHR-25 ChIP-seq profile in the matrix.

To further examine the contribution of the tissue of expression on the co-association between peak profiles for TFs, tissue-specific expression levels for the examined factors from Cao *et al*, 2017 were juxtaposed to the co-association matrix. Expression levels and co-association values somewhat correlate since factors expressed in the same tissue can conceivably be regulating an array of the same targets. These results advocate that the observed overlaps found between LIN-22 and NHR-25 peaks in TaDa are most likely due to target regulation by both factors because of their shared role in controlling aspects of epidermal development.

3.2.6 LIN-22 and NHR-25 DNA-binding motif identification by TaDa

Since TaDa peaks likely represent genomic regions of genuine DNA binding sites, identification of the DNA binding motif for each of the factors was attempted next. To this end, selecting an appropriate set of peaks to be used for motif identification is crucial. The challenge here is that TF DNA binding motif identification has so far most commonly been performed using the narrow summits of ChIP-seq peaks that allow for greater resolution than the broad TaDa peaks. Due to their smaller size a recurring candidate motif is more likely to be found as it is not diluted by other surrounding sequences. For NHR-25, DNA binding sequences and motifs have been previously reported by ChIP-seq and functional studies (Barkoulas *et al.*, 2016; Araya *et al.*, 2014; Shao *et al.*, 2013) and follow the consensus sequence 5'-TGACCTTG-3'. As a side note, the motif identified in Shao *et al*., 2013 does not resemble the expected supporting the hypothesis that profiles from that study are of somewhat suboptimal quality. In addition, being the homologue of the broadly studied across systems FTZ-F1 factor, motifs are known for multiple of its orthologues. Therefore, identification of an NHR-25 motif that matches the reported using TaDa peaks was attempted at first in an effort to determine the minimum set of peak filtering criteria to enrich the interrogated group of sequences for the motif and allow for identification.

Peak intensity and size are often used for peak filtering (Araya *et al.*, 2014) and previous DamID studies have also used conservation as well (Southall *et al.*, 2014). Here, peak intensity reflects TaDa enrichment scores, size is restricted either by filtering the length of peaks (usually between 50-500 bp) or by isolating GATC fragments of highest enrichment and conservation is assessed based on PhastCons7way scores (Spieth, Hillier & Wilson, 2005).The effect of these different criteria on the identified NHR-25 motif using two algorithms is presented in detail in the Appendix C.4. Selection of the peaks showing the highest enrichment appears to be the criterion with the highest impact in reproducing the NHR-25 motif.

For the presented motifs in this study, sequences were restricted to those that overlap between L2 and L4 peaks for each factor to increase the probability that they reflect real binding sites. The top 200 of those with the highest averaged enrichment score were used. The TaDa *de novo* identified motif for NHR-25 agrees very well with the reported (Figure 3.7A left) and when run against a database of known motifs, showed significant similarity amongst others (Appendix B.2) to those of the *nhr-25* human orthologue NR5A1 (*p*=3.04e-6) and mouse orthologue *Nr5a2* (*p*=2.35e-5) (Figure 3.7B left).

LIN-22 as a bHLH *Hes*-related factor was expected to show preference for an N-box (5'-CACNAG-3') or class C E-box (5'- CACGAC-3' or 5'- CACGCG-3') binding sequence (Ohsako *et al.*, 1994; Takebayashi *et al.*, 1994). The identified motif matched an E-box sequence (5'-CANNTG-3') (Figure 3.7A right) and is very likely to constitute the amalgamation of different E-box and N-box sequences. *Hes*-related factors have been shown to act on a variety of targets either by direct binding or by antagonising other bHLH factors for binding sites or dimerising partners (Sasai *et al.*, 1992; Kageyama & Ohtsuka, 1999), resulting in binding to different sequences, likely explaining the "noisy" make-up of the identified motif. The identified motif is also similar to reported motifs for the human orthologue HES1 (Lichtenberg *et al.*, 2018).

Comparison to known motifs showed significant similarity amongst others (Appendix B.3) to that of the human HEY1 (Hes-related Family bHLH transcription factor with YRPW Motif 1) factor (*p*=3.29e-3), that shows moderate homology to *lin-22* (DIOPT weighted score 1.9) and the *C. elegans* HLH-1 (*p*=2.45e-4), homologue of the mammalian *MyoD* which *Hes1* is known to antagonise for sites and binding partners (Sasai *et al.*, 1992) (Figure 3.7B right).

Using the positional weight matrices represented in the logos for the two identified motifs the genome was scanned for instances of motif occurrence. To assess if TF binding as determined in TaDa showed preference for the identified motifs, aggregate signal was mapped onto motif sites. For the NHR-25 motif the *nhr-25:dam* L2 and L4 signal showed clearly increased average preference for regions that include the motifs as opposed to neighbouring sequences (Figure 3.7C). In contrast, the *lin-22:dam* L2 and L4 signal did not show any enrichment preference in relationship to the NHR-25 motif. The reverse relationship was observed when the signal profiles were mapped onto the LIN-22 motif. *lin-22:dam* L2 and L4 aggregate profiles exhibited increased average signal at the site of the motif whereas no such preference was particularly evident for the *nhr-25:dam* L2 and L4 signal (Figure 3.7D). Some limited increase is observed
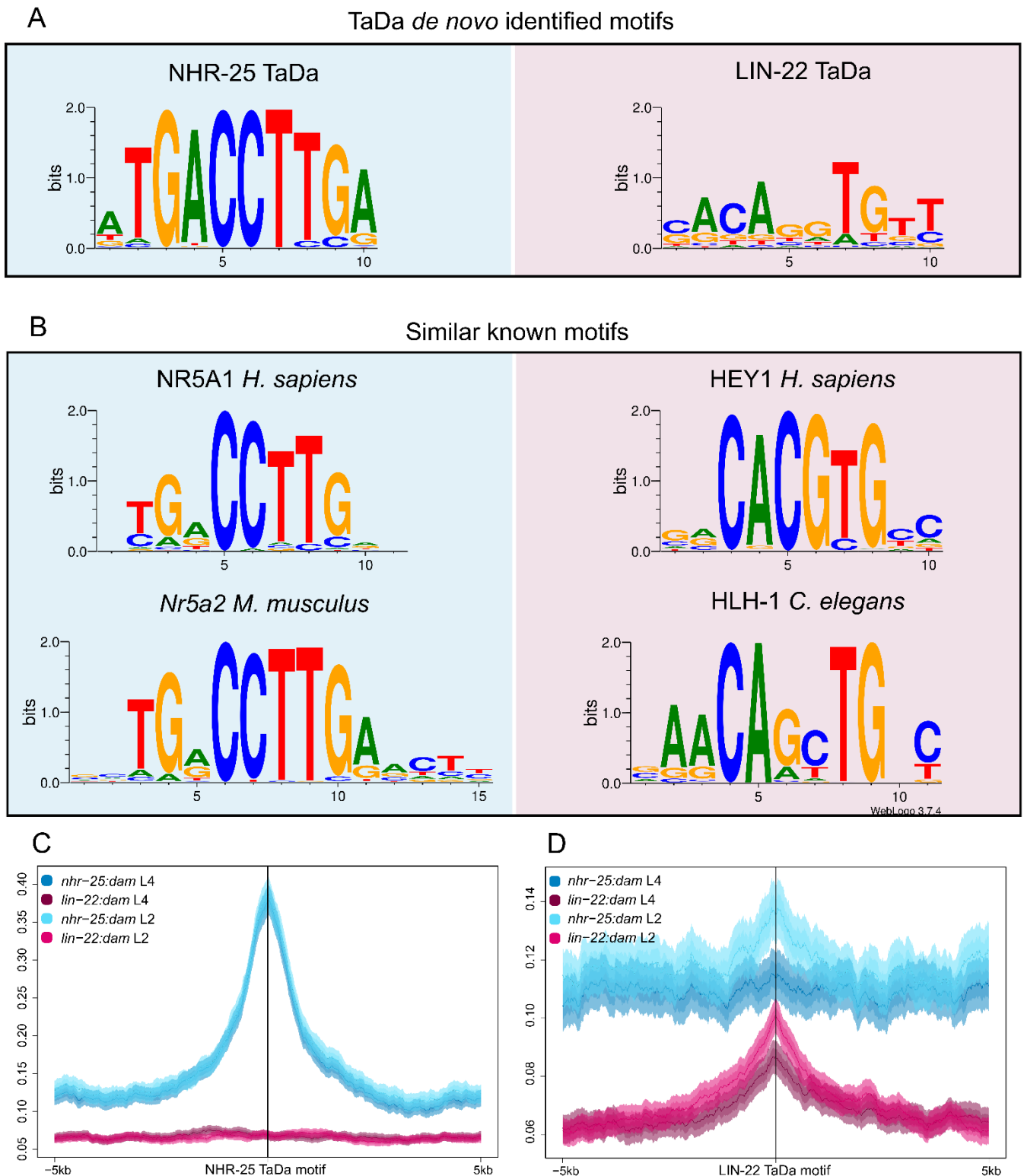
Figure 3. 7 NHR-25 and LIN-22 DNA-binding motifs identified by TaDa peaks. (A) *De novo* identified DNA-binding motifs for NHR-25 (left) and LIN-22 (right) from the top 200TaDa peaks for each TF respectively. **(B)** Significantly similar known motifs to the TaDa-identified NHR-25 (left) and LIN-22 (right) motifs from available databases. **(C-D)** Aggregation plots of TaDa signal for all TFs and available stages over instances of the NHR-25 TaDa motif (C) and the LIN-22 TaDa motif (D). Both TFs show strong preference for their respective motif compared to the alternative. ±5 kb around the peak centres have been plotted, the Y-axes represent mean enrichment scores for the plotted sequences and shaded areas represent 95% confidence intervals.

for *nhr-25:dam* L2 which could be attributed either to the described overlap of NHR-25 and LIN-22 TaDa peaks or to the multiple possible permutations of the composition of the "noisy" LIN-22 motif that make it more frequent on the genome (24077 instances compare to 16044 for the NHR-25). In terms of occurrence within significant peaks, the motifs for each respective factor were found in 41% of the LIN-22 L2, 33% of the LIN-22 L4, 47% of the NHR-25 L2 and 51% of the NHR-25 L4 total TaDa peaks. The co-occurrence was statistically significant for all by Monte Carlo simulations with *p*-values of 4.5e-53, 2.9e-19, 2.3e-264, <1e-320 for the above respectively.

3.2.7 TaDa-identified target genes relate to known functions of LIN-22 and NHR-25

As described in section 3.2.4, significant peaks were assigned to nearby genes that constituted putative LIN-22 or NHR-25 targets. To ensure that target genes were not being missed, instead of assigning peaks to a single gene based on proximity as the prioritisation factor, a less stringent approach was used where peaks were assigned to all genes that fulfilled the assignment criteria, that is the centre of the peak lying within 6 kb upstream of the TSS to 1 kb downstream of the TES of the gene. This assignment resulted in a set of 2809, putative targets of LIN-22 at L2 and 2833 genes at L4. Similarly, 3552 and 3724 genes were identified for NHR-25 at L2 and L4 respectively.

As expected by the extent to which peaks overlap, the majority of identified genes were shared between developmental stages for each factor. More specifically, 63.2% and 63.7% of genes for *lin-22:dam* L2 and L4 repsectively and 64.9% and 68% of genes for *nhr-25:dam* L2 and L4 respectively, were shared between the stages (Figure 3.8A, B). For both overlaps the intersections were highly statistically significant with *p*<1e-320 with a hypergeometric distribution test. Genes in the overlap are more likely to be *bona fide* targets since they have been independently reproduced by experiments in both stages.

To assess if the identified genes are likely to be true targets of LIN-22 or NHR-25, the gene sets were analysed for enrichment for genes belonging to gene ontology (GO) terms with relevance to previously known LIN-22 and NHR-25 biological roles. Tables with all the recovered GO terms are presented in Appendix B.4-B.10. Multiple GO terms related to LIN-22 functions or biological processes were found to be enriched both in L2 and L4 gene-sets. Subsets of target genes were participating in various developmental ontologies with terms like "cellular developmental process", "post-embryonic development" and "post-embryonic animal organ development" recovered both for L2 and L4 sets (Figure 3.8Ci, ii). *lin-22* is known to control aspects of epidermal development by instructing the correct establishment of division symmetry or asymmetry in the seam cells and allowing for division stochasticity in the P3.p cell of the ventral epidermis (Katsanos *et al.*, 2017). In addition, it participates in seam cell maintenance, promoting the hypodermal fate in dividing cells and anterior/posterior patterning by regulating Hox genes (Wrischnik & Kenyon, 1997), therefore such terms reflect this capacity of LIN-22. Furthermore, *lin-22* is also involved in various neurogenesis or neural morphogenetic events. It supresses PDE and PVD neuron formation in seam cells anterior to V5 and *lin-22* mutants have supernumerary mating rays that are part of the male mating organ (Katsanos *et al.*, 2017; Wrischnik & Kenyon, 1997; Yip

& Heiman, 2016). Consequently, LIN-22 is expected to mostly supress genes that drive neurogenesis and this is reflected in the GO terms with significantly enriched terms like "neurogenesis", "neuron development", "regulation of neuron differentiation" and "male anatomical structure morphogenesis" in both L2 and L4 sets.

All of the above GO terms were also enriched for the genes in the overlap between L2 and L4 stages, albeit with slightly lower enrichment, indicating that a lot of the genes regulating those ontologies are shared between datasets (Figure 3.8Ciii). Of the selected GO terms presented in Figure 3.8C none were recovered from the exclusively L4 genes but a few ("neurogenesis", "neuron development", "regulation of neuron differentiation", "cellular developmental process") were enriched in the L2 only genes (Appendix B.6). In addition, the L2 dataset was also found to be enriched for genes belonging to the Wnt-signalling pathway (KEGG and WikiPathways databases with adjusted p-values of 3.28e-5 and 3.25e-3 respectively) which was linked for the first time to a *Hes*-related factor in Katsanos *et al*., 2017, where *lin-22* was shown to antagonise Wnt signalling by suppressing the expression of the *frizzled* receptor *lin-17*. These data suggest that LIN-22 might be also regulating other components (e.g. from WikiPathways Wnt-signalling genes: *mom-1, unc-37, bar-1, lit-1, lin-17, mig-5, mom-5, gsk-3, mig-14, mab-5, pop-1*) having a more profound effect on the Wnt-signalling pathway in general.

Lastly, in Katsanos *et a*l., 2017 we had hypothesised based on lineaging data that *lin-22* is likely to control aspects of division timing probably by acting on the heterochronic pathway. In the identified gene-sets for LIN-22, both the *lin-14* and *lin-28* heterochronic genes were found amongst the putative targets. Furthermore, LIN-22 binding peaks were also found near the heterochronic micro-RNA genes *lin-4* and *let-7* (Appendix C.5). Overall, a large portion of the heterochronic pathway machinery were found to be putative targets of LIN-22 by TaDa. It is worth noting that *let-7* was found as a target only in the L4 dataset, coinciding with the *let-7* expression onset which is known to be at L4 (Slack & Ruvkun, 1997) and could point to an activating role by LIN-22.

Similarly, the NHR-25 sets of genes were encouragingly found to be enriched for multiple GO terms related to *nhr-25* functions (Figure 3.8D). Terms for "structural constituents of the cuticle" and "molting cycle" genes were found to be amongst the most significantly enriched, representing one of the main biological processes driven by NHR-25 which is regulation of larval molting and new cuticle formation (Hayes, Frand & Ruvkun, 2006; Chen, Eastburn & Han, 2004; Gissendanner & Sluder, 2000; Chisholm & Hsiao, 2012). Moreover, *nhr-25* has been reported to play an important role in epidermal development by regulating cell-cell junction and fusion in epidermal cells like the seam cells, regulating their size and shape required for re-establishment of cell contacts and the asymmetry and number of divisions they execute, all important for their normal postembryonic development (Chen, Eastburn & Han, 2004; Gissendanner & Sluder, 2000; Šilhánková, Jindra & Asahina, 2005; Hajduskova *et al.*, 2009). Such functions were reflected in the target gene lists by their enrichment for developmental GO term like "post-embryonic development" and "cellular developmental process" (Figure 3.8D). Neuronal related GO terms like "neurogenesis", "neuron development" and "regulation of neuron differentiation" were also significantly enriched and were
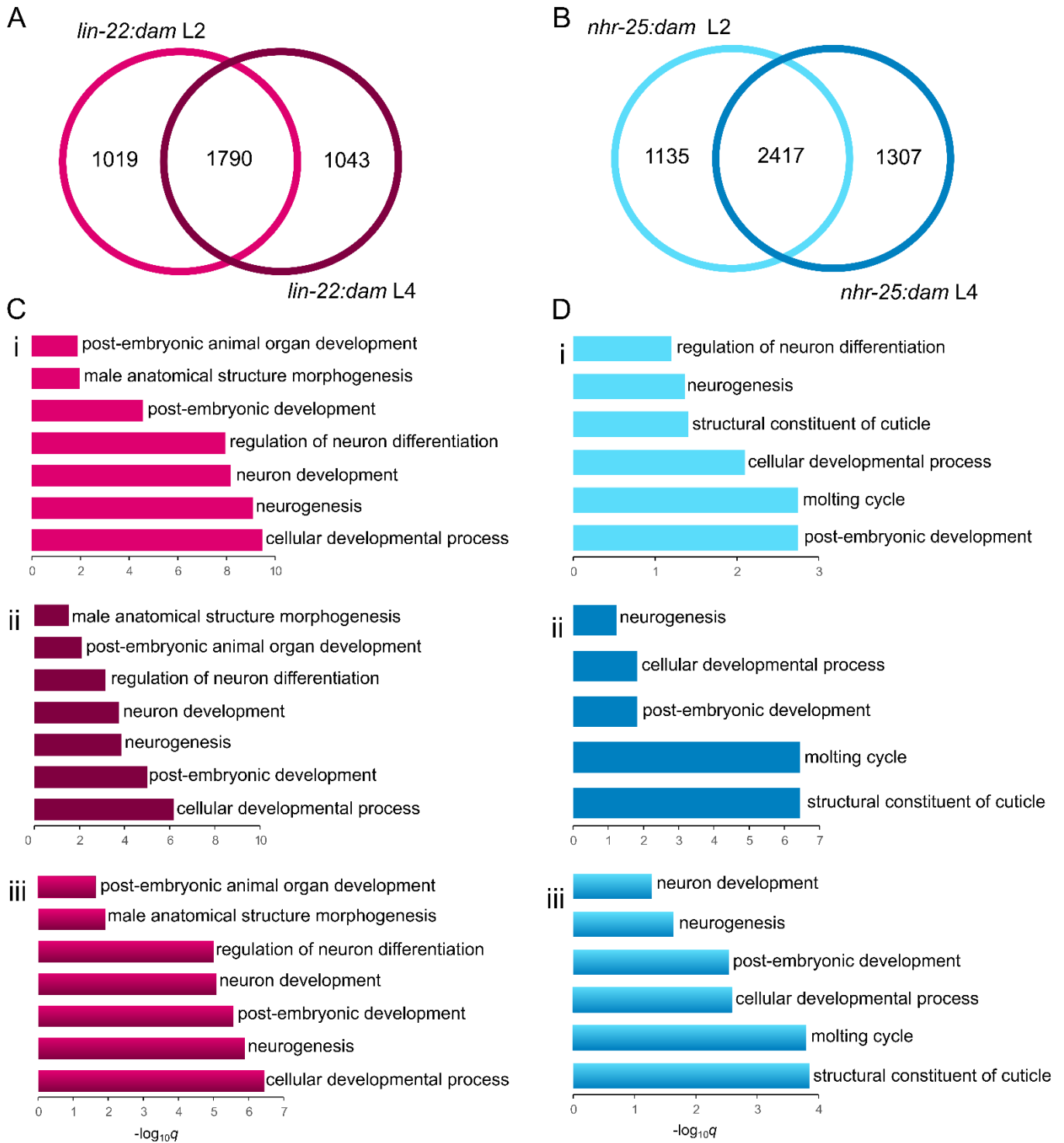
Figure 3. 8 Enriched GO terms for LIN-22 and NHR-25 putative targets relate to their known functions. **(A-B)** Venn diagrams for putative target gene-sets intersections between L2 and L4 for the *lin-22:dam* (A) and *nhr-25:dam* (B) identified genes (significance of intersection size with a hypergeometric distribution test *p*<1e-320 for both). **(C-D)** Plots of selected significantly enriched terms from GO-term analysis for the genes in the L2 **(i)**, L4 **(ii)** or intersection datasets **(iii)**. For both TFs at all stages, relevant terms are recovered with neurogenesis and development related terms for *lin-22:dam* **(C)** and molting and development related terms for *nhr-25:dam* **(D)**.

found reproducibly in both stages. These could, for example, correspond to subsets of target genes participating in the correct determination of the neural fate in the T seam cell descendants, as previously described (Hajduskova *et al.*, 2009).

Just like LIN-22, the above GO terms were found to be enriched both in L2 and L4 (Figure 8D i, ii) datasets while most of them were also recovered for the set of genes in the overlap between the two stages (Figure 3.8D iii). However, in contrast to LIN-22 the exclusively L2 genes were not further enriched for these GO terms whereas the exclusively L4 were still significantly enriched for "molting cycle" and "structural constituent of the cuticle genes" (Appendix B.9, B.10). Lastly, amongst the putative miRNA targets of NHR-25, the seam cell expressed miRNA cluster *mir-42*, *mir-43* and *mir-44* and the hypodermally expressed *mir-47* (Martinez *et al.*, 2008) were also found (Appendix C.5).

3.2.8 Target genes show extensive overlaps with existing datasets and reveal novel developmental links

To further investigate the make-up of the identified gene lists and evaluate their character as true targets of LIN-22 and NHR-25, the overlaps they show with existing datasets were assessed. The LIN-22 TaDa target genes for L2 and L4 were intersected with a list of *in silico* predicted genetic interactors of *lin-22* (Zhong & Sternberg, 2006)*.* The list included 52 genes that were predicted to be downstream of *lin-22* and it was reasoned that as for other previously shown genetic interactions (e.g. *mab-5*), it was likely that some could be proven to be direct targets by TaDa. Indeed, 22 genes for L2 and 15 for L4 (significant overlaps for both with a Fisher's exact test *p*=5.21e-7 and *p*=4.15e-3 respectively) were identified by LIN-22 TaDa as potential direct targets (Figure 3.9A). These included *lin-32* and *mab-5,* shown earlier in the chapter to be putative targets but also other genes known to participate in seam cell development like *unc-62* (Hughes *et al.*, 2013) and *nmy-2* (Ding & Woollard, 2017)*,* which had no previous links to *lin-22* as their regulator. An important putative target of LIN-22, with a pivotal role in seam cell fate that was identified in the intersections was *rnt-1*. *rnt-1* is the Runx homologue of *C. elegans* which plays a central role in seam cell fate determination by overriding the Wnt signalling asymmetry to promote symmetric division in L2 (van der Horst *et al.*, 2019). Regulation of *rnt-1* by LIN-22 could explain why *lin-22* mutants have a propensity to show ectopic symmetric seam cell divisions.

In the absence of *C. elegans* ChIP-seq datasets for LIN-22 and in an effort to more broadly validate the identified TaDa target genes as genuine LIN-22 targets, a ChIP-seq dataset for HES1, the human homologue of *lin-22*, from the Encode project (Accession: ENCSR109ODF) was utilised, as well as a list of selected targets of *Hes1*, the mouse homologue of *lin-22*, from a ChIP-chip experiment (Kobayashi *et al.*, 2009). Since the ChIP-seq data had the form of significant peaks across the human genome, they were assigned to protein coding genes with the approach used for *C. elegans* here and the resulting gene-list was converted (by DIOPT (Hu *et al.*, 2011)) to a list of the *C. elegans* putative orthologues, which was used here as the set for comparisons. The ChIP-chip shortlist was also converted the same way. The final

ChIP-seq list contained 3224 genes, some of which represented the same human homologue with different ranks/percentage of homology. All were kept to ensure that targets were not being missed.

Intersections with TaDa gene-sets showed considerable overlaps for both stages (approximately 23.4% of genes for both L2 and L4 datasets) that were statistically significant with a Fisher's exact test ($p=$ 1.88e-28 for L2 and $p=$ 2.58e-29 for L4) (Figure 3.9A). In addition, GO term analyses for the gene-sets of the overlaps found enrichment for multiple of the same GO terms identified just by the TaDa gene-sets, indicating that those targets are to some extent represented in the human dataset as well and likely constitute genuine conserved regulatory targets of LIN-22/HES1 across the two species (Figure 3.9A right). The mouse *Hes1* dataset was a curated list that did not include all the identified targets, however in the overlap with the TaDa datasets the known target of *Hes1* factors *lin-32* (*achaete-scute* homologue) was found, as well as the Wnt signalling components *lit-1* and *pop-1,* showing that the LIN-22 regulation of the Wnt signalling might be conserved. Moreover, a few cell cycle genes were found in the overlap like the cyclins *cyd-1* and *cya-1,* while the homologue of the cell cycle inhibitor *cki-1* had been shown in the same research that produced the list, as well as in others, to be supressed by Hes1 (Kobayashi *et al.*, 2009; Murata *et al.*, 2005), suggesting a role for *lin-22* in cell cycle regulation. *cki-1* has also been shown to be downregulated in *lin-22* mutants (Katsanos *et al.*, 2017) and is found here to be a putative direct target of LIN-22.

For assessment of the NHR-25 TaDa-identified target genes, the available ChIP-seq datasets presented earlier (section 3.2.5) were utilised. In the NHR-25 ChIP-seq L1 study (Shao *et al.*, 2013) a list of target genes was available, while from the ChIP-seq L2 study (Araya *et al.*, 2014) a profile of significant peaks was available that were assigned to genes using the same approach as for the TaDa data. Highly significant overlaps were found for all pairwise and higher grade intersections between the datasets (Figure 3.9B) indicating that a plethora of TaDa-identified genes as targets of NHR-25 were also found by ChIP-seq and are likely true targets of NHR-25. More specifically, a somewhat limited but statistically significant overlap was found with the 1377 protein coding genes of the L1 dataset, with 11.2% of the L2 and 11.1% of the L4 TaDa-identified genes, present in the overlap (significant with Fisher's exact test $p=$ 4.88e-27 for L2, $p=$ 6.77e-28 for L4). In contrast, the overlaps with the ChIP-seq L2 dataset of 7438 genes were large, with 62% of L2 and 61.6% of the L4 TaDa genes present in the overlaps, almost as many as in the overlap between the TaDa datasets for the different stages (Figure 3.9B). Both overlaps were statistically significant with a Fisher's exact test, with $p=$ 3.76e-248 for the L2 and $p=$ 1.86e-255 for the L4 overlap.

As in the case of the number of overlapping peaks between the two methods, the overlap of genes represents <30.8% of the ChIP-seq L2. Besides ChIP-seq artefacts or targets missed by TaDa, the rest of the ChIP-seq L2 dataset could contain targets of NHR-25 outside the *wrt-2* expression domain that is examined in TaDa. Supporting that hypothesis, tissue enrichment analysis for genes in the overlap between ChIP-seq L2 and TaDa showed association with multiple epidermal tissues, whereas exclusively ChIP-seq L2 genes showed enrichment for neuron and reproduction related tissues, likely representing the rest of the *nhr-25* expression domain (Appendix B.11, B.12).
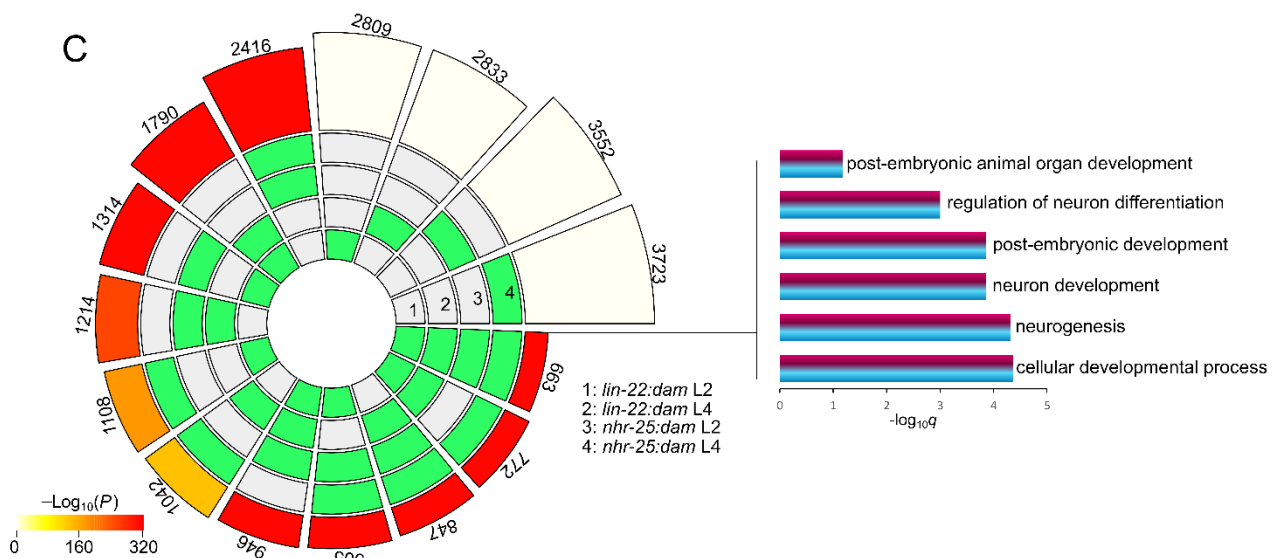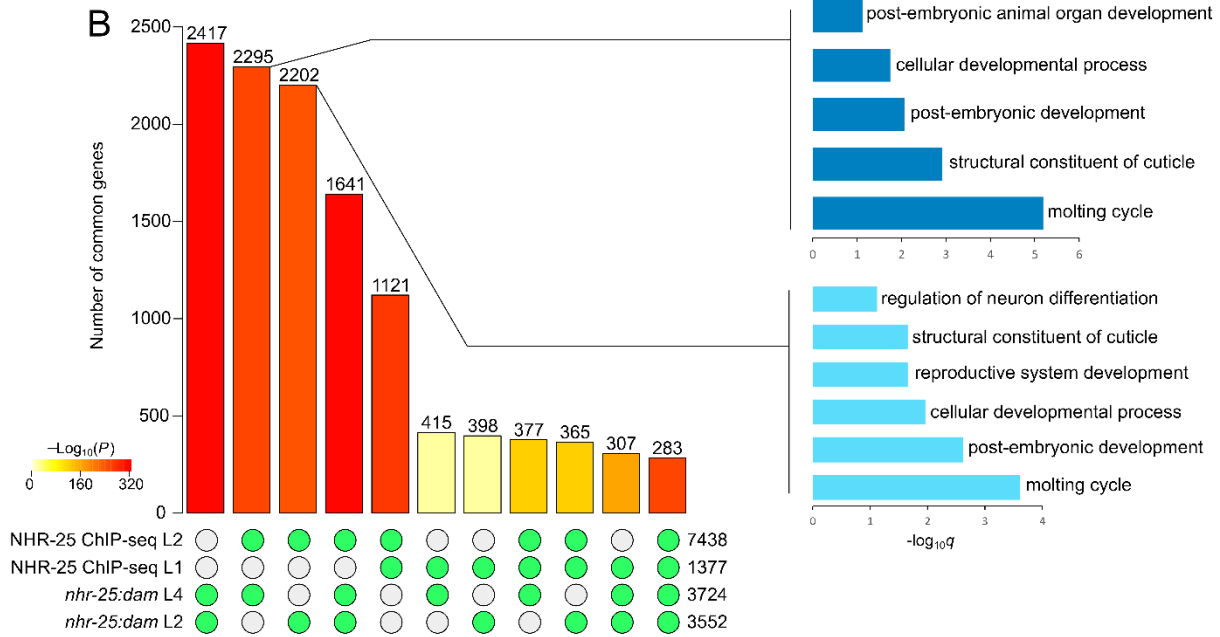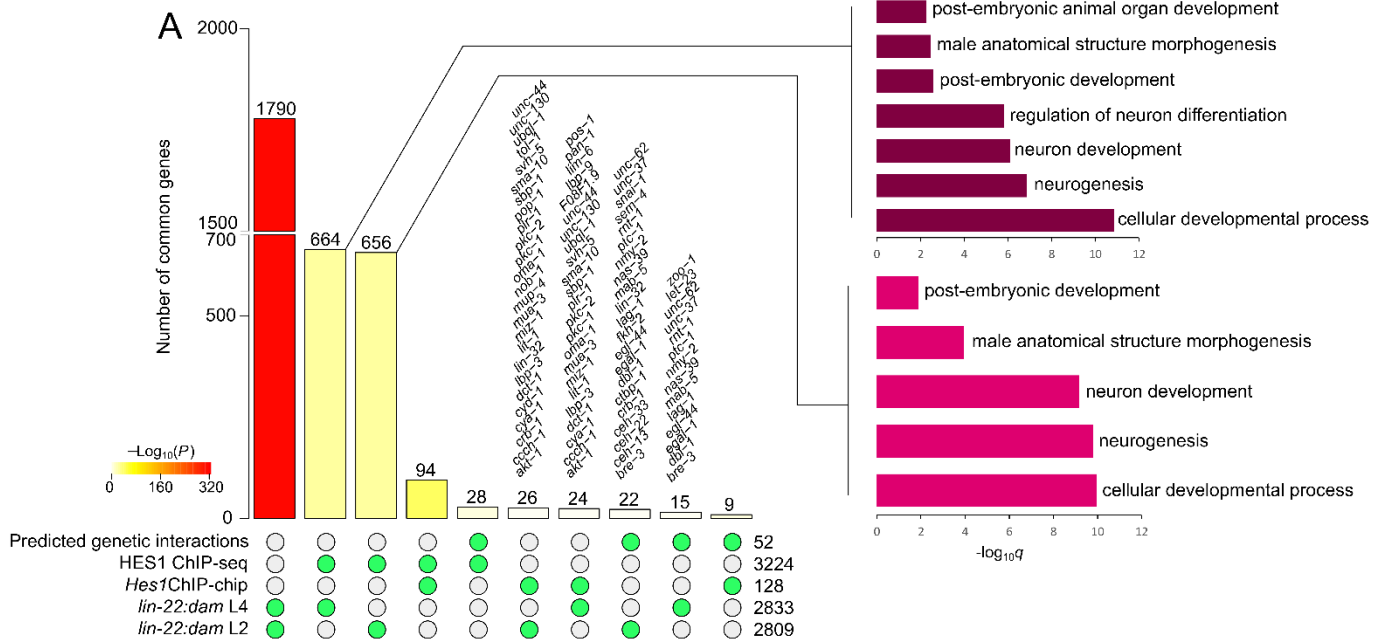
Figure 3. 9 TaDa identified sets of targets for LIN-22 and NHR-25 show significant overlaps between them and with available datasets. (A) Barplot of pairwise intersection sizes between: identified target genes for *lin-22:dam* L2 and L4, genes predicted to be downstream genetic interactors of *lin-22* (Zhong & Sternberg, 2006), *C .elegans* orthologues of HES1 (*H. sapiens*) ChIP-seq identified targets and *C. elegans* orthologues of a subset of *Hes1* (*M. musculus*) ChIP-chip identified target genes (Kobayashi *et al.*, 2009). Genes in the intersection between TaDa and ChIP-chip or predicted interactors are listed in full above the bars. Selected enriched GO-terms for gene-sets from the large intersections between orthologues of HES1 ChIP-seq targets and TaDa targets are shown. The lists of orthologues might contain more than one *C. elegans* homologue for each human or mouse gene of the original list **(B)** Barplot of all possible intersection sizes between: *nhr-25:dam* L2 and L4 identified target genes, NHR-25 ChIP-seq L1 identified target genes (Shao *et al.*, 2013) and NHR-25 ChIP-seq L2 (Araya *et al.*, 2014) peaks assigned to genes using the method used in this study. Selected enriched GO-terms for gene sets from the large intersections between TaDa and ChIP-seq L2 are shown. **(C)** Circular plot of all possible intersections between TaDa identified target genes for LIN-22 and NHR-25 at L2 and L4. Selected enriched GO-terms are shown for the genes common in all datasets. In A and B, the size of each individual gene-set is printed at the bottom right of each graph. In A, B and C the statistical significance of each intersection assessed with a Fisher's exact test is indicated by colour hue. Of all observed intersections shown, only the *lin-22:dam* L4 gene set intersection with the ChIP-chip subset gene set was not significant with a *p*-value of 0.082335 (>0.05).

Encouragingly, the genes in the overlaps with the ChIP-seq L2 datasets were enriched for GO terms identified previously and relating to *nhr-25* functions, like molting and cuticle structure related terms, as well as developmental and neurogenesis related terms (Figure 3.9B right). The fact that genes in the overlap maintain enrichment for relevant GO terms, suggest that large proportions of the TaDa target genes, also reproduced in ChIP-seq, are very likely to be genuine targets of NHR-25 in the epidermis. It also highlights that the peak overlaps found in section 3.2.5 most likely represent binding of true target genes that perhaps participate in known *nhr-25* ontologies. The small overlap observed with the ChIP-seq L1 target genes could be further evidence of the suboptimal quality of signal in that study as previously stated.

Considering that LIN-22 and NHR-25 had significant peak overlaps and the fact that their sets of identified target genes were enriched for shared GO terms, pointed towards shared target genes. To assess this, the LIN-22 and NHR-25 TaDa putative target gene-sets were intersected to identify overlaps. As expected substantial significant overlaps were found in all pairwise comparisons across factors with 37-39.4% of LIN-22 L2 and 42.9-46.4% of LIN-22 L4 genes overlapping NHR-25 datasets and 29.3-34% of NHR-25 L2 and 29.8-35.3% of NHR-25 L4 genes overlapping LIN-22 gene-sets (Figure 3.9C). The statistical significance for the pairwise overlaps across factors with a Fisher's exact test was $1e\text{-}320 \leq p \leq 3.90e\text{-}140$. Interestingly, the intersection between all the TaDa datasets which contained 663 shared genes was enriched for those common GO terms related to neurogenesis and development. This finding indicates that at least to some extent shared functions of the two factors, for example on epidermal development, could be executed by the same array of target genes, which are regulated by both to achieve the correct developmental outcome.

3.2.9 TaDa confirms that *mab-5* and *lin-17* are LIN-22 targets

Closer inspection of specific target genes showed that the posterior Hox gene *mab-5* was a candidate for direct regulation by LIN-22, with a significant peak in its upstream promoter (Figure 3.10A), as shown in section 3.2.4. *mab-5* expression is known to be inhibited in anterior to V5 seam cells by LIN-22 (Wrischnik & Kenyon, 1997). Quantitative smFISH data from Katsanos *et al.*, 2017 and also reproduced in this study (Figure 3.10B), showed that in WT animals *mab-5* is expressed in the posterior daughter of V5 during the symmetric division and in none of the anterior seam cells (H0-V4) where *lin-22* was shown to be expressed (Katsanos *et al.*, 2017). In the absence of *lin-22* the amount of signal within the V5 lineage did not change (Katsanos *et al.*, 2017) but its pattern of expression was replicated by more anterior seam cells (V1-V4), with the posterior daughter of the symmetric division showing *mab-5* expression (Figure 3.10B). These results in combination prove the direct suppression of *mab-5* expression in anterior to V5 seam cells by LIN-22, confirming it as a direct target identified by TaDa.

Similarly, the *frizzled* receptor gene *lin-17* was shown in Katsanos *et al.*, 2017 to be restricted to posterior to the vulva seam cell lineages by *lin-22*. TaDa identified significant enrichment for LIN-22 binding on the *lin-17* promoter (Figure 3.10C) indicating that is a likely direct target of LIN-22. smFISH for *lin-17* both in this study and reported in Katsanos *et al.*, 2017, showed near absence of *lin-17* expression in anterior seam cells (H0-V1) during the L3 division in WT animals (Figure 3.10D). In a putative *lin-22* null mutant, expression of *lin-17* was strikingly increased in anterior to the vulva seam cells (Katsanos *et al.*, 2017), overall indicating that LIN-22 acts on *lin-17* by directly supressing its expression in anterior body seam cells.

Examination of the LIN-22 signal enrichment over the *lin-17* promoter exhibits extensive enrichment across the promoter, with what looks like two sites of increased enrichment more evident in the L2 profile. Two GATC fragments, one at 598-856 bp and another one at 3456-3713 bp upstream of the *lin-17* ATG are locally the most enriched (Figure 3.10E top) potentially indicating, as proposed earlier, the specific location of LIN-22 binding sites. To further study the *lin-17* promoter, an rVista analysis for conservation (Loots & Ovcharenko, 2004) was performed by comparing the sequence of the promoter between *C. elegans* and the related Caenorhabditis species *C. briggsae* and *C. brenneri.* The analysis revealed two conserved regions of the *lin-17* promoter that could potentially harbour regulatory elements and were termed CRE1 and CRE2 (Figure 3.10E bottom). Surprisingly, these elements almost precisely overlapped with the two most enriched GATC fragments described above (Figure 3.10E), suggesting that these might contain the *cis*-regulatory elements that recruit LIN-22 to supress *lin-17*.

To test this hypothesis, transcriptional reporters were built with either CRE1 or CRE2 fused to a minimal core promoter from *pes-10,* driving the expression of histone bound GFP. Multi-copy transgenes in the form of extrachromosomal arrays were created for each element (*icbEx177* for CRE1 and *icbEx180* for CRE2) in WT (N2) and were crossed in a putative null *lin-22(icb49)* mutant. Both reporters showed capacity to drive expression in cells of the posterior to the vulva epidermis in WT animals (Figure 3.10F,

G), in the region were *lin-17* expression was reported to be observed by smFISH (Katsanos *et al.*, 2017). Expression was observed both in the seam cells and the hypodermis. Some low level expression in posterior intestinal cells is likely to be an artefact of the *unc-54* 3'UTR used in the transgenes.

In the absence of LIN-22, in *lin-22(icb49)* mutants, the expression by both reporters was very frequently notably expanded to anterior to the vulva seam cells and hypodermis (Figure 3.10F, G), reminiscent of the expansion of *lin-17* expression in the anterior seam cells as assessed by smFISH (Katsanos *et al.*, 2017). Populations of transgenic animals of WT and *lin-22(icb49)* background were quantified at the L4 stage for the proportion of animals showing expression of the two transgenes in anterior to the vulva epidermal cells (Figure 3.10H). For both conserved putative regulatory elements a significantly higher proportion of animals showed anterior expression in the *lin-22(icb49)* background compared to the WT ($p<0.001$ with a Fisher's exacts test). Specifically, the proportion almost doubled for CRE1 from 34.3% to 68%, while for CRE2 it increased from 66.7% in WT to the entire transgenic population in *lin-22(icb49)* (100%). Taken together, these data indicate that LIN-22 most likely acts through these discovered regulatory elements of the *lin-17* promoter to supress its expression, while at the same time demonstrate the value of TaDa signal in revealing more detailed features of the regulatory landscape.

3.2.10 LIN-22 activates *cki-1* and supresses *rnt-1* in V1-V4 seam cells

Based on the GO term analysis and the overlaps with existing datasets, a few putative targets with biological relevance for seam cell development were selected to perform further confirmation experiments. With regards to LIN-22 TaDa, these genes were the previously mentioned *cki-1* and *rnt-1,* that showed significant signal enrichment in their promoter regions and *lin-39* that had been previously linked to *lin-22* regulated developmental events (Wrischnik & Kenyon, 1997) and showed significant enrichment mostly in a downstream region.(Figure 3.11A).

CKI-1 is a cyclin-dependent kinase inhibitor of the Cip/Kip family that has been found to act on the cell cycle by instructing G1 arrest in multiple developing tissues (Hong, Roy & Ambros, 1998). CKI-1 driven quiescence is required for differentiation programs to progress and lack of *cki-1* in the vulval precursor cells results in extra proliferative cell divisions (Buck, Chiu & Saito, 2009; Hong, Roy & Ambros, 1998; Matus *et al.*, 2015). *cki-1* expression has also been shown to follow a developmentally regulated patterns of expression in the V lineage of the seam cells, with higher expression levels between molts (Hong, Roy & Ambros, 1998). In *lin-22* mutants, whole-animal RNA-seq experiments had shown a significant reduction of *cki-1* expression suggesting that an activating relationship might exist (Katsanos *et al.*, 2017). In mammals the homologue of *cki-1*, p27[Kip1] has been shown to be directly repressed by Hes1 suggesting an opposite regulatory relationship. Here, smFISH experiments at the late L1 stage detected *cki-1* expression in all seam cells in WT animals and *lin-22(icb49)* putative null mutants (Figure 3.11B) (Appendix C.9). Quantification of transcript in the seam cells revealed a significant reductions of expression levels in V1-V4 cells of *lin-22(icb49)* mutants (Figure 3.11C) in keeping with the whole-animal RNA-seq prediction.
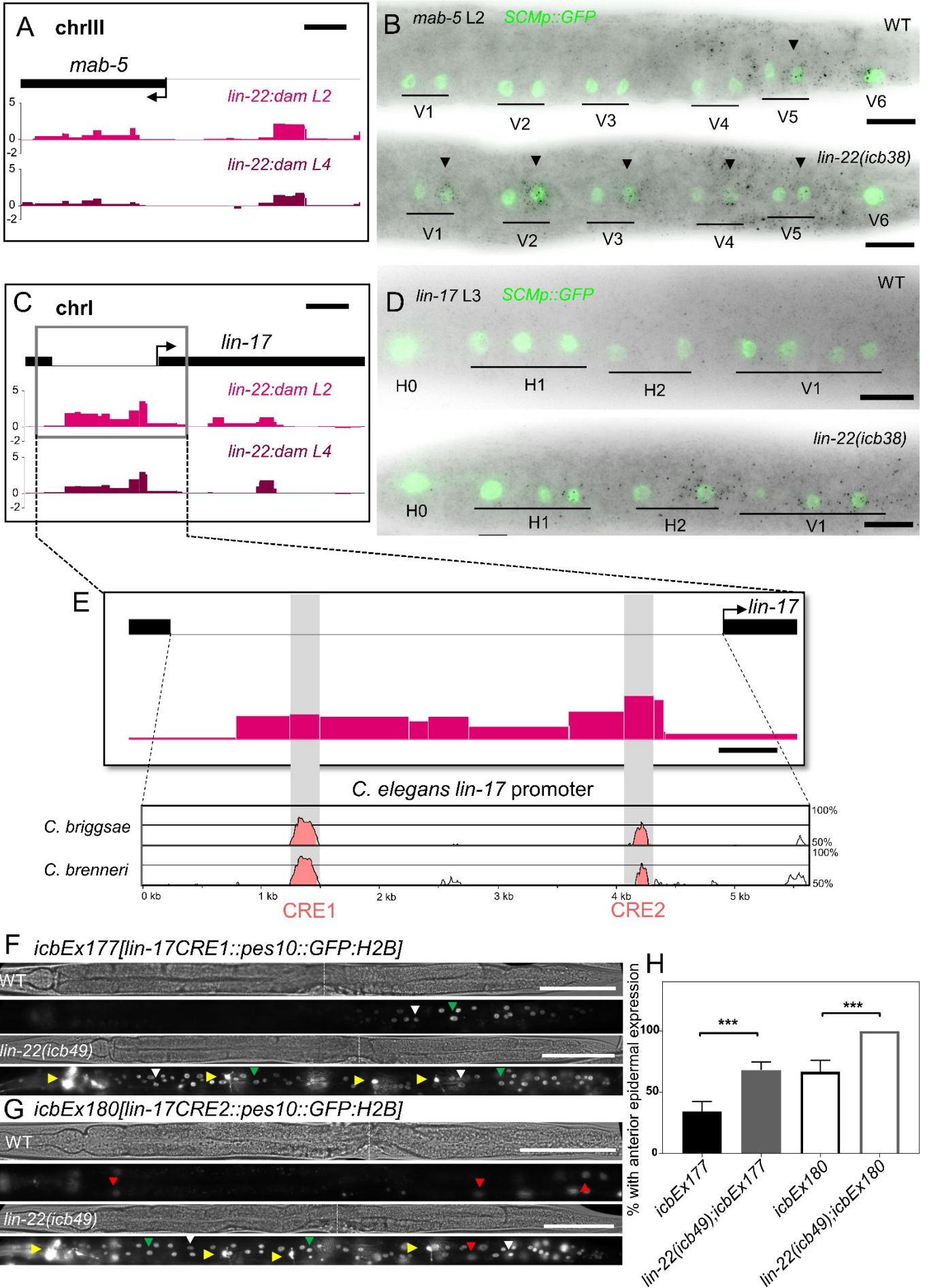
Figure 3. 10 Previously suspected LIN-22 targets are confirmed by TaDa. (A) *lin-22:dam* signal enrichment profile in regions with significant peaks in the promoter of *mab-5*. **(B)** Representative *mab-5* smFISH images from WT and *lin-22(icb38)* mutants show expansion of the WT V5-like expression of *mab-5* (indicated by arrowheads) in V1-V4 cells during the symmetric division of L2. **(C)** *lin-22:dam* signal enrichment profile in regions with significant peaks in the promoter of *lin-17*. **(D)** Representative *lin-17* smFISH images from WT and *lin-22(icb38)* mutants showing markedly increased *lin-17* expression in anterior seam cells (H1-V1) during the L3 division. **(E)** rVista analysis of the *lin-17* promoter identified two conserved, putative regulatory elements (*CRE1* and *CRE2*) between the *Caenorhabditis* species indicated and *C. elegans*. The position of the elements is shown on the *C. elegans* sequence. They are both overlapping GATC fragments with local maximum enrichment scores. **(F-G)** Representative brightfield and fluorescence images of L4 transgenic animals carrying transcriptional reporters for the *lin-17 CRE1* (*icbEx177* transgene) (F) and *CRE2* (*icbEx180* transgene) (G) sequences fused to *GFP:H2B* in WT and *lin-22(icb49)* mutant backgrounds. Expression is restricted to few posterior cells in WT and notably expands to the anterior epidermis in mutants. Green arrowheads indicate representative expression in the seam cells, white arrowheads in the hypodermis, red arrowheads in intestinal cells. White dashed lines indicate the position of the vulva and yellow arrowheads expression of the neuronal marker *dat-1p::GFP* linked to the *lin-22(icb49)* mutation. Anterior is to the left and the dorsal side is facing up. **(H)** Quantification of the proportion of transgenic animals with epidermal expression anterior to the vulva in WT (n=35 for *CRE1* and n=24 for *CRE2*) and *lin-22(icb49)* (n=47 for *CRE1* and n=20 for *CRE2*) mutant animals carrying the above reporters. Error bars indicate the standard error of the proportion. Black stars show statistically significant differences of the proportion with a Fisher's exact test, *** $p<0.001$. In B and D, seam cells are labelled with *SCMp:GFP* and black spots correspond to investigated mRNAs. Y-axes in A and C represent normalised $\log_2$(*lin-22:dam*/*NLS-GFP:dam*) scores. Scale bars are 2 kb in A and C, 10 μm in B and D, 500 bp in E, 100 μm in F and G.

These data suggest that LIN-22, even though commonly thought as a transcriptional suppressor, most likely acts to directly activate *cki-1* expression.

The Runx transcription factor homologue of *C. elegans rnt-1* was also identified by TaDa as a putative target. RNT-1 is known to be a major determinant of seam cell fate and along with its binding partner BRO-1 (homologue of CBFβ) is responsible for overriding the Wnt/β-catenin asymmetry signalling pathway (Wβa) cues and instructing the execution of symmetric proliferative seam cell divisions at the L2 stage (Nimmo, Antebi & Woollard, 2005; Kagoshima *et al.*, 2007; van der Horst *et al.*, 2019). Lineaging data for *lin-22* mutants had shown occurrence of ectopic symmetric divisions at later stages making *rnt-1* a promising candidate for direct regulation by LIN-22, in light of the TaDa identification.

smFISH experiments and quantifications for *rnt-1* at the late L1 stage detected low levels of expression in the seam cells of WT animals, in line with reporter data (van der Horst *et al.*, 2019) (Figure 3.11D, E). In the absence of LIN-22, in *lin-22(icb49)* mutants, the expression of *rnt-1* was found to be significantly increased in V1-V4 seam cells (Figure 3.11D, E), where symmetrisation of divisions had been found to occur most frequently (Katsanos *et al.*, 2017). Therefore, *rnt-1* is very likely to be a direct target of LIN-22, which represses its expression in V1-V4 seam cells in WT animals.

As presented in previous sections, *lin-22* was predicted and proven here to directly repress the Hox gene *mab-5* in anterior to V5 seam cells. Another Hox gene of the midbody that had been considered to be inhibited by LIN-22 is *lin-39. lin-39* is expressed in V2-V5 seam cells overlapping the *lin-22* expression domain of H0-V4. Moreover, genetic evidence had suggested a putative interaction because the multiple PDE phenotype of *lin-22* mutants is supressed in *lin-22; lin-39* double mutants (Wrischnik & Kenyon, 1997). A significant area of LIN-22 binding was found downstream of *lin-39,* potentially signifying expression
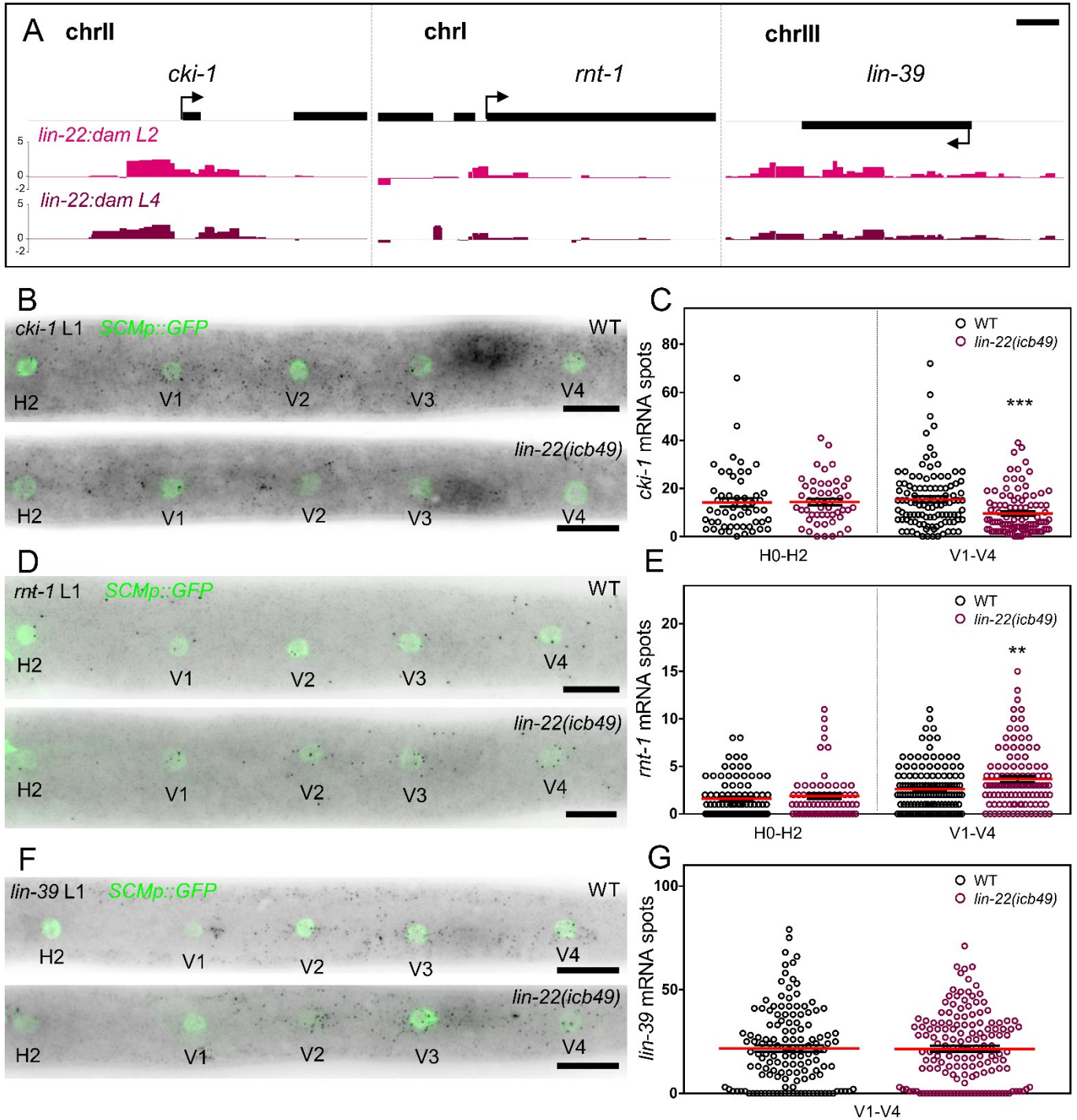
Figure 3. 11 Confirmation experiments of newly predicted LIN-22 targets: LIN-22 activates *cki-1* and represses *rnt-1.* (A) *lin-22:dam* signal enrichment profiles in regions with significant peaks around genes identified as putative LIN-22 targets. **(B-C)** Representative *cki-1* smFISH images from WT and *lin-22(icb49)* animals at the late L1 stage (B) and quantification of *cki-1* mRNA spots in H0-V4 seam cells (C), showing a significant reduction of expression in V1-V4 cells (50≤n≤126). **(D-E)** Representative *rnt-1* smFISH images from WT and *lin-22(icb49)* animals at the late L1 stage (D) and quantification of *rnt-1* mRNA spots in H0-V4 seam cells (E), showing a significant increase in expression in V1-V4 cells (65≤n≤167). **(F-G)** Representative *lin-39* smFISH images from WT and *lin-22(icb49)* animals at the late L1 stage (F) and quantification of *lin-39* mRNA spots in V1-V4 seam cells (G), showing no changes in expression in *lin-22(icb49)* mutants (154≤n≤166). In B, D, F the seam cells are labelled with the *SCMp:GFP* marker and black spots correspond to the respective mRNA. In C, E, G quantifications are presented in pools of H0-H1 and V1-V4 cells, the error bars indicate the SEM and black stars statistically significant differences to the WT mean with a *t*-test (** *p*<0.01, *** *p*<0.001). Scale bars are 2 kb in A and 10 μm in B, D, F.

regulation though a downstream enhancer. smFISH for *lin-39* confirmed the expression pattern in V2-V5 seam cells at the late L1 stage, both in WT and *lin-22(icb49)* mutant animals (Figure 3.11F) (per cell quantification in Appendix C.9). Quantification of transcript in the V1-V4 cells found no difference between WT and *lin-22(icb49)* mutants (Figure 3.11 G), indicating that LIN-22 is likely not a regulator of *lin-39* at least at the examined stage. Based on this evidence *lin-39* might be required for the ectopic PDE phenotype of *lin-22* mutants without directly interacting with *lin-22*.

3.2.11 NHR-25 supresses *egl-18* and *elt-1* in V1-V4 seam cells

Similar to LIN-22, TaDa-identified putative targets of NHR-25 with known roles in seam cells development, were selected for further validation. These genes were *egl-18*, *elt-1* and *rnt-1* all showing significant signal enrichment for NHR-25 binding in regions immediately upstream or overlapping their TSS (Figure 3.12A). All the above genes are crucial regulators of seam cell fate and no known postembryonic connections between them and *nhr-25* have been described, with *nhr* factors in general thought as downstream regulators in seam cell developmental networks (Koh & Rothman, 2001; Thompson *et al.*, 2016).

egl-18 is a GATA transcription factor, target of the activation of the Wβa signalling pathway that specifies seam cell fate during asymmetric seam cell divisions, usually in the posterior daughter cell (Gorrepati, Thompson & Eisenmann, 2013). Expansion of *egl-18* expression in daughters that should adopt the hypodermal differentiation program correlates with ectopic adoption of the seam cell fate (Gorrepati, Thompson & Eisenmann, 2013; Katsanos *et al.*, 2017).To test whether *egl-18* is a target of NHR-25, smFISH experiments were pursued. In these experiments, *nhr-25* RNAi treatment was used, which has been proven to be very effective and generates a more potent phenotype than the reference *ku217* allele. Experiments were performed at L3 to allow time for the treatment to take effect.

As previously reported, *egl-18* was expressed mostly in the posterior V daughters (Vn.papp and Vn.pppp) following asymmetric divisions of each dividing V1-V4 seam cell in control animals (Katsanos *et al.*, 2017) but was often observed to be expressed in anterior daughters as well in *nhr-25* RNAi treated animals (Figure 3.12B). This ectopic expression was more frequently observed in anterior V lineages (V1 or V2) (Appendix C.9). Quantification of smFISH spots showed an overall increase in *egl-18* expression in all daughters of V1-V4 at the L3 division (Figure 3.12C). Based on this evidence, NHR-25 appears to be a direct regulator of *egl-18* acting to repress its expression. Considering that *egl-18* has a seam cell specifying capacity this finding could be proposing a new mechanism of how reduction-of-function of *nhr-25* in animals leads to supernumerary seam cells mostly in the anterior body. Inability to form cell-cell junctions in the seam cells was previously considered to be the primary mechanism (Chen, Eastburn & Han, 2004; Šilhánková, Jindra & Asahina, 2005).

ELT-1 is another GATA transcription factor which is considered to be the master epidermal fate regulator in *C. elegans* and is known to regulate *nhr-25* in the embryo (Chisholm & Hsiao, 2012; Gilleard & Mcghee, 2001). Post-embryonically, *elt-1* is expressed in the seam cells and specifies seam cells fate
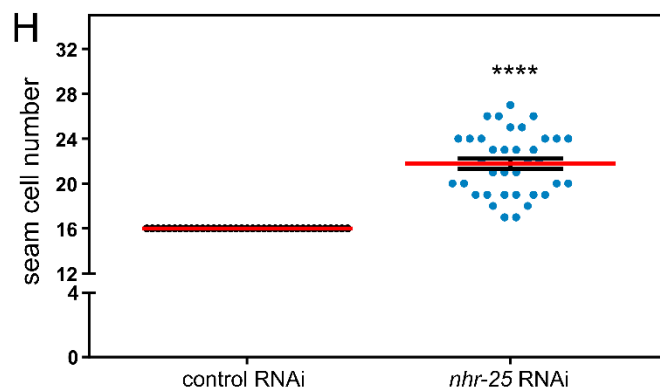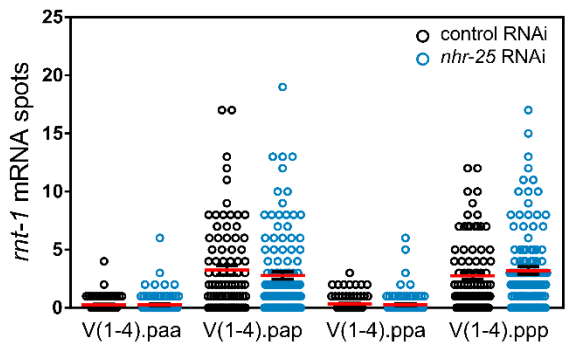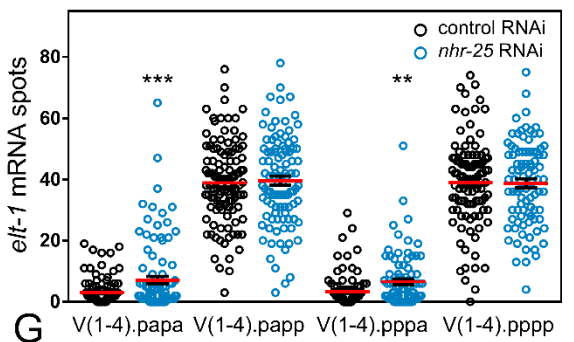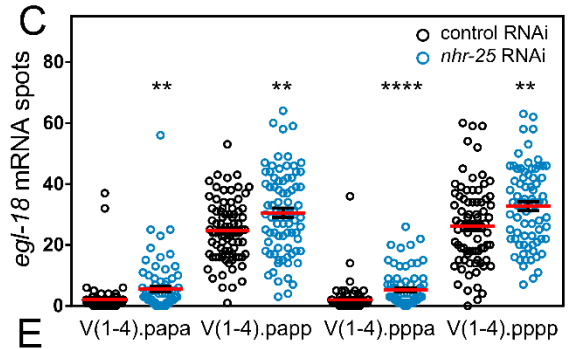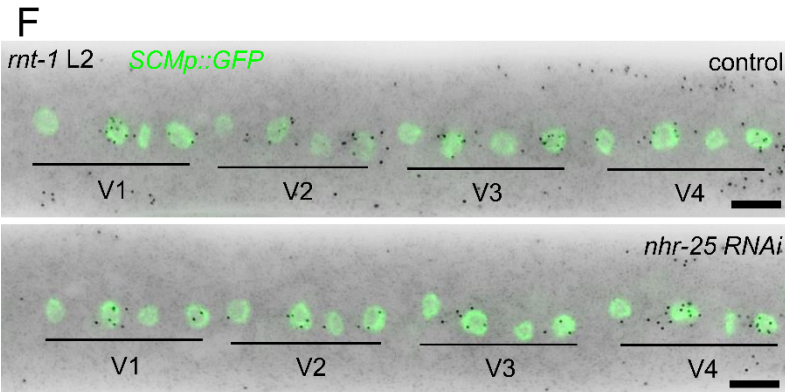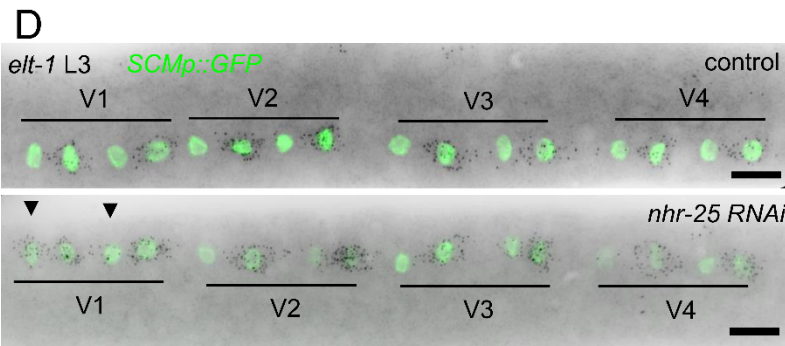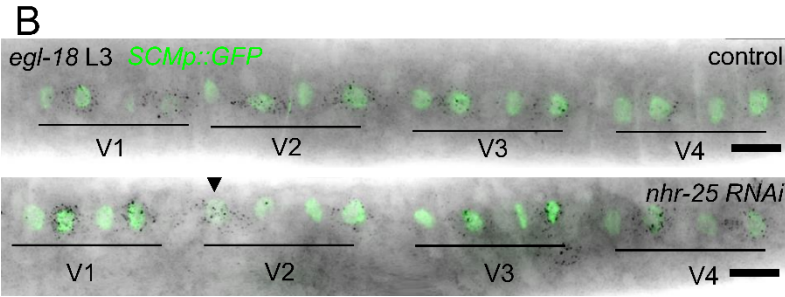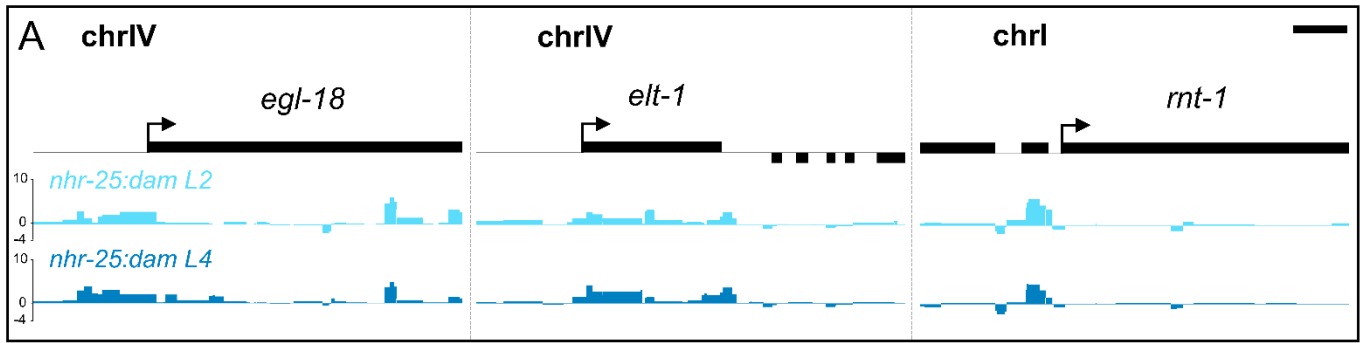
Figure 3. 12 Confirmation experiments of newly predicted NHR-25 targets: NHR-25 represses *egl-18* and *elt-1*. (A) *nhr-25:dam* signal enrichment profiles in regions with significant peaks around genes identified as putative NHR-25 targets. **(B-C)** Representative *egl-18* smFISH images from control and *nhr-25* RNAi treated animals during the L3 division (B) and quantification of *egl-18* mRNA spots in the V1-V4 L3 division daughter cells (C), showing a significant increase in expression in all cell lineages (60≤n≤88). **(D-E)** Representative *elt-1* smFISH images from control and *nhr-25* RNAi treated animals during the L3 division (D) and quantification of *elt-1* mRNA spots in the V1-V4 L3 division daughter cells (E), showing a significant increase in expression in the anterior daughters of the L3 V1-V4 asymmetric divisions (.papa and .pppa lineages) (82≤n≤116). **(F-G)** Representative *rnt-1* smFISH images from control and *nhr-25* RNAi treated animals during the L2 asymmetric division (F) and quantification of *rnt-1* mRNA spots in the V1-V4 L2 asymmetric division daughter cells (G), showing no difference in expression between control and *nhr-25* RNAi treated animals (88≤n≤124). **(H)** Seam cell number scoring in control and *nhr-25* RNAi treated animals based on *SCMp::GFP* marker expression. Animals were grown on the same plates used for smFISH to assess treatment efficacy (n≥36). In B, D, F the seam cells are labelled with the *SCMp:GFP* marker and black spots correspond to the respective mRNA. In C, E, G, H the error bars indicate the SEM and black stars statistically significant differences to the WT or control mean with a *t*-test (** $p<0.01$, *** $p<0.001$, **** $p<0.0001$). Scale bars are 2 kb in A and 10 μm in B, D, F.

(Brabin, Appleford & Woollard, 2011). smFISH for *elt-1* performed here, shows expression in those seam cell daughters of the L3 asymmetric division that are expected to retain their seam cell fate in control animals (Figure 3.12D). *nhr-25* RNAi treated animals often showed notably increased expression in anterior differentiating daughters (Figure 3.12D). This was captured in the quantification of *elt-1* transcript in all V1-V4 L3 division daughters, showing a significant increase in expression only in the anterior lineages (Vn.papa and Vn.pppa) that are normally expected to differentiate in the wild-type (Figure 3.12 E). Similarly to *egl-18,* this was more frequent in the anterior V lineages (i.e. V1 and V2) where more extra seam cells are observed by *nhr-25* knockdown (Hajduskova *et al.*, 2009). These data indicate that NHR-25 likely directly regulates *elt-1* by repressing its expression in anterior differentiating daughters. They also lend further support to the hypothesis that in the absence of NHR-25 seam cell specifying factors like EGL-18 and ELT-1 lead to extra seam cells either by ectopic maintenance of fate or by initiating symmetric divisions.

rnt-1, that was confirmed above as a direct target of LIN-22 was amongst the shared putative targets also identified in the NHR-25 TaDa. To assess if *rnt-1* was a target of NHR-25 as well, smFISH detection of its transcript was performed during the asymmetric division of L2, since reproducible expression patterns could not be acquired at L3 for this particular probe. Expression of *rnt-1,* both in control and *nhr-25* RNAi treated animals, was in posterior daughters that would maintain the seam cell fate (Figure 3.12F). Quantification did not reveal any differences in expression levels between the treatments (Figure3.12G), likely indicating that *rnt-1* is not directly regulated by NHR-25. Closer inspection of the *nhr-25:dam* and *lin-22:dam* signal near *rnt-1* revealed enrichment in a slightly different region for each factor perhaps belonging to the upstream gene in the case of NHR-25.

A technical caveat that could have masked an effect in this experiment was also the execution at L2 which might have not allowed enough time for the RNAi treatment to knock-down *nhr-25* sufficiently. However, to make sure that the RNAi treatment performed for the smFISH experiments had a strong effect, the efficacy was tested by growing animals on the same RNAi plates and scoring the seam cell number at

late-L4 as a proxy. A significant increase of the mean seam cell number from 16 to 21.7 was observed in *nhr-25* RNAi treated animals, in line with previous reports (Hajduskova *et al.*, 2009) (Figure 3.12H) (*p*<0.0001 with a t-test).

3.3 Discussion

3.3.1 Transcription factor target identification by TaDa in *C. elegans* is a powerful new methodology

Deciphering the make-up of gene regulatory networks controlling patterning events is a pivotal question in developmental biology (Davidson, 2010). Such networks of interconnected transcription factors (TFs) and their targets instruct fate decisions like those between differentiation and fate maintenance, or a self-renewing and a proliferative fate in the *C. elegans* stem-cell like seam cells (Joshi *et al.*, 2010; Brabin & Woollard, 2012). Understanding how participating TFs regulate each other and finding their targets can shed light onto conserved stem-cell behaviour.

To begin dissecting the seam cell/epidermis underlying regulatory network I adapted here targeted DamID (Southall *et al.*, 2013). In this chapter, I presented the first application of the targeted DamID (TaDa) (Southall *et al.*, 2013) method for identification of protein-DNA interactions in *C. elegans*. More specifically, I used it to identify targets of the transcription factors LIN-22 and NHR-25 in the epidermis by performing next-generation sequencing on the amplification products from methylated gDNA, making this also the first DamID-seq experiment for a transcription factor in *C. elegans*.

The experimental and genetic construct design was initially adapted and assessed in its characteristics in the context of *C. elegans*. Since our focus is on seam cell development, the target identification by TaDa was restricted by using the promoter of *wrt-2,* which is expressed predominantly in the seam cells and to a lesser extent in the hypodermis and rectal cells (Aspock *et al.*, 1999; Cao *et al.*, 2017; Pani & Goldstein, 2018), to drive expression of the Dam-fusions. This would allow for identification of targets bound by the two TFs only within the epidermis, rather than their complete native expression domain, increasing the likelihood they would be relevant to the tissue's development. *lin-22* is expressed specifically in the H0-V4 seam cells (Cao *et al.*, 2017; Katsanos *et al.*, 2017), already within the *wrt-2* expression domain but *nhr-25* has a broader expression domain and was selected here amongst other reasons to assess if only a subset of targets is identified in TaDa.

Tissue-specific or constitutive levels of expression of Dam-fusions have been shown to lead to toxicity and saturated non-targeted methylation (Southall *et al.*, 2013; van Steensel & Henikoff, 2000; Schuster *et al.*, 2010). However, the TaDa construct configuration/design overcomes these obstacles to permit tissue specificity (Southall *et al.*, 2013), which is also demonstrated here to be true in the epidermis of *C. elegans*. Expression of Dam-fusions as a secondary ORF alleviated the detrimental effect in fecundity and the excessive methylation that resulted from *wrt-2p* levels of expression of the fusions as primary

ORFs. The *wormCherry* primary ORF in TaDa transgenes also allowed for visualisation of successful expression in the expected *wrt-2* expression domain.

Importantly, both TFs, that were fused to the N-terminus of Dam to prevent obstruction of their N-proximal binding sites, showed functionality as well as capacity to produce gDNA GATC methylation. Multi-copy transgenes for both factors affected phenotypes related to their function. The *lin-22:dam* fusion partially rescued the multiple PDE phenotype of a *lin-22* mutant (Katsanos *et al.*, 2017; Wrischnik & Kenyon, 1997; Yip & Heiman, 2016), albeit not to the extent previously reported by overexpression of *lin-22* (Katsanos *et al.*, 2017) which can be attributed to the reduced expression levels driven by the TaDa configuration constructs even at a multi-copy level. The *nhr-25:dam* fusion potentiated the increased seam cell number phenotype of an *nhr-25* reduction-of-function mutant rather than mitigating it. This showed that the fusion retained some seam cell related functionality of *nhr-25*. In the case of *nhr-25*, it could also be due to self-regulation of the factor, found here and in previous studies (Araya *et al.*, 2014; Shao *et al.*, 2013) leading to a decrease in expression. The mode that NHR-25 acts in the anterior and/or posterior daughter during an asymmetric division is not yet known and therefore the *wrt-2p* driven overexpression might distribute transcript in a potentially ectopic manner, making this phenomenon hard to interpret mechanistically.

Overall, the technique required simple molecular cloning and the amenability of *C. elegans* to transgenesis allowed quick establishment of stable TaDa strains by MosSCI (Frøkjær-Jensen *et al.*, 2014) and simple confirmation experiments. Moreover, even though methylation was occurring only in the epidermis, isolation and amplification was sensitive and generated enough product for downstream processing even from ~2000 animals that can be easily grown on *dam⁻* food source.

Next-generation sequencing results had on average approximately 27 and 17 million 150 bp unique mappable reads for the LIN-22 and NHR-25 experiments respectively, showing that the protocol followed here produces sequencing depth well beyond previously described thresholds of around 10 million 50 bp reads (Askjaer, Ercan & Meister, 2014; Marshall *et al.*, 2016). It also demonstrates the potential for further multiplexing of samples that could significantly reduce the cost of the experiment. Aligned read-count normalised sequencing results, prior to calculation of $\log_2$ ratios, showed very good reproducibility between biological replicates within the experiments and good correlation across developmental stages. Results from control samples (*NLS-GFP:dam*) showed good correlation between them, both within but also across experiments, since to a large extent they capture chromatin state which is not random within a tissue. This suggest that technical inter-experimental variation does not hinder sample reproducibility and is not a major contributor of the resulting differences in read count maps, indicating that comparisons across experiments can be made with confidence. In the context of the present study, it illustrates that separate clustering exhibited by results for the two factors, acquired from separate experiments, most likely reflects true differences in binding preferences rather than experimental disparities.

After calculation of a single, averaged per GTAC fragment, $\log_2$(*TF:dam/NLS-GFP:dam*) normalised signal track for each TF at each stage, a key question was how much of the identified signal and in

particular the peaks of statistically significant enrichment, reflected genuine binding of LIN-22 or NHR-25. Available ChIP-seq data for NHR-25 were used to approach this question and only comparisons between the L2 datasets are discussed here.

ChIP-seq is a more established technique for TF target identification, with multiple applications for various TFs in *C. elegans* (Kudron *et al.*, 2018; Araya *et al.*, 2014), in contrast to the sole example of DAF-16 that has been performed by DamID (Schuster *et al.*, 2010). ChIP-seq signal peaks are generally accepted to represent TF binding sights, thus they were compared to TaDa reasoning that significant overlaps must exists. Signal profiles from the two methods showed qualitatively good agreement which was also reflected by the increased average TaDa signal over the locations of ChIP-seq peaks, in comparison to surrounding regions, across all the peaks. The TaDa NHR-25 dataset contained 2044 peaks compared to 5980 of the ChIP-seq. The TaDa peaks showed significant non-random overlaps with ChIP-seq peaks, with over a third of them overlapping with approximately 16% of the total ChIP-seq peaks. The difference in size of the initial datasets can be partly attributed to the increased resolution of ChIP-seq peaks (one TaDa peak can overlap multiple ChIP-seq peaks) (Aughey, Cheetham & Southall, 2019). Interestingly, it can also be attributed to the difference in the examined expression domains, with ChIP-seq target identification happening across all the tissues where NHR-25 is expressed (epidermis, glia, vulva) while TaDa is restricted to the epidermis. The common target genes between TaDa and ChIP-seq seem to regulate epidermis related functions and associate with epidermal tissues. In contrast, exclusively ChIP-seq target genes associated with expression in nerve and reproductive tissues and were not found to participate in similar biological processes.

This result provides evidence to the capacity of TaDa to dissect TF expression domains and perform identification of targets that are more likely to participate in functions of the studied tissue. TaDa peaks not overlapping with ChIP-seq could also be a result of identification in a subdomain of expression since tissue-specific binding might be more prominent when it is not diluted or averaged across all tissues where binding is occurring, making their discovery easier in TaDa. Lastly, based on the above the common peaks between TaDa and ChIP-seq are likely to be genuine targets. Lack of correlation in peak intensity between ChIP-seq and TaDa for those peaks, has been previously described (Cheetham *et al.*, 2018), and might reflect differences in the inherent biases between the two methods relating to the fusions used (Ramialison *et al.*, 2017), chemical crosslinking, antibody accessibility, PCR biases (Meyer & Liu, 2014) and the examined expression domains. However, here we see that peak intensity/score was the most influential criterion in identifying a good fitting motif for NHR-25 when applied as a filter to selects peaks of the TaDa profile, indicating that there is likely to be biologically relevant information in the quantitative aspect of the profiles. Therefore, caution is required when deciding whether to treat the data as qualitative or quantitative across methods.

3.3.2 TaDa-identified binding of LIN-22 and NHR-25 in the epidermis happens in putative regulatory regions and reflects factor-specific targets

TaDa signal for the transcription factors is expected to represent binding sites of LIN-22 or NHR-25 across the genome (Aughey & Southall, 2016). TFs often exhibit specific preferences in their genome-wide occupancy as they are recruited to *cis*-regulatory elements that usually reside in promoters or enhancers, often found upstream of genes that they regulate (Spitz & Furlong, 2012). Such enrichment localisation preferences were assessed for the TF signal profiles generated here, both as a measure of meaningfulness and biological relevance of the TaDa data and to better understand how LIN-22 and NHR-25 exert their regulatory action across their targets in the epidermis.

Encouragingly, initial inspection of TaDa signal in the profiles of both TFs, evidently exhibited frequent enrichment in intergenic regions often upstream of genes. Genome-wide assessment across all protein coding genes confirmed this preference, with substantially higher average signal in upstream to the TSS regions, as previously seen by DamID for DAF-16 binding (Schuster *et al.*, 2010), than within the gene or downstream of the TES. The most increased average signal was found proximally upstream to the TSS, as it has been previously observed for a multitude of transcription factors in ChIP-seq experiments (Araya *et al.*, 2014). Importantly, statistically significant TaDa peaks for both factors maintained this propensity for localisation. Almost half of the total peaks for both TFs at both stages were localising upstream of genes. More than half of those were overlapping with the TSS and of the rest half resided within the first 2 kb of the upstream to the TSS region. These results are in keeping with the finding that the majority of bound promoters in ChIP-seq experiments exhibit binding within the first 2 kb upstream of genes (Araya *et al.*, 2014). These regions are very likely to be promoters or enhancers, harbouring *cis*-elements that LIN-22 and NHR-25 would bind. In addition, peaks localising within genes showed substantial preference for introns compared to exons, as previously described both in DamID and ChIP-seq experiments in *C. elegans* (Schuster *et al.*, 2010; Fuxman Bass *et al.*, 2014). Introns are known to carry enhancers in *C. elegans* (Fuxman Bass *et al.*, 2014) and the preference likely reflects binding of those enhancers.

Further assessment of the regulatory potential of regions where TaDa peaks localised, using annotated accessible chromatin data (Jänes *et al.*, 2018), showed notable overlaps of at least 72% of total peaks with open chromatin elements. Coding promoters and putative enhancers were most often found to overlap with TaDa peaks for both TFs. Localisation characteristics and overlaps with known regulatory sequences exhibited by all TaDa peak profiles, highly resemble the expected occupancy of TFs and solidify the hypothesis that TaDa signal largely represents genuine binding for LIN-22 and NHR-25.

Interestingly, TaDa signal for LIN-22 was notably increased on average over locations of NHR-25 peaks and vice versa. This was further investigated to clarify if it resulted from TaDa related biases in the epidermis, which would manifest in profiles for both factors, or if it reflected genuine co-occurrence of LIN-22 and NHR-25 binding, due to their roles in epidermal development. Peaks for the two TFs showed

significant genome-wide overlaps with 20%-33% of peaks for each factor at each stage displaying across-factor overlaps. The vast majority of those overlaps were found to occur within promoter regions with less than a third of them in HOT regions, indicating that they are likely to be due to co-regulation of the same targets not widely shared with other transcription factors, thus potentially relating to the participation of both in epidermal development.

To further assess this, global pairwise co-association calculations of peak localisation patterns were performed across a selection of peak profiles of TFs related to epidermal or neuronal development and biological processes. Such analyses, so far mostly performed on ChIP-seq data, have previously shown that factors with peak localisation patterns that associate closely, often have shared ontologies or correlate with expression in the same tissue (Chikina & Troyanskaya, 2012; Kudron *et al.*, 2018; Araya *et al.*, 2014). LIN-22 and NHR-25 TaDa peak profiles were found to be most similar to other epidermal regulators than neuronal regulators. Moderate co-association seen between epidermal factors compared to neuronal factors might suggest more distinct groups of targets for each of them. Intriguingly, the TaDa NHR-25 profile showed far lower co-association with neuronal factors than the ChIP-seq profile strongly suggesting that TaDa successfully restricted target identification to the epidermis, excluding likely targets in the glia were *nhr-25* is also expressed. LIN-22 and NHR-25 TaDa profiles showed highest similarity with each other which along with the above supports the hypothesis that they have shared targets. Nevertheless, more TaDa profiles for other transcription factors will need to be acquired to assess how much of this similarity could also be method-driven.

Lastly, advocating to the common target hypothesis, when peaks for the two factors were assigned to sets of genes and those were intersected, the common set of genes was enriched for gene ontology terms, relating to development and neurogenesis, which are shared by both. Overall, these findings suggest that both LIN-22 and NHR-25 likely converge to the regulation of certain genes, to execute aspects of their functional roles in the epidermis.

3.3.3 TaDa reveals novel developmental links for LIN-22 with the heterochronic and Wnt signalling pathways and the cell cycle

The sets of genes identified by TaDa as putative targets of LIN-22 at L2 and L4 were explored in their content by performing enrichment analysis and comparing them with available datasets. Mutations in *lin-22* were first recovered from screens for their ectopic neurogenesis in V1-V4 seam cells, showing multiple PDE and PVD neuronal structures and aberrant mating rays and had been mostly studied on those grounds (Waring, Wrischnik & Kenyon, 1992; Wrischnik & Kenyon, 1997; Waring & Kenyon, 1990; Yip & Heiman, 2016). The Hox gene *mab-5,* a known factor involved in the ectopic neurogenesis was found here by TaDa and confirmed by smFISH (Katsanos *et al.*, 2017) to be directly repressed by LIN-22. Besides preventing neurogenesis in anterior to V5 seam cells, *lin-22* had been shown to play some role in seam cell number regulation (Wrischnik & Kenyon, 1997). We expanded on this in Katsanos *et al.*, 2017,

showing that *lin-22* is required for the correct establishment of seam cell division symmetry or asymmetry, potentially by antagonising Wnt signalling through suppression of the *frizzled* receptor *lin-17*. Encouragingly, these developmental and neurogenesis related functions of *lin-22* were reflected by the enriched GO terms for its putative targets at both stages. In addition, large overlaps were found with homologues of targets of HES1, the human homologue of LIN-22, from ChIP-seq experiments in cell lines. These overlaps reproduced the enrichment for some of those GO terms, suggesting that a large number of the discovered regulatory relationships might be conserved across species and overall highlight the biological relevance of the TaDa-identified targets.

Intriguingly, the sets of putative LIN-22 targets were also found to be enriched for genes encoding multiple components of the Wnt pathway strengthening the connection previously mentioned. More specifically, LIN-22 was found to potentially target components at various levels of the Wnt signalling cascade, like Wnt secretion *(mom-1, mig-14)*, the destruction complex (*mig-5, gsk3*), β-catenin (*bar-1*) and Wnt receptors (*mom-5, lin-17*) (Sawa & Korswagen, 2013). The broad array of targeted components indicates that the reported *lin-22* antagonism of Wnt signalling (Katsanos *et al.*, 2017) is likely to be both in the form of direct regulation and it is more extensive than previously thought. Moreover, the expression of *lin-17* had been shown to be inhibited in the anterior epidermis by *lin-22* (Katsanos *et al.*, 2017) and TaDa signal presented here indicated that it was most likely a result of direct repression, making this the first described instance of a *Hes*-related factor regulating the Wnt signalling pathway. In this study, the TaDa-predicted binding of LIN-22 on the *lin-17* promoter was used to further dissect the regulatory relationship and led to the identification of two conserved regulatory elements of the *lin-17* promoter, found to be under the control of LIN-22. These likely correspond to the *cis*-regulatory sequences that recruit the factor and lead to repression of *lin-17* expression.

Exhaustive lineaging of post-embryonic seam cell development in a *lin-22* mutant had pinpointed patterning defects that seemed to associate with division timing errors, suggesting a potential link between *lin-22* and the heterochronic pathway as previously hypothesised (Katsanos *et al.*, 2017). Specifically, both symmetric and asymmetric, temporally ectopic divisions were observed. TaDa identified the protein-coding genes *lin-14* and *lin-28* and the miRNA genes *lin-4* and *let-7* as targets of LIN-22, all of which are major regulators of the heterochronic pathway. *lin-14* is required for the L1 division and supresses division symmetry, while *lin-4* acts to deplete *lin-14* transcript to allow for the symmetric division of L2 to progress which is instructed by *lin-28* (Harandi & Ambros, 2015; Slack & Ruvkun, 1997). Expression of *let-7* at L4 cues the end of divisions and the onset of terminal differentiation (Reinhart *et al.*, 2000; Hayes, Frand & Ruvkun, 2006). Confirmation of the regulation by LIN-22, which is predicted by TaDa, could explain a plethora of seam cell patterning errors observed in *lin-22* mutants.

Ectopic symmetric divisions could also result from misregulation of the Runx homologue *rnt-1,* which is amongst the identified targets of LIN-22 and had been previously predicted to be a downstream genetic interactor (Zhong & Sternberg, 2006). The RNT-1/BRO-1 complex overrides the Wnt/β-catenin asymmetry pathway to execute the symmetric division of L2 (van der Horst *et al.*, 2019). *rnt-1* expression is found

here by smFISH to be increased in V1-V4 seam cells in the absence of *lin-22* suggesting that LIN-22 normally acts to repress its expression, likely confirming the TaDa identification. Additionally, *unc-62* is also amongst the identified targets and had been shown to be upstream of *rnt-1* in performing the symmetric division (Hughes *et al.*, 2013). These results highlight the high complexity in the biological regulation of the underlying gene network that will be interesting to further dissect.

A potential link with the cell cycle is also proposed by the data presented here, with two cyclins (*cya-1* and *cyd-1*) found in overlaps with mouse targets of *Hes1,* the *M. musculus* homologue of *lin-22*, as well as significant TaDa enrichment found around the cell cycle inhibitor *cki-1*. *cki-1* is expressed in the seam cells and knockdown by RNAi increases their number (Hong, Roy & Ambros, 1998; Buck, Chiu & Saito, 2009). The homologue of *cki-1*, p27^{Kip1} has been shown to be directly supressed by Hes1 in mice (Murata *et al.*, 2005) and is significantly downregulated in *lin-22* mutants (Katsanos *et al.*, 2017), conceivably contributing to the extra seam cells often observed in those mutants. smFISH experiments showed reduction of *cki-1* expression in V1-V4 seam cells in a *lin-22* mutant, validating the TaDa identification and indicating an activating role for LIN-22. *Hes*-related factors are thought to act as repressors (Kageyama, Ohtsuka & Kobayashi, 2007), as seen here in the case of *mab-5, lin-17* and *rnt-1*, making this an interesting finding demonstrating that LIN-22 can also likely act as a transcriptional activator. LIN-22 is an unusual hes factor lacking both the Groucho interacting WRPW domain and the Orange domain that associate with repressive roles, thus potentially allowing LIN-22 to act as an activator (Schlager *et al.*, 2006).

3.3.4 NHR-25 targets identified by TaDa uncover a new role in mediating differentiation programmes in the seam cells

As in the case of LIN-22, GO term enrichment analysis on sets of the TaDa-identified NHR-25 targets revealed involvement in functions previously associated with *nhr-25*. Like its homologue FTZ-F1 in *Drosophila*, NHR-25 plays a central role in cuticle formation and regulation of molting throughout the stages (Hayes, Frand & Ruvkun, 2006; Šilhánková, Jindra & Asahina, 2005; Gissendanner & Sluder, 2000), which is reflected by the enriched GO terms. In addition, apart from its role in embryonic epidermal development (Chen, Eastburn & Han, 2004), it is required postembryonically for a multitude of developmental events in the epidermis. It has been shown to mediate differentiation events required for vulva formation and hypodermal differentiation and fusion of seam cell daughters (Chen, Eastburn & Han, 2004; Gissendanner & Sluder, 2000). It also acts to determine a neural fate in a T seam cell division descendant (Hajduskova *et al.*, 2009). All such developmental and neural related functions of *nhr-25* are represented as enriched GO terms, both in the TaDa sets of targets identified here and in the significant overlaps they exhibit with NHR-25 ChIP-seq-identified targets. Relevant GO terms and the substantial overlaps that TaDa-identified targets demonstrate with ChIP-seq-identified targets, point to the compelling likelihood that the TaDa-identified targets constitute *bona fide* regulatory targets of NHR-25.

Specific identified targets with known roles in seam cell development were used to perform confirmation experiments. Two such targets were the GATA factors ELT-1 and EGL-18. ELT-1 is considered to be a master epidermal regulator (Chisholm & Hsiao, 2012), it is expressed postembryonically in the seam cells and specifies the seam cell fate, likely through promotion of seam cell fate determinants and also suppression of differentiation factors (Brabin, Appleford & Woollard, 2011). *egl-18* is a target for activation by the Wβa signalling pathway post-embryonically and is required to specify the seam cell fate during asymmetric divisions in non-differentiating daughters (Koh & Rothman, 2001; Gorrepati, Thompson & Eisenmann, 2013). Overexpression or ectopic expression of *egl-18* in daughters that should be committed to differentiation, results in ectopic seam cell fate maintenance and is associated with increased seam cell numbers (Gorrepati, Thompson & Eisenmann, 2013; Katsanos *et al.*, 2017).

smFISH experiments in this study found that during the asymmetric division of L3 animals with *nhr-25* knockdown showed significantly increased expression of *elt-1* in the daughters that should differentiate to the hypodermal fate. Similarly, *egl-18* expression by smFISH was shown to be increased in all division daughters in *nhr-25* knockdown. These findings likely confirm NHR-25 as a direct repressor of *elt-1* and *egl-18,* which were identified by TaDa and provide some proof of the biological relevance of the identified targets. More importantly, they propose a new mechanisms via which *nhr-25* is implicated in seam cells development.

So far, NHR-25 was thought to act on seam cell patterning by regulating seam cell shape and establishing cell-cell contacts that are required for tissue morphogenesis (Šilhánková, Jindra & Asahina, 2005; Chen, Eastburn & Han, 2004). The absence of *nhr-25* leads to increased seam cell numbers mostly in the anterior body as cells can divide ectopically or fail to differentiate (Chen, Eastburn & Han, 2004; Šilhánková, Jindra & Asahina, 2005; Hajduskova *et al.*, 2009). Results here indicate that NHR-25 likely actively participates in the differentiation process following asymmetric division, by repressing seam cell specifying factors in daughters that are destined to acquire the hypodermal fate. This regulation seems to have a stronger impact in more anterior lineages (V1-V2) as the agreement between the domain of increased expression of *egl-18* and *elt-1* and of the *nhr-25* knockdown phenotype suggest. These results highlight that TaDa and smFISH can provide a powerful combination to dissect new biological mechanisms underlying stem cell maintenance and differentiation.

Chapter 4

Profiling epidermal gene expression in *C. elegans* by assaying RNA-polymerase occupancy using targeted DamID

4.1 Introduction

Multicellular organisms are comprised of a plethora of differentiated tissues and cell-types that perform specific functions to support the organism's physiology. These specialised cell fates are determined during development through the establishment of distinct patterns of gene expression, by precise control of the selection of expressed genes and the temporal and quantitative characteristics of their expression. This is achieved by the establishment of a cell-type specific epigenetic regulatory state for the genome that results into transcriptional programs that are highly characteristic and descriptive of the tissue. Elucidating tissue or cell-type specific transcriptional profiles has been a persistent pursuit in biology, especially since the rise of high-throughput omics technologies (Kolodziejczyk *et al.*, 2015), as it both confers and constitutes the genetic basis of cellular identity in the context of a whole organism. Moreover, examining differences between gene expression profiles across tissues can help us uncover genes involved in specialised tissue-related functions. More importantly, differences between differentiated cells and their progenitors can pinpoint key cell-fate determinants involved in how cell-fate is acquired, maintained and altered when differentiation occurs. Understanding how transcriptome changes mediate such pivotal events is central to stem cell biology and can shed light to the mechanisms underlying their biology.

The *C. elegans* epidermal seam cells provide a good model to investigate such questions. They undergo stem-cell like divisions to proliferate by dividing symmetrically, or give rise to differentiated daughters while maintaining their fate by dividing asymmetrically (Sulston & Horvitz, 1977), constituting an ideal platform to examine the transcriptional states that specify the stem cell or differentiated cell fates. Seam cell asymmetric division daughters primarily differentiate towards the hypodermal fate making both progeny cell-types of the epidermis. The ability to compare the gene expression profiles between the two could allow for discovery of genes broadly implicated to conferring their epidermal character, but also specialised gene batteries involved in determining the stem-cell character of seam cells and regulating aspects of their patterning program. Acquiring such cell-type specific gene expression profiles remains however challenging.

Traditionally, for multicellular organisms this is achieved by isolating the tissue or cell-type of interest prior to mRNA extraction and sequencing (Kolodziejczyk *et al.*, 2015). In *C. elegans* larvae this is particularly difficult due to the physical barrier formed by the cuticle surrounding all the tissues and associating especially tightly with the seam cells and hypodermis (Page, 2007). Mechanical or laser based dissection is particularly challenging especially for cuticle associated cells and to separate tissues chemical dissociation of the cuticle is required to produce cell suspensions (Zhang & Kuhn, 2013) that can undergo selection. Methods based on cell or nuclei sorting require a lot of material and suffer from selection related biases (Spencer *et al.*, 2014; Deal & Henikoff, 2011), while mRNA-tagging methods can be toxic (Yang, Edenberg & Davis, 2005). The valuable sci-RNA-seq approach has high resolution but is challenging to perform and analyse and is expensive for more focused questions (Cao *et al.*, 2017).

Tissue-specific gene expression profiling can also be performed by assaying the occupancy of RNA pol using TaDa (Southall *et al.*, 2013). The expression of the Dam-RNApol fusion can be driven in any tissue or tissue subdomain or cell-type for which promoters are available with no requirement for cell isolation or any manipulation prior to nucleic acid extraction. The methylation marks that indicate expression occur *in vivo* in almost wild type conditions since the design of the system prevents toxicity. Importantly, it requires only a fraction of the material needed for other methods and shares protocols and reagents with all other TaDa applications (i.e. for identification of TF targets and accessible chromatin).

In this chapter, gene expression profiling by TaDa is used for the first time in *C. elegans* towards identifying seam cell specific factors that are involved in the tissue's development. This is attempted by acquiring and comparing gene expression profiles for the seam cells and hypodermis. Since the two cell-types are developmentally linked and differentiated seam cells fuse to the hypodermis, very strict seam cell and hypodermal promoters were generated to avoid artificial overlaps and allow for meaningful intersections of the profiles. Versatile tools to aid the production of TaDa transgenics were built and used to generate lines for seam cell and hypodermal assaying of RNApol occupancy at the L2 and L4 stage.

The resulting occupancy profiles were assessed for signatures signifying the capturing of gene expression and for agreement with existing knowledge on patterns of spatiotemporal specificity of expression. The biological relevance of the TaDa-acquired sets of expressed genes was examined by enrichment analysis for their tissue-specificity and involvement in expected tissue-related ontologies, as well as by comparison to existing seam cell and hypodermis specific transcriptomes. The exclusively seam cell expressed genes were mined for TFs and chromatin factors as potential regulators of seam cell development, for which an RNAi screen looking at seam cell numbers was performed for functional assessment of their roles. A versatile platform for tissue or cell type specific knockdown was built and tested in the epidermis to support candidate confirmation experiments. Lastly, the capacity to identify tissue specific expression of miRNAs and their involvement in epidermal development was investigated. Overall, this chapter demonstrates the feasibility of performing gene expression profiling by TaDa in *C. elegans*, provides evidence of the comparability to other methods of tissue-specific transcriptomics and expands the selection of factors involved in seam cell development by identifying new participating TFs, Chromatin factors and miRNAs.

4.2 Results

4.2.1 Promoter specificity to resolve seam cell and hypodermal gene expression using TaDa

The seam cells and the hypodermis are developmentally linked. After each asymmetric larval division one daughter, for most of the seam cells, differentiates to hypodermal fate and fuses to the hyp7 syncytium (Sulston & Horvitz, 1977). In addition, both the seam cells and hypodermis are part of the epidermis and

are therefore expected to share aspects of their transcriptional programs to some extent (Chisholm & Hsiao, 2012; Cao *et al.*, 2017). Since TaDa is based on using tissue-specific expression of Dam-fusions, in order to perform gene expression profiling in the epidermis, promoters that can discriminate between the seam cells and hypodermis are required. Identifying such promoters is somewhat challenging in our system due to the common epidermal fate and the need for homogenous spatiotemporal expression within each cell-type.

Moreover, the exclusivity of the expression domains driven by the promoters is crucial especially because the fusion of differentiated seam cells to the hypodermis creates another level of complexity. Specifically, the content of each seam cell diffuses into the hyp7 syncytium after every differentiation and fusion event, creating the potential for remaining seam cell expressed Dam-RNApol fusions to methylate genes expressed in the hypodermis, reducing the specificity of the seam cell profiles. Therefore, the chosen promoters for both tissues need to be highly specific with particular emphasis on the hypodermal promoter fully excluding the seam cell domain. This would allow for subtraction from the seam cell profiles of both commonly expressed genes and falsely identified genes resulting from the transmission of the fusion.

In the case of the seam cells, the commonly used promoter of *wrt-2,* employed in chapter 3, is predominantly expressed in the seam cells but is also actively expressed in the hypodermis (Aspock *et al.*, 1999; Cao *et al.*, 2017; Pani & Goldstein, 2018). Assaying RNApol occupancy using it, would produce profiles covering both cell-types. On the other hand, the traditional *SCM* enhancer is large in size to be ideal for cloning and single-copy transgenesis (Hope, 1991; RM *et al.*, 1995), thus alternative options had to be explored.

The promoter of the nucleotide sugar transporter gene *srf-3* has been reported to drive strong specific expression in all seam cells, from the late embryo until the terminal differentiation, using reporter gene analysis (Höflich *et al.*, 2004). Cell-type-specific transcriptomic data show that it is strongly expressed in the seam cells (37th most highly expressed gene of the tissue based on (Cao *et al.*, 2017)) with at least 5.3x higher expression than in the next most highly expressing tissue (Cao *et al.*, 2017).

In the current genome annotation there are 3 major isoforms for *srf-3* with potentially different promoters. The promoter previously reported to drive expression in the seam cells is the complete sequence from the 3'UTR of the upstream gene until the ATG of isoform b (Höflich *et al.*, 2004) (*srf-3bp*) and fully includes the putative promoter of isoform a (*srf-3ap*) (Figure 4.1A). To pinpoint the sequence with the seam cell expression driving capacity, the promoters *srf-3ap* and *srf-3bp* were inserted into reporters of histone localised *C. elegans*-optimised GFP (GFPo) and single-copy transgenics were produced as for insertion of TaDa constructs. The putative promoter of isoform a could drive expression in intestinal cells, hypodermal cells, the germline and spermatheca, with only faint expression in the seam cells (Figure 4.1B top). This pattern did not agree with the reported expression pattern from the transcriptomics data (Cao *et al.*, 2017) or reporter data for the isoform b putative promoter sequence (Höflich *et al.*, 2004). This discrepancy was further resolved here with the *srf-3b* promoter construct achieving a substantial shift

towards seam cell expression in comparison to the other tissues without completely abolishing non-seam cell expression (Figure 4.1B middle).

Since the inclusion of the sequence of the 1st intron of *srf-3* isoform a (*srf-3i1*) (Figure 4.1A) had such a striking effect in shifting the expression towards seam cell preference, the hypothesis that it contains a seam cell specific enhancer element was tested. The *srf-3i1* sequence was fused to a minimal core *pes-10* promoter to ensure expression initiation capacity and was inserted in a reporter of the same configuration as the other promoters. Single-copy transgenic animals were generated. Remarkably the *srf-3i::pes-10* promoter drove strong and specific expression in the seam cells without evidence of expression in other tissues (Figure 1B bottom). The expression was observed in all seam cells and throughout post-embryonic development up to terminal differentiation. Having taken this element out of its native genomic context evidently led to isolation of the seam cell expression capacity and abolishment of the expression in other tissues, which is likely regulated by the rest of the *srf-3* promoter in the endogenous context. On this basis, the *srf-3i1::pes-10* promoter was selected to drive expression of the Dam-RNApol fusion for TaDa gene expression profiling in the seam cells and is abbreviated in the rest of this study to *srf-3i1* for simplicity.

With regard to the hypodermal-specific promoter, a popular option is the promoter of the collagen gene *dpy-7* that has been widely used and is commonly associated with the hypodermal fate (Johnstone, Shafi & Barry, 1992; Brabin, Appleford & Woollard, 2011; Blazie *et al.*, 2017). However, careful microscopic observations revealed that *dpy-7p* drives low expression levels in the seam cells, which was more evident during or right after seam cell divisions (Figure 4.1C). More specifically 217 out of 223 observed seam cells from various larval stages showed expression of nuclear *mCherry-H2B* driven under *dpy-7p* from a single-copy transgene. This observation is in keeping with sci-RNA-seq data reporting almost equal amounts of transcript for the two cell-types (Cao *et al.*, 2017).

A closer inspection of the short 309 bp *dpy-7* promoter revealed three putative GATA binding sites positioned at -6, -135 and -250 bp from the ATG of the gene (Figure 4.1D). Two of those sites (-6 and -250 bp) were AGATAA in sequence while the -135 bp was TGATAA. The TGATAA sites have been associated with hypodermal expression and have been identified as the binding sites for the GATA factor ELT-3 that regulates the hypodermal fate (Gilleard & Mcghee, 2001; Shao *et al.*, 2013). In contrast, a point mutation altering an AGATAA site of the *lin-22* promoter *(ot269 allele)* has been shown to significantly hinder the native seam cell expression of *lin-22* resulting to a mutant phenotype (Katsanos *et al.*, 2017). This indicates a likely regulatory role in promoting seam cell expression. Based on that hypothesis a synthetic version of the *dpy-7* promoter named *dpy-7syn1* was built, replacing the AGATAA sites with TTGATAA sites (Figure 4.1D). The promoter was inserted into a reporter driving histone-localised *mCherry* and was assessed for expression in the seam cells as a multi-copy transgene. The expression driven by *dpy-7syn1* was found to be exclusively hypodermal with 0/220 observed seam cells showing *mCherry* expression. Accordingly, the *dpy-7syn1* was selected to drive expression of the Dam-RNApol fusion for

A

Chromosome IV: 12848221-12854660

txt-19 3'UTR

srf-3ap

srf-3i1

srf-3bp

srf-3 isoform a

srf-3 isoform b

B

srf-3ap::GFPo-H2B

srf-3bp::GFPo-H2B

srf-3i1::pes-10::GFPo-H2B

C

arf-3:pes-10::GFP-CAAX
SCMp::GFP

dpy-7p::mCherry-H2B

D

dpy-7p   CTTATCT   TTATCA   AGATAA

dpy-7syn1   TTGATAA   TTATCA   TTGATAA

-309 bp   -250 bp   -135 bp   -6 bp   ATG

E

arf-3:pes-10::GFP-CAAX
SCMp::GFP
dpy-7syn1::mCherry-H2B
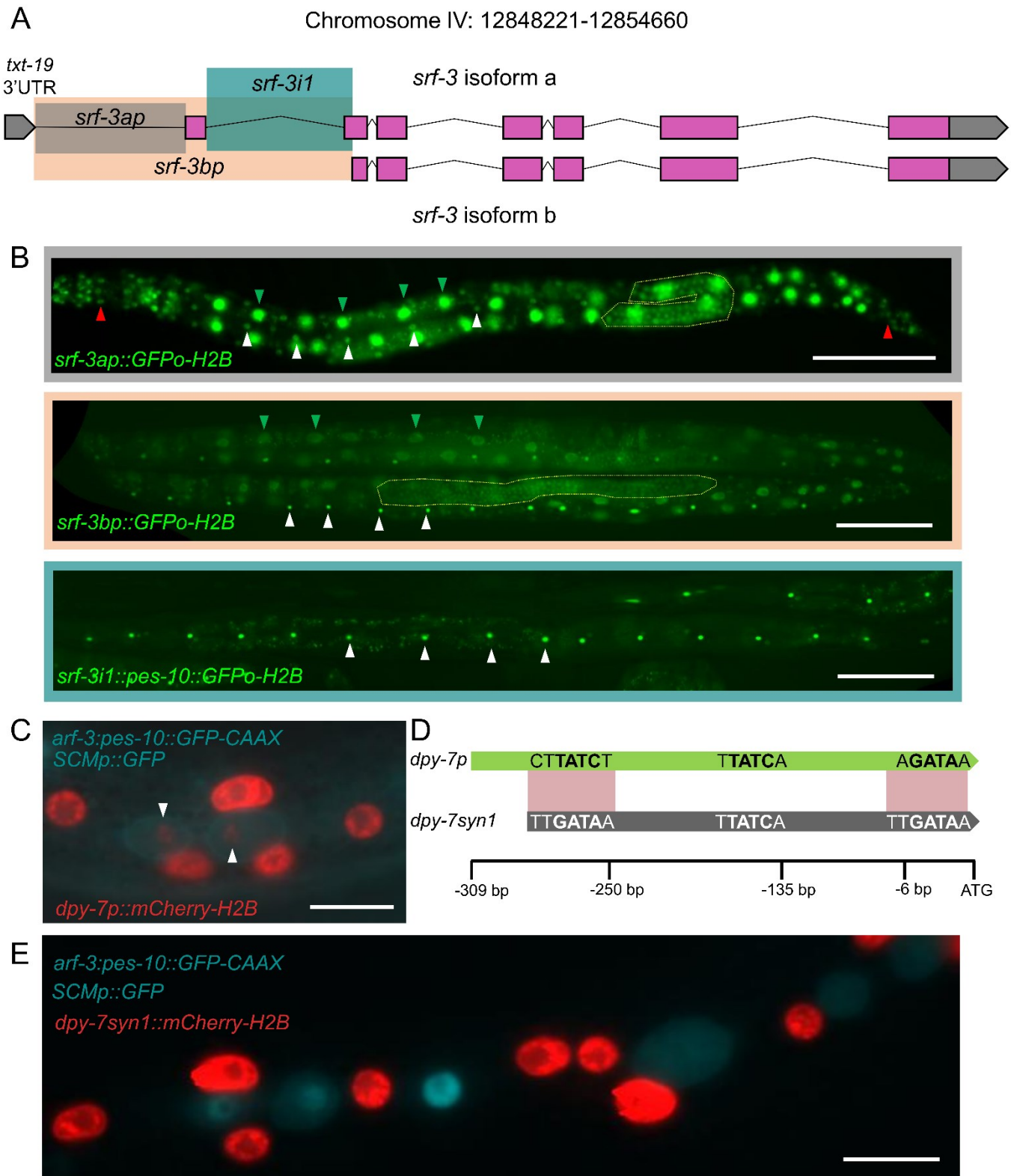
Figure 4. 1 Discovery of seam cell and hypodermis specific promoters for RNApol TaDa. (A) Illustration of the genomic locus at the position of the gene *srf-3* on chromosome IV. Pink blocks signify exon sequences and grey pointers 3' UTRs. Two isoforms of *srf-3* (isoform a and b) are shown. Shaded areas near the 5', mark putative promoter or regulatory sequences tested for seam cell specific expression. With grey, the 1093 bp putative promoter sequence of isoform a (*srf-3ap*), extending from the end of the upstream gene *txt-19* 3'UTR to the *srf-3* isoform a start codon. With peach, the 2246 bp sequence starting at the same position and extending to the start codon of isoform b (*srf-3bp*) and with teal, the 1081 bp first intron of isoform a (*srf-3i1*). **(B)** Representative fluorescence images of late L4 transgenic animals carrying single-copy transgenes of transcriptional reporters driving expression of *GFP-H2B* under the *srf-3ap* promoter (grey frame), the *srf-3bp* promoter (peach frame) and the *srf-3i1::pes-10* promoter (teal frame).White arrowheads indicate examples of expression in seam cell nuclei, green in intestinal and red in hypodermal. Yellow outlined areas indicate expression in the germline. **(C)** Representative fluorescence image showing expression of *mCherry-H2B* under the promoter of *dpy-7* from a single-copy transgene, mainly in hypodermal nuclei. Seam cells, marked in cyan by membrane (*arf-3:pes10::GFP-CAAX*) and nuclear (*SCMp::GFP*) reporters, show expression of *dpy-7p:mCherry-H2B*, indicated by white arrowheads, which is more prominent during divisions. **(D)** Illustration of the *dpy-7* promoter with the positions, given in distance from the endogenous ATG, of AGATAA sites altered to TTGATAA, indicated by red shading, to produce the synthetic *dpy-7syn1*. The sequences are given as they are on the forward strand. **(E)** Representative fluorescence image showing expression of *dpy-7syn1::mCherry-H2B* from a multi-copy transgene, in hypodermal nuclei but absence from seam cell nuclei at various stages before during and after divisions. Seam cells are marked as in C. Scale bars are 100 µm in B and 10 µm in C and E.

TaDa gene expression profiling in the hypodermis. In addition, further evidence of the association of AGATAA sites with seam cell expression was produced.

4.2.2 Generation of transgenic lines using a versatile TaDa cloning platform

As outlined in section 1.6.1 the TaDa transgene design permits the low-level, tissue-specific expression of a protein of interest fused to Dam in an orientation that does not obstruct its DNA-interacting capacity, by placing it downstream of a primary ORF like *mCherry*. To mediate quick and simple assembly of such constructs for any tissue or protein of interest and assist generation of the transgenes used in this chapter, a versatile genetic construct was designed and built.

This TaDa cloning platform carried all the key features required to build and insert in the genome, as single-copy, any TaDa construct for *C. elegans* applications (Figure 4.2A). In more detail, it was based on the pCFJ151 universal MosSCI backbone (Frokjaer-Jensen *et al.*, 2014) and therefore the Dam-fusion expression unit is flanked by recombination sequences for the universal MosSCI locus *ttTi5605.* This allowed for single-copy insertion on any chromosome, depending on experimental needs and transgenic selection based on rescue by the included *cb-unc-119* cassette. It also contains an *attR4-attL1* Gateway® cloning docking site that includes a *ccdb* bacterial-lethal gene and a chloramphenicol resistance gene. This permits the restriction digest-free insertion of any promoter of interest with an LR reaction, taking advantage of the gateway design of the *C. elegans* promoterome project (Dupuy *et al.*, 2004). The promoter cloning site is followed by the *C. elegans* codon-optimised *wormCherry* as a primary ORF, which is in turn followed by two STOP codons, a nucleotide for frameshift and a 6 bp unique restriction site of XmaJI before the ATG of the *dam* gene. On the 3' end, *dam* is fused to a *myc*-tag followed by three unique

A

XmaJI

AAG TAG TAA C CCTAGG ATG

ttTi5605_R attR4 CmR attL1 wormCherry dam unc-54 3'UTR Cb-unc-119 ttTi5605_L

CTG CG GCATGC A TTAATTAA GT GCTAG
PaeI   PacI   NheI

AmpR

B

i srf-3i1 wormCherry 2x STOP dam ama-1   iv dpy-7syn1 wormCherry 2x STOP dam ama-1

ii srf-3i1 wormCherry 2x STOP dam rpb-6   v dpy-7syn1 wormCherry 2x STOP dam rpb-6

iii srf-3i1 wormCherry 2x STOP dam NLS-GFP   vi dpy-7syn1 wormCherry 2x STOP dam NLS-GFP

C

D

replicate 1   replicate 2

L2   L4   L2   L4

3 kb —

100 bp —

iii ii vi v iii ii vi v   iii ii vi v iii ii vi v

Figure 4. 2 A novel *C. elegans* TaDa cloning platform is used to generate Dam-RNApol transgenic lines that show cell-type-specific expression and reproducible gDNA methylation. (A) Graphic illustration of the versatile *C. elegans* universal TaDa cloning platform with its key features. From left to right the plasmid contains: universal MosSCI recombination sites (ttTi5605_R and L), an LR attR4-attL1 Gateway® cloning site for promoter insertion, a *wormCherry* primary ORF followed by 2 x STOP codons, indicated in red, a frameshift in yellow and a unique XmaJI restriction site followed by *dam*, unique restriction cites for PaeI, PacI, NheI and an in-frame STOP codon prior to an *unc-54* 3'UTR. An *unc-119* expression cassette is also included to aid screening of transgenics. **(B)** Illustrations of the key features of single-copy transgenes used in this study for RNApol occupancy probing by TaDa in the seam cells and hypodermis. For the transgenes i, ii, iii the *srf-3i1:pes-10* promoter has been abbreviated to *srf3i1*. **(C)** Confirmation of single-copy transgene expression in the expected tissue of interest, for transgenes in B, using *wormCherry* expression as a proxy. Animals were imaged at the L4 stage. White arrowheads indicate expression in the seam cells while white arrows indicate absence of expression in the seam cells in animals expressing in the hypodermis, indicated by the yellow outline. Scale bars are 20 μm. **(D)** Representative electrophoresis of amplification products from methylated gDNA from animals carrying the transgenes indicated at the bottom, showing 200-2000 bp smears. Extractions were performed at L2 and L4 stages with each combination of promoter driving expression of *dam:rpb-6* or *dam:NLS-GFP* at each stage being represented by two biological replicates. The same volume of amplification product was loaded for each sample in each gel. Note the reproducibility of observable band patterns seen in *dam:rpb-6* samples.

restriction sites (PaeI, PacI, NheI) before the STOP codon and the 3'UTR from *unc-54*. The upstream and downstream restriction sites allow choice between N- or C- terminal fusions of the protein of interest with Dam, depending on the topology of the DNA interacting domain. In addition, the two different modes of cloning for the protein of interest and promoter, prevent any potential

clash between sequences and unique restriction sites as long as the promoter is inserted second. This universal TaDa cloning platform can be a valuable tool for any TaDa application in *C. elegans* and was used to readily create 6 transgenes for this chapter.

Tissue-specific gene expression profiles are generated by TaDa based on assaying genome-wide RNApol II occupancy. To acquire such occupancy profiles, using the above platform, *dam* was fused upstream of the major RNApol II subunit gene *ama-1* (*dam:ama-1*), that has been previously used in ChIP-seq experiments (Araya *et al.*, 2014) and of *rpb-6* (*dam:rpb-6*), encoding the subunit 6 that participates in all RNApol complexes and its homologue in *Drosophila* has been successfully used in DamID experiments (Filion *et al.*, 2010). Transgenic lines were created driving these fusions as well as a control (*dam:NLS-GFP*), in a TaDa configuration under the two selected promoters for seam cell and hypodermis expression, *srf-3i1* and *dpy-7syn1*. The key features of these transgenes are illustrated on Figure 4.2B.

The expression of *wormCherry* from the primary ORF was used as proxy to confirm that all single-copy transgenes drove expression in the expected tissues. Microscopy at the L4 stage for all transgenic lines showed expression in the seam for the *srf-3i1* transgenes and in the hypodermis, with noticeable exclusion of expression in the seam cells, for the *dpy-7syn1* transgenes (Figure 4.2C). These expression domains named after their promoters are used interchangeably with the tissue names that they correspond to in the rest of this study.

The capacity for methylation of the fusions was tested, as is typical for DamID, by extraction of methylated gDNA, amplification and electrophoresis of the product. The *dam:ama-1* fusion failed to

produce any detectable methylation in any of the stages, tissues or experimental replicates or repeats attempted (not shown). Moreover, sequencing results for both tissues for this fusion at L2 and L4 produced on average only 1.9 million unique mappable reads (Appendix B.1) that do not suffice for downstream analysis. The AMA-1 *Drosophila* homologue RpII215 has produced methylation patterns in a fusion with Dam at the same orientation (Southall *et al.*, 2013). However, it remains unclear whether in *C. elegans* the AMA-1 protein structure conformation somehow obstructs Dam from interacting with DNA effectively. In contrast, the *dam:rpb-6* fusions produced efficient observable methylation in all stages and tissues (Figure 4.2D). Moreover, *rpb-6* is substantially smaller in size than *ama-1* (coding sequences: 414 bp and 5571 bp respectively) making it easier to work with in terms of both cloning and transgenesis, as well as it is expected to participate in all 3 RNApol complexes (Shpakovski *et al.*, 1995) allowing for assaying of total transcription. For these reasons the *dam:rpb-6* fusions were chosen to assay RNApol occupancy for profiling of gene expression in the seam cells and hypodermis.

gDNA was collected at L2 and L4 stages and for each promoter, driving the *dam:rpb-6* or control *dam:NLS-GFP* fusions*,* two biological replicates were performed in parallel. The amplification products of the methylated DNA for each sample are presented on Figure 4.2D and show smears of various intensities between 2 kb and 200 bp for all samples. Reproducible band patterns are observed in products from *dam:rpb-6* fusions across the promoters and stages, which are not observed in control samples. These are different from the patterns observed in *lin-22:dam* fusion samples indicating that they may correspond to GATC fragments showing frequent methylation in the epidermis, potentially as parts of highly expressed genes.

4.2.3 Sequencing results reveal substantial similarities in RPB-6 occupancy across developmental stages and epidermal cell types

The amplification products for all samples underwent next-generation sequencing to identify profiles of RNApol occupancy. For all samples, between 11.7 to 24.2 million 150 bp-long unique mappable reads were found with genomic coverage ranging between 17 to 25x times, significantly higher than previously reported thresholds (Askjaer, Ercan & Meister, 2014; Marshall & Brand, 2015). The genome-wide sequence alignment read count-normalised maps generated from the sequencing results of *dam:rpb-6*, showed very high correlation between replicates and across stages. Maximum Pearson's correlation coefficient values of 1 were found for both *srf-3i1* and *dpy-7syn1 dam:rpb-6* samples (Figure 4.3A, B), indicating strong reproducibility between replicates but also providing preliminary evidence for increased similarity of occupancy profiles across stages. The control *dam:NLS-GFP* fusion samples showed higher correlation between them than with the *dam:rpb-6* samples, resulting in separate clustering of their maps for both promoters (Figure 4.3A, B), lending support to the biological meaningfulness of the profiles.

Correlation coefficient values between the control samples where more modest ranging between 0.69 and 0.92 for *srf-3i1* and 0.56 and 0.90 for *dpy-7syn1*. This could be somewhat explained by the more

random nature of methylation produced by the non-targeted *dam:NLS-GFP*. Nevertheless, this contradicts the previously observed trend between the *wrt-2p* driven *NLS-GFP:dam* control samples, in chapter 3, that showed higher correlation (Pearson's correlation coefficient values of 0.9 to 0.98). The extent to which these differences are meaningful and have biological underpinnings or are technical in nature remains to be seen.

Pearson's correlation comparisons between the sequence alignment read count maps were also made across the two expression domains for all samples. As expected controls samples clustered separately from the *dam:rpb-6* samples that strikingly showed very high correlation even across cell-types with all coefficient values approximately 1 (Figure 4.3C). This was further demonstrated by a principal component analysis where all *dam:rpb-6* samples grouped very tightly and separately from the control samples (Figure 4.3D). These observations indicate that the discovered read count maps are highly similar between the cell-types, which is very likely to reflect their common epidermal character that involves shared expression of multiple genes, as has been previously reported (Cao *et al.*, 2017).

It should be noted that since RPB-6 participates in all RNApol complexes the extensive similarity observed particularly across cell types, demonstarated with the evident overlap on the PCA space, is also likely to be driven by the expression of RNApol I or III loci. These genomic regions are expected to be expressed at high levels and approximately equally across cell types thus increasing the correlation between *dam:rpb-6* read-count maps. In contrast, the control samples are more sparsely scattered on the PCA space which is in agreement with their more reduced Pearson's coefficient of correlation values. This could further support the more stochastic methylation produced by these fusions and the fact that these samples are expected to reflect chromatin accessibility that is likely to be somewhat different across cell-types.

As in chapter 3, the meaningful genome-wide RPB-6 occupancy signal is calculated from stage and promoter-matched *dam:rpb-6* and *dam:NLS-GFP* sequencing results and reported in normalised $\log_2$(*dam:rpb-6/dam:NLS-GFP*) scores per GATC fragment of the genome. As previously described, the availability of 2 replicates per sample allowed for 4 pairwise calculation per promoter and stage (representative examples in Appendix C.6), with the resulting profiles averaged into a single RPB-6 occupancy signal profile for each promoter at each stage. A representation of those four profiles across chromosome I is presented in Figure 4.3E, labelled just based on the expression domain and the stage (complete genome-wide signal tracks are available in Appendix C.7, C.8). They show peaks that are expected to cover genic regions of expressed genes. These profiles were used for all downstream analyses presented here.
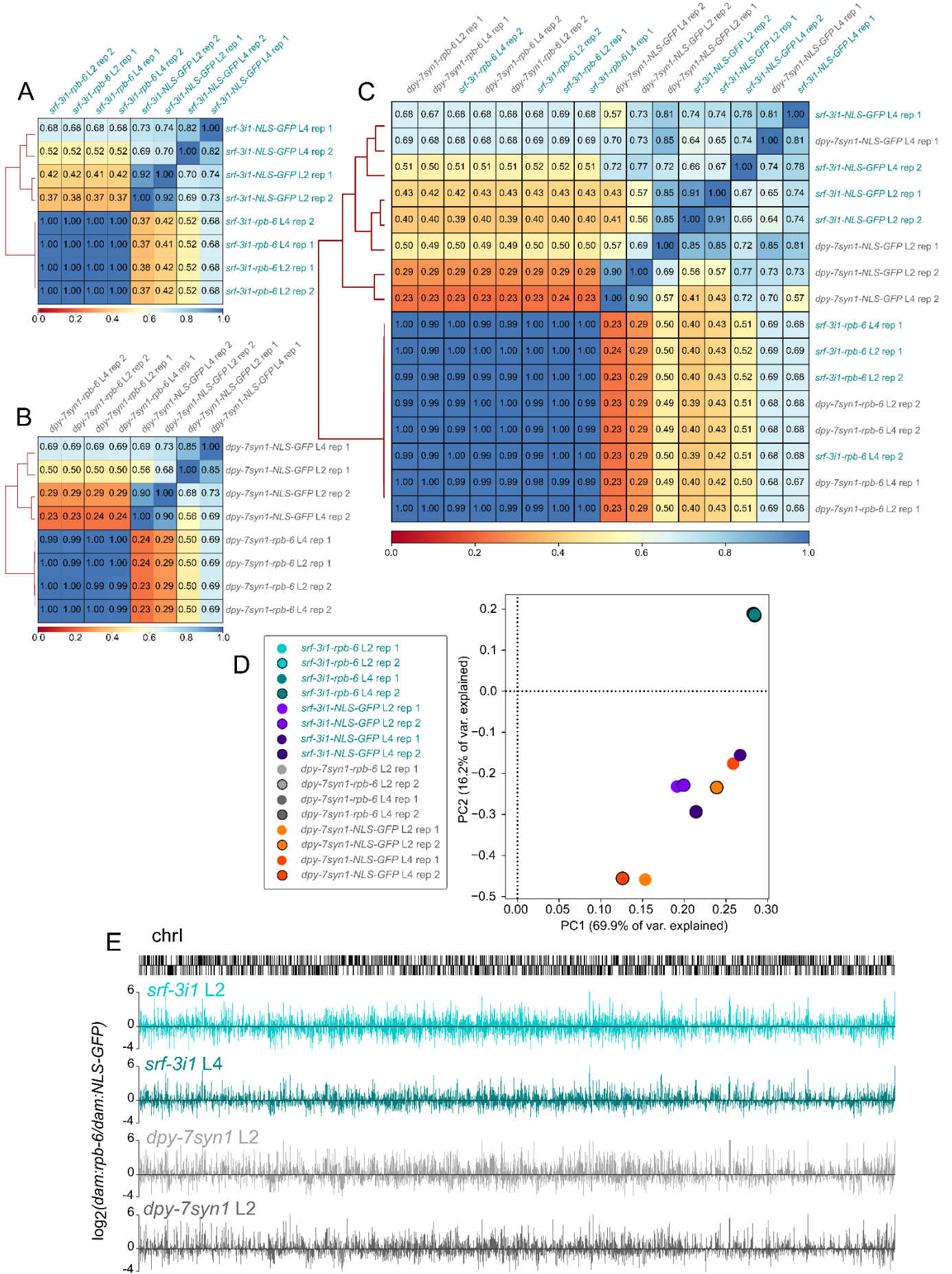
Figure 4. 3 Sequencing results indicate highly similar RPB-6 occupancy signatures across tissues. (A-B) Pearson correlation heatmaps based on normalised aligned read count maps for the *srf-3i1* (A) and *dpy-7syn1* (B) RNA pol TaDa samples. The correlation coefficient for each pairwise comparison is printed in each respective cell of the heatmaps. For both promoters, the *dam:rpb-6* samples show very high correlation coefficients between replicates and stages and low correlation with the respective control *dam:NLS-GFP* samples, which show moderate correlation between them and cluster separately. **(C)** Summary heatmap of Pearson correlations between all samples for the RNApol TaDa performed in this study. All normalised aligned read count maps from *dam:rpb-6* samples show very high correlation across the interrogated tissues, clustering separately from the control samples that are moderately correlated across tissues. **(D)** Principal component analysis on normalised aligned read count maps for all samples shows tight clustering of *dam:rpb-6* samples that form a separate group from control fusion samples for both promoters. **(E)** Example of averaged signal enrichment profiles for *dam:rpb-6* occupancy across chromosome I for the seam cell *srf-3i1* and hypodermal *dpy-7syn1* promoters in L2 and L4 stages. The Y-axes represent normalised $\log_2(dam:rpb-6/dam:NLS-GFP)$ scores. Scale bar length is 2 Mb.

4.2.4 RPB-6 occupancy occurs in gene bodies with spatiotemporal specificity that follows known gene expression patterns

To assess the biological relevance and the spatiotemporal specificity of the acquired occupancy signal, closer inspection of the signal enrichment profiles was performed at the loci of selected genes presented on Figure 4.4. Cases of significant occupancy across the gene body (FDR<0.05) that constitutes the TaDa determination of an actively expressed gene are mentioned below.

The major seam cell fate regulators *elt-1*, *egl-18* and *elt-6* that are known to be expressed in the seam cells (Katsanos *et al.*, 2017; Cao *et al.*, 2017; Gorrepati, Thompson & Eisenmann, 2013) show signal enrichment within their sequences and are found to be significantly occupied and therefore expressed, only in the seam cell profiles. Furthermore, *srf-3* and the complex of groundhog genes *grd-13*, *grd-3* and *grd-10* show signal enrichment and significant occupancy only for the seam cells. All of the above are known to be expressed in the seam cells but not the hypodermis (Höflich *et al.*, 2004; Aspock *et al.*, 1999; Cao *et al.*, 2017) further confirming the cell-type specificity of the acquired profiles. Similarly, the seam cell but not hypodermis expressed terminal differentiation fusogen *aff-1,* which mediates fusion of the seam cells into a single syncytium at the late L4 stage (Sapir *et al.*, 2007), is significantly occupied and shows enrichment only at the L4 stage as would be expected. Furthermore, *nhr-25* which has been reported to be expressed in both cell-types (Gissendanner & Sluder, 2000), shows signal enrichment and significant occupancy in all profiles. Likewise, *elt-3* a major regulator of the hypodermal fate (Gilleard & Mcghee, 2001) is known to be primarily expressed in the hypodermis in larvae but also shows expression in the seam cells that is 4.5x lower according to L2 sci-RNA-seq data (Cao *et al.*, 2017). Accordingly, signal enrichment and significant RPB-6 occupancy was found for all profiles but signal is qualitatively reduced in the seam cells.

Genes that have been shown with reporters to be hypodermis specific like the osmotic stress factor *osm-7* (Wheeler & Thomas, 2006) and the warthog *wrt-8* (Aspock *et al.*, 1999)*,* were found here to show enrichment and significant occupancy only in the *dpy-7syn1* expression domain. Considering that in
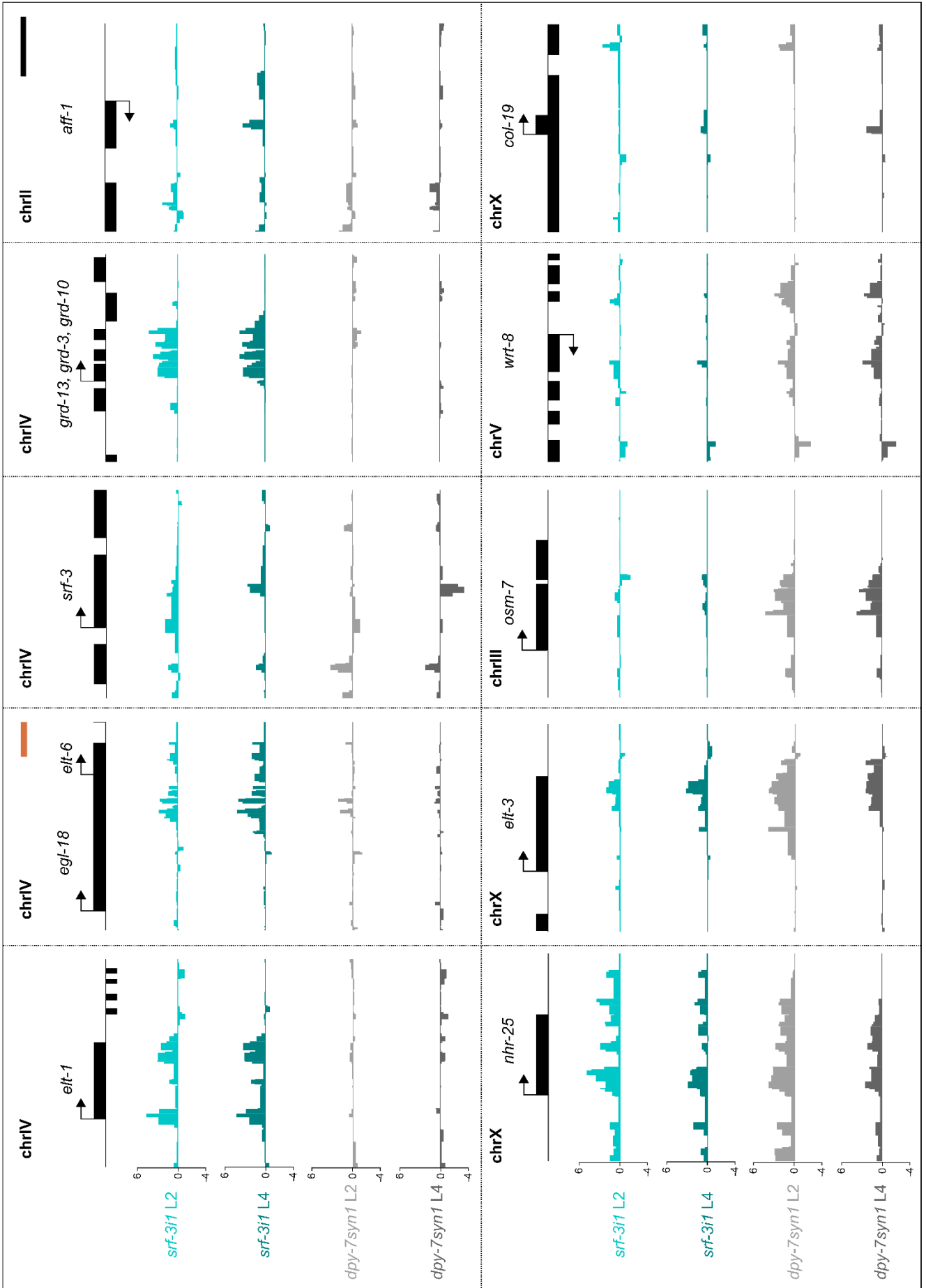
Figure 4. 4 Examples of signal enrichment across genes with known tissue and stage specificities Examples of the signal enrichment profiles over genes showing statistically significant (FDR<0.05) RPB-6 occupancy in samples specified below. The genes *elt-1, egl-18, elt-6, srf-3, grd13, grd-3, grd-10* and *aff-1* show expression, as assessed by RPB-6 occupancy, in the *srf-3i1* but not the *dpy-7syn1* samples, with *aff-1* only at L4, while the genes *osm-7* and *wrt-8* show expression only in *dpy-7syn1* with *col-19* only at L4. *nhr-25* and *elt-3* are found to be expressed in both *srf-3i1* and *dpy-7syn1* domains. The Y-axes represent normalised $\log_2(dam:rpb-6/dam:NLS-GFP)$ scores. Both black and orange scale bars indicate 2 kb and all panels with the exception of the *egl-18/elt-6* panel are described by the black scale bar.

animals expressing the *dam:rpb-6* fusion in the seam cells some amount ends up in the hypodermis due to differentiation events, we could reasonably expect to find hypodermis specific genes showing enrichment in the seam cells as well. *osm-7* based on sci-RNA-seq data falls in the top 20th percentile in terms of expression levels in the hypodermis, amongst genes most highly expressed in that tissue (Cao *et al.*, 2017). The fact that it does not show enrichment in the seam cells despite its likely active high levels of expression could advocate against the extensive contamination of the seam cell profiles for hypodermal genes. Nevertheless, timing of the onset of expression for certain hypodermal genes after differentiation is likely to be a significant contributor to this. Lastly, similar to *aff-1* the L4/adult specific collagen *col-19* which is primarily expressed in the hypodermis but also the seam cells (Liu, Kirch & Ambros, 1995), is found here to show strong signal enrichment in the *dpy-7syn1* L4 profile and is significantly occupied in both *srf-3i1* and *dpy-7syn1* domains only at the L4 stage.

At the genome-wide level RPB-6 occupancy maintains on average a localisation preference for genic sequences rather than intergenic, as is also seen in the specific examples above (Figure 4.4). More specifically, this was observed for the occupancy signal in both expression domains and stages (Figure 4.5A, B), highlighting the biological meaningfulness of the occupancy of all profiles and further supporting that the signal is most likely to reflect active expression. Interestingly, the RPB-6 occupancy showed a preference for increased signal enrichment towards the 3' of genes and a depletion below average signal levels for the 5' of genes, near the TSS, in all profiles (Figure 4.5A, B). This is different to the average occupancy across genes seen by ChIP-seq for the major RNApol II subunit AMA-1, in both L2 and L4 stages (data from(Araya *et al.*, 2014)), which reproduces the increased occupancy of the 3', near the TES, but also shows increased occupancy near the TSS regions (Figure 4.5A, B right). Likewise, RNApol II TaDa in *Drosophila* using the homologue of AMA-1 showed high average occupancy of TSS and TES (Southall *et al.*, 2013). In principle, ChIP-seq is more likely to capture positions where RNApol pauses since it increases the chance of being captured in those positions, whereas TaDa labels all of the transcribed sequence.

Gene coordinates, used to assess average occupancy here, are based on the longest transcript produced by a gene, with up to 94% of those genes however having other isoforms (Tourasse, Millet & Dupuy, 2017). Alternative start and termination sites have been shown to produce most isoforms in humans (Reyes & Huber, 2018), so the likelihood that the 3' enrichment was due to positioning of transcribed isoforms was investigated. Average positional occupancy by isoforms of annotated gene
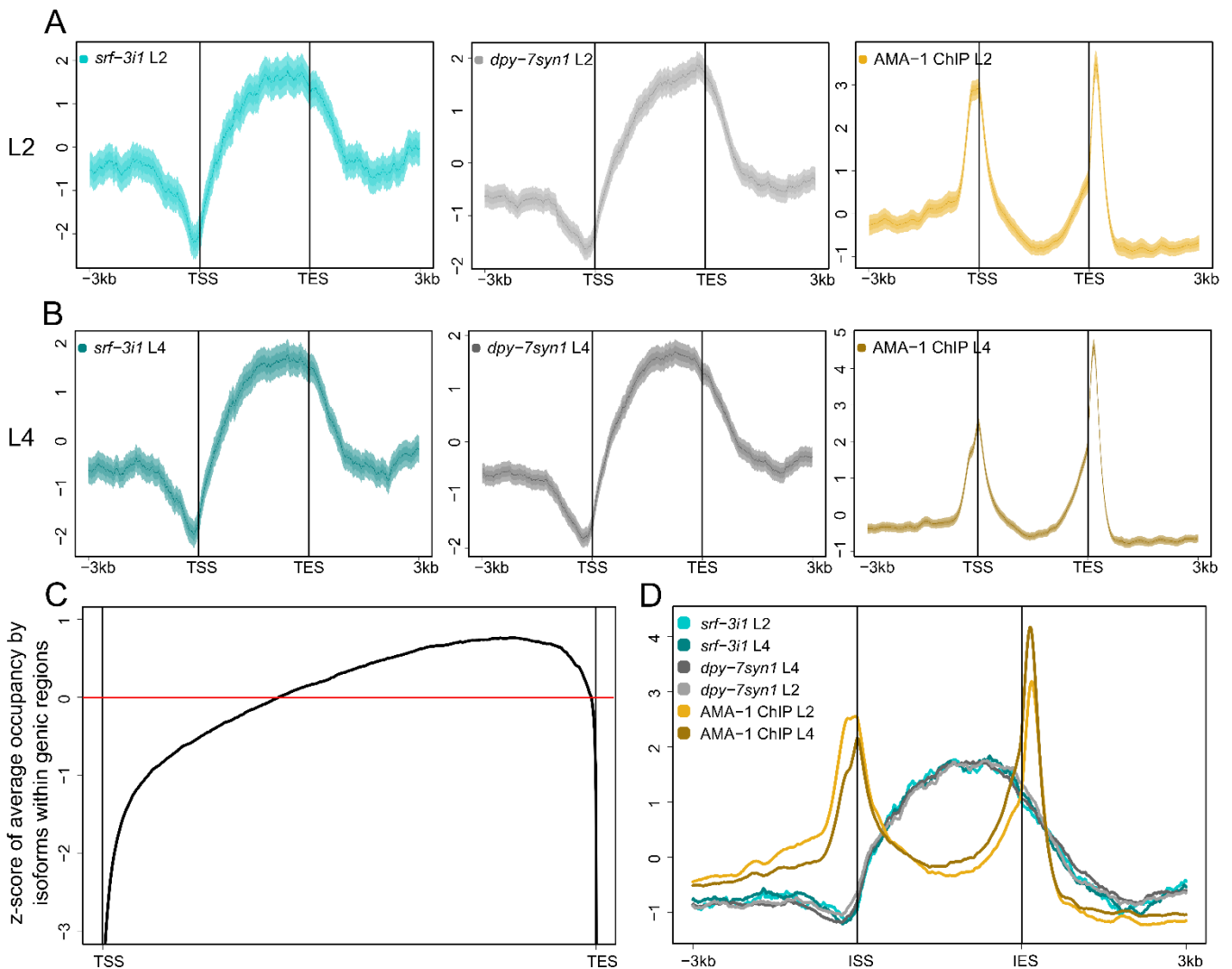
Figure 4. 5 TaDa RPB-6 occupancy is increased across gene bodies with 5' depletion and 3' preference (A-B) Aggregation plots showing average TaDa RPB-6 signal for the *srf-3i1* and *dpy-7syn1* expression domains or whole-animal ChIP-seq AMA-1 signal in 10 bp bins for regions up to 3 kb upstream of the TSS to 3kb downstream of the TES, over all protein coding genes pushed into a pseudo-length of 3kb from L2 (A) and L4 (B) samples. All TaDa RPB-6 samples show increased average signal in genic sequences with preference for 3' regions and depletion near the TSS while the AMA-1 ChIP-seq signal shows peaks of increased average enrichment both over the TSS and at the 3' of genes. **(C)** Deviation from the average occupancy of genic sequences by isoforms shows increased coverage of 3' regions compared to 5' regions of gene sequences annotated based on the largest transcript. Red line indicates the average isoform occupancy across the gene length. **(D)** Aggregation plot for the signal presented in A and B showing average signal anchored at the start and end sites (ISS and IES respectively) of all isoforms of *C. elegans* protein coding genes, pushed into a pseudo-length of 3kb, along with a region 3kb upstream and 3kb downstream. In A, B and D, Y-axes are z-scores for the plotted sequence length and shaded areas in A and B represent 95% confidence intervals.

sequences, showed preference for 3' positions (Figure 4.5C). To examine if this could explain the 3' preference for RPB-6 occupancy within genes, the average occupancy signal was assessed across isoforms instead of genes by anchoring aggregation plots at isoform start and end sites. All TaDa and the ChIP-seq profiles exhibited approximately the same patterns of average signal enrichment, with 3'

preference and 5' depletion for RPB-6 occupancy, indicating that isoform positioning is unlikely to be a contributor.

The 5' depletion could indicate either a transcriptional post-initiation recruitment of RPB-6 in the RNApol II complex or a structural conformation of the initiation complex that prevents the DAM:RPB-6 from producing methylation. The below average signal at TSS and the fact that the yeast homologue Rpb6 has been found to participate in the initiation complex (Ishiguro *et al.*, 2000), points to likely obstruction of DAM:RPB-6 methylation in the initiation complex. DamID-chip experiments for RpII18 (Filion *et al.*, 2010), the Drosophila homologue of RPB-6, lack the resolution to reveal any preference (Appendix C.10).

4.2.5 TaDa-identified expression profiles for seam cell and hypodermis are involved in epidermal tissue-related functions

The protein-coding genes of the genome were assessed for statistically significant (FDR<0.05) RPB-6 occupancy across their length, as previously described (Marshall & Brand, 2015; Southall *et al.*, 2013), to identify expressed genes and produce the final gene expression profiles per tissue and stage. The gene expression profiles are lists of the expressed genes with an average occupancy value for each. For the seam cells, the gene expression profiles of the *srf-3i1* expression domain contained 2227 genes at the L2 and 2446 genes at the L4 stage. The majority of genes for both sets were common between the stages showing a significant overlap (p<1e-320 with a hypergeometric distribution test), with 59.6% of the L2 and 54.3% of the L4 genes being shared by both profiles (Figure 4.6A). The hypodermal lists of expressed genes from the *dpy-7syn1* expression domain contained 2756 genes at the L2 and 2681 genes at the L4 stage. Similarly, the gene sets between stages showed extensive significant overlap (p<1e-320 with a hypergeometric distribution test) with 74.3% of L2 and 76.4% of L4 genes common between both stages (Figure 4.6B). The large overlaps in expressed genes between stages for both expression domains is in agreement with the very high correlation between occupancy profiles seen in section 4.2.3.

To broadly assess the contents of the expression profiles for their biological relevance and association with the cell-type they were found to be expressed within, gene-set enrichment analyses were performed (see Appendix B.13-B.16 for complete results). Interestingly, when tested for enriched gene ontology terms, all gene-sets showed highly significant enrichment for terms pertaining to the synthesis of the cuticle and the molting process, which are key epidermal functions (Page, 2007). The enriched terms were "structural constituent of the cuticle" and "molting cycle", highly relevant as both tissues are known to be involved in timing and execution of molts, as well as to be expressing scores of collagen genes that build the cuticle (Page, 2007; Chisholm & Hsiao, 2012) (Figure 4.6C, D).

The seam cell gene-sets showed significant enrichment for neuronal GO terms like "neuron development" and "regulation of neuron differentiation" that are relevant to this cell-type, as it gives rise to cells that differentiate into neurons or neural precursors (Sulston & Horvitz, 1977; Chisholm & Hsiao, 2012) (Figure 4.6C). The seam cells also give rise to sensory rays of the male mating organ

A

srf-3i1 L2

899

1328

1118

srf-3i1 L4

B

dpy-7syn1 L2

708

2048

633

dpy-7syn1 L4

C

GO terms

neuron development

male anatomical structure morphogenesis

structural constituent of cuticle

molting cycle

regulation of neuron differentiation

male anatomical structure morphogenesis

molting cycle

structural constituent of cuticle

D

reproductive system development

multicellular organism growth

structural constituent of cuticle

molting cycle

reproduction

multicellular organism growth

molting cycle

structural constituent of cuticle

$-\log_{10}q$

E

Tissue enrichment

epithelial system

ABarpppaa (H2 precursor)

ABprapapa (V5 precursor)

PVD

ABarpppaa (H2 precursor)

ABarppapp (V4/V6 precursor)

PVD

epithelial system

F

hyp5

hyp4

hyp6

epithelial system

hyp5

hyp4

hyp6

epithelial system

$-\log_{10}q$

G

Motif in promoters of seam cell expressed genes

Motif in promoters of hypodermis expressed genes

Figure 4. 6 TaDa-identified transcriptomes for seam cells and hypodermis show enrichment for relevant ontologies and tissues (A-B) Venn diagram of sets of genes found to be expressed in the *srf-3i1* domain (A) or the *dpy-7syn1* domain (B) by TaDa, based on significant RPB-6 occupancy in L2 and L4 stages. The majority of the genes for both domains are present in the intersection between the sets for the 2 stages. **(C-D)** Plots of selected significantly enriched gene ontology terms for the *srf-3i1* (C) and *dpy-7syn1* (D) expression domains for the L2 (top) and L4 (bottom) gene sets. **(E-F)** Plots of selected over-represented tissues with expression patterns significantly enriched for similarity to the *srf-3i1* (E) or the *dpy-7syn1* (F) gene-sets at the L2 (top) and L4 (bottom) stage. **(G)** *De novo* identified motifs enriched in the putative promoter sequences up to 2kb upstream of the TSS of the genes in the intersection between L2 and L4 for the *srf-3i1* (left) and *dpy-7syn1* (right) expression domains.

(Sulston, Albertson & Thomson, 1980), which is reflected in the discovered gene-sets, with the term "male anatomical structure morphogenesis", found to be significantly enriched (Figure 4.6C).

Likewise, the gene-sets from the hypodermal *dpy-7syn1* expression domain showed enrichment for the terms "reproduction" and "reproductive system development" (Figure 4.6D). The ventral hyp7 and its precursors participate in the formation of the egg-laying apparatus (Lints & Hall, 2004), a crucial structure of the reproductive system, indicating the relevance of the terms. Moreover, the hypodermis is the major driver of growth in *C. elegans* through DNA endoreduplication in the nuclei of its syncytium (Chisholm & Hsiao, 2012). This is reflected in the *dpy-7syn1* gene-sets of both stages with significant enrichment for the term "multicellular organism growth" (Figure 4.6D).

Similarly, when assessing enrichment for genes with known spatial expression domains by a tissue enrichment analysis, the "epithelial system" was found to be significantly enriched in all gene-sets (Figure 4.6E, F). Other recovered tissue-over-representation terms were very likely to relate to either the *srf-3i1* or *dpy-7syn1*expression profiles. More specifically, in both L2 and L4 *srf-3i1* gene-sets the "PVD" neuron that arises from the V5 seam cell was found to be significantly enriched, along with the seam cell precursor cells "ABarpppaa", "ABprapapa" and " ABarppapp" (distance of one division from the right H2, two divisions from the right V5 and one division from left V4/V6 respectively). In the case of the hypodermis, in both L2 and L4 the anterior hypodermal cells "hyp4", "hyp5", "hyp6" that are highly similar to hyp7 and are within the expression domain of *dpy-7syn1* were significantly enriched. Overall, the enrichment analyses for both GO terms and tissue-similarity, advocate that the acquired expression profiles are likely to be biologically meaningful.

Since the discovered gene-sets were demonstrating signs pointing to expression in the tissues of interest, identification of motifs in their putative promoters was attempted. In more detail, for the genes in the intersections between L2 and L4 stages that are very likely to be truly expressed, the regions from their start site up to 2kb upstream were analysed for enrichment of motifs that could correspond to TF binding-sites. This analysis was performed both for the *srf-3i1* and *dpy-7syn1* expression domain gene-sets (complete results in Appendix C.11, C.12). Interestingly, for the hypodermal gene-sets a TGATAA motif was found to be significantly enriched (Figure 4.6G right). In section 4.2.1 conversion of AGATAA sites to TGATAA on the promoter of *dpy-7* led to exclusively hypodermal expression. In addition, comparison of the motif against available databases showed striking and significant similarity to the binding

motif for ELT-3 (p=5.81e-6), which as previously stated is a major regulator of the hypodermal fate (Gilleard & Mcghee, 2001).

In the case of the intersection gene-set for *srf-3i1*, amongst the enriched motifs was an AGATAG motif (Figure 4.6G left). An AGATAG motif resides in a sequence of the *lin-22* promoter that when deleted (including the previously mentioned AGATAA motif) completely abolished the seam cell expression of *lin-22* (Katsanos *et al.*, 2017). It is of note that an AGATAG motif is also found in the intron 1 of *srf-3* isoform a, which is used here as a seam cell specific enhancer. Comparison of the motif against databases found very significant similarity with that of GATA2 (p=7.44e-4), a human GATA factor with homology to the seam cell fate regulators *elt-1* and *elt-6* (Koh & Rothman, 2001; Smith, McGarr & Gilleard, 2005). The discovery of these motifs on the promoters of the TaDa-identified expressed genes further supports the biological relevance of the acquired expression profiles.

4.2.6 TaDa-determined gene expression lists show extensive overlap across cell types and with previously established datasets

One of the principal aims of this study is to identify genes expressed in the seam cells but not the hypodermis, hoping to uncover factors participating in seam cell fate determination or development. To that direction, multiple intersections across all the identified sets of expressed genes was performed. Intriguingly, the largest subset that formed was by far the overlap between all the gene-sets (1035 genes). Specifically, between 37.5%-46.4% of genes from each set were found in the overlap (Figure 4.7A). Moreover, the vast majority of the genes for all gene sets, between 68.9%-77%, were found in overlaps across expression domains (Figure 4.7A). All these pairwise and higher order overlaps were found to be highly significant for all overlaps (p<1e-320 for all overlaps with a Fisher's exact test) (Figure 4.7B). The substantial sharing of expressed genes across expression domains is in agreement with the high correlation seen between occupancy maps (Figure 4.3C) and is very likely to reflect the common epidermal nature of the two tissues.

As a proxy to asses the above hypothesis the gene-sets were examined for their content of cuticle collagen genes, since the epidermis is secreting the cuticle and regulating the molts. Out of the 173 currently known collagens that participate in cuticle formation (Teuscher *et al.*, 2019), 106 in total were found in the gene-sets. Specifically, 54 and 97 in *srf-3i1* L2 and L4 respectively and 75 and 101 in *dpy-7syn1* L2 and L4 respectively. 50 of those genes (between 49.5% and 92%) were shared amongst all of the sets, indicating the level of similarity driven due to common epidermal functions. From the intersections between the gene sets, a subset of 1090 seam cell-only genes was determined and was used for further identification of seam cell developmental factors.

Having produced evidence that the identified gene sets for the *srf-3i1* and *dpy-7syn1* expression domains, likely correspond to true expression profiles for the seam cells and hypodermis respectively, they were compared against transcriptomes acquired for these cell-types with alternative methods. The

published datasets used were based on tissue-specific RNA-seq. For the PAT-seq datasets tissue-specificity was achieved by mRNA-tagging in the seam cells (*grd-10* promoter) or the hypodermis (*dpy-7* promoter), in mixed stage animals grown in liquid cultures(Blazie *et al.*, 2017), whereas for sci-RNA-seq datasets by combinatorial barcoding and single-cell transcriptome clustering at L2 (Cao *et al.*, 2017). The aim was to explore the comparability of TaDa with more traditional RNA-seq based approaches and to further assess the biological relevance of the identified expression profiles based on across-method reproducibility.

For PAT-seq the threshold for an expressed gene was 1 FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and in sci-RNA-seq 10 TPM (Transcripts per million). All gene-set intersections for the seam cells across methods were highly significant ($p \leq 2.2\text{e-}162$ with a Fisher's exact test) (Figure 4.7C). 72.4% of the TaDa seam cell L2 and 79% of the L4 genes were also found in sci-RNA-seq, showing a compelling agreement between the two methods and underscoring the quality of the TaDa identification. The overlaps with PAT-seq were smaller (33.8% for L2 and 33.9% for L4) but significant. In the case of the hypodermal datasets, similarly all intersections were found to be highly significant ($p<1\text{e-}320$ for all intersections with a Fisher's exact test) (Figure 4.7D). The proportions of the TaDa genes found to be expressed by the other methods, were more similar for the hypodermal datasets, with 76.1% and 79.2% overlapping with sci-RNA-seq and 67.4% and 71.3% overlapping with PAT-seq for L2 and L4 gene sets respectively.

It should be noted that the sci-RNA-seq datasets are significantly larger than the TaDa for both cell-types, which could contribute to the size of the overlaps. The sci-RNA-seq genes not identified by TaDa could either be misattributed to seam cell or hypodermal expression in sci-RNA-seq or have been missed by TaDa. The extent to which each of these contributes to the difference will require further work and the acquisition of TaDa profiles for other tissues as well. The PAT-seq hypodermis dataset is also substantially larger than the TaDa but it was acquired using the unmodified promoter of *dpy-7* which has been shown here to also drive expression in the seam cells. RNA collection for PAT-seq was performed using mixed stage animals, which could also explain the broader dataset.

Both PAT-seq and sci-RNA-seq are generally accepted as quantitative approaches that report expression levels for each of the transcripts/genes they identify. In TaDa, genes with significant RNApol II occupancy are given an average occupancy value for the whole gene. These values could potentially convey information about the levels of expression, as higher frequency of transcription is expected to produce more robust methylation in more individuals, leading to higher signal enrichment. To assess this the correlation between sci-RNA-seq or PAT-seq expression levels and TaDa occupancy scores was examined (Figure 4.7E). The assessment was performed for both seam cells and hypodermis using only L2 datasets, to stage-match the sci-RNA-seq datasets and was based on the common genes between the sets. Interestingly, all comparisons showed statistically significant correlation despite the R^2 values indicating low goodness of fit to the linear model (Pearson's correlation test: $p<0.0001$ for all except the comparison to PAT-seq for the seam cell dataset $p=0.0217$). This advocates towards some amount of
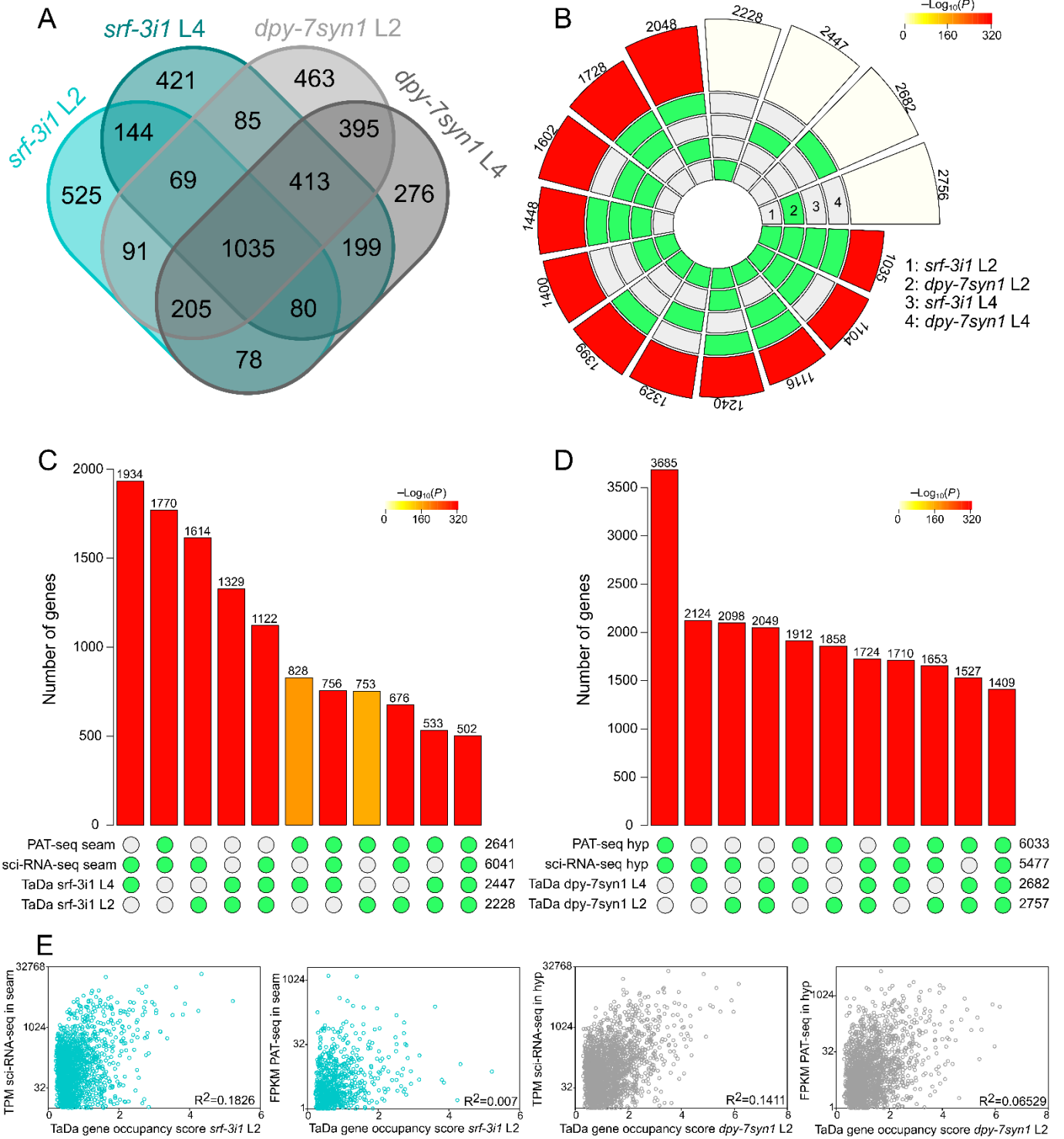
Figure 4. 7 Sets of TaDa-identified expressed genes overlap significantly across cell-types and with published transcriptomes. (A-B) Multiple intersections of all the acquired gene-sets expressed in the *srf-3i1* and *dpy-7syn1* experession domains in both developmental stages. The Venn diagram (A) presents the number of genes that are shared across datasets or are unique to the expression domain and/or stage. The circular plot (B) reports the sizes of all pairwise and higher-order intersections between the sets and indicates that they are highly significant with a Fisher's exact test, highlighting the similarity of the sets. **(C-D)** Barplots of the sizes and statistical significance, assessed by a Fisher's exact test, of all possible intersections between TaDa, sci-RNA-seq (Cao *et al.*, 2017) genes over 10 TPM and PAT-seq (Blazie *et al.*, 2017) identified sets of expressed genes in the seam cells (C) and the hypodermis (D). All overlaps are highly significant. **(E)** Correlation scatterplots of expression levels for L2 genes common between *srf-3i1* (left) or *dpy-7syn1* (right) sets and the sci-RNA-seq or the PAT-seq datasets for seam cells and hypodermis. For TaDa the $log_2(dam:rpb-6/dam:NLS-GFP)$ scores are used as a measure of expression levels, for sci-RNA-seq the values are transcripts per million reads (TPM) and for PAT-seq fragments per kilobase of transcript per million reads (FPKM). All correlation analyses showed significant correlation across methods with $p<0.0001$ for all except the *srf-3i1*/seam cell PAT-seq for which the $p=0.0217$. The R^2 value is indicated.

quantitative information available from TaDa expression profiling, which could be employed to study non-binary differences in expression levels between tissues. It is noteworthy that the PAT-seq and sci-RNA-seq expression levels for their common seam cell expressed genes did not show correlation ($p=0.0529$) even though both methodologies are thought to be quantitative.

4.2.7 TaDa reveals *efl-3* as a seam cell expressed TF that is regulated by LIN-22

From the intersection between expression profiles acquired by TaDa, a set of 1090 genes putatively expressed in the seam cells but not the hypodermis were identified. The seam cell only set could potentially contain seam cell expressed factors that participate in the tissue's development, the fate determination and set the seam cell identity apart from the hypodermal. Based on that hypothesis the seam cell-only gene-set was mined for transcription and chromatin factors, with the reasoning being that they are more likely to participate in fate determination networks or regulate expression that executes seam cell developmental events.

A published dataset for all the *C. elegans* genes predicted to encode TFs, containing 988 genes (Haerty *et al.*, 2008), was manually curated to 907 genes by removing those encoding chromatin factors, based on the annotation of the Ahringer chromatin factor RNAi library (Source Biosciences). Employing the above dataset 58 transcription factors were found in the seam cells-only set (Appendix B.17). Amongst those factors, major known regulators of the seam cell fate were found, like *elt-1, ceh-16, egl-18, elt-6* (Smith, McGarr & Gilleard, 2005; Huang *et al.*, 2009; Gorrepati, Thompson & Eisenmann, 2013; Koh & Rothman, 2001) and the known seam cell expressed *nhr-73* and *nhr-74* (Cassata *et al.*, 2005; Koh & Rothman, 2001; Miyabayashi *et al.*, 1999). All these factors have previously demonstrated seam cell-specificity of expression (Katsanos *et al.*, 2017; Gorrepati, Thompson & Eisenmann, 2013) in larvae, making this an encouraging observation relating to the tissue-specificity of the discovered genes. In addition, it suggested that other such factors with previously unknown roles might be part of that set of TFs. To investigate this further, direct functional confirmation was attempted for members of the 58 TFs,
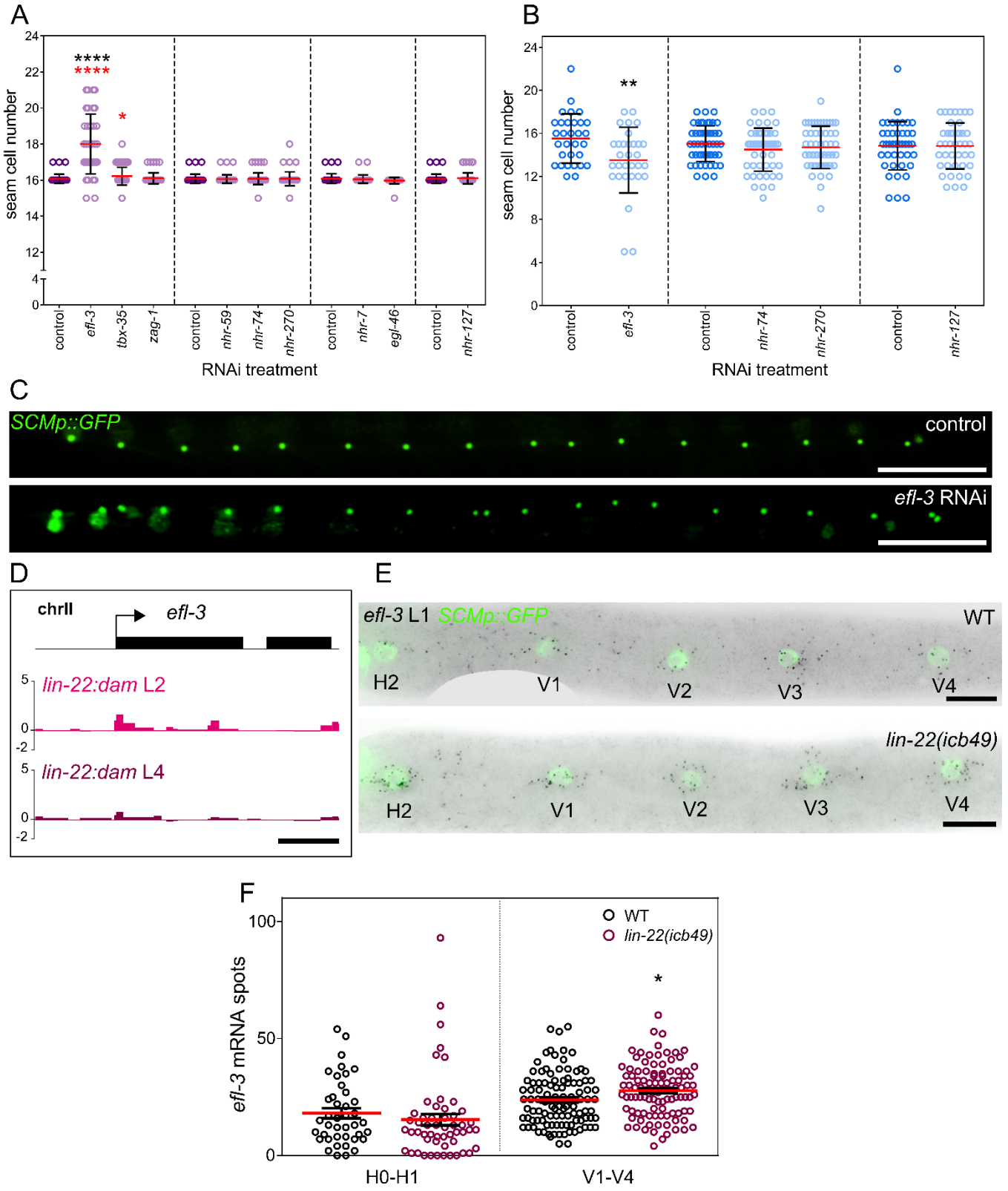
Figure 4. 8 The E2F factor *efl-3* is identified as a novel seam cell fate regulator under the control of LIN-22 (A-B) Quantification of seam cell number at the late L4/Early adult (EA) stage of RNAi treated animals carrying the *SCMp::GFP* reporter in a WT background (34≤n≤59 animals per treatment) (A) or an *elt-1(ku491)* hypomorphic mutant background (31≤n≤50 animals per treatment) (B). RNAi treatments are for the TF genes indicated on the X-axis and sets of treatments performed on the same day are grouped and separated from others by dashed lines, having their respective control. **(C)** Representative fluorescence images of seam cell nuclei marked with *SCMp::GFP* at the L4/EA stage depicting the increase in seam cell number from 16 in the control to 19 in the *efl-3* RNAi treated animal. For both the ventral side is down and anterior to the left. **(D)** LIN-22 TaDa signal enrichment near the TSS of *efl-3* forms a significant peak of putative binding. The Y-axes represent $\log_2$*(lin-22:dam/NLS-GFP:dam)* scores. **(E)** Representative images of *efl-3* smFISH in WT and *lin-22(icb49)* mutants at the late L1 stage, showing *efl-3* expression as black spots corresponding to mRNAs in the seam cells, marked by *SCMp::GFP*. An observable increase in expression in *lin-22(icb49)* is captured. **(F)** Quantification of *efl-3* mRNA spots in WT and *lin-22(icb49)* mutants at late L1 in H0-V4 seam cells, pooled for H (n≥41) and V cells (n≥102), showing a significant increase in expression levels in V1-V4 seam cells. In A, B and F, red line indicates the mean. Error bars are ± Standard Deviation (SD) for A and B and ± Standard Error of the Mean (SEM) in F. Black stars indicate statistically significant differences to the mean with either a one-way ANOVA and a Dunnett's post-hoc or a t-test. Red stars indicate statistically significant differences in variance with a Levene's median test. In both cases * *p*<0.05, ** *p*<0.01, **** *p*<0.0001. Scales bars are 100 μm in C, 2 kb in D and 10 μm in E.

by a small-scale RNAi screen. Based on availability of RNAi clones in our lab and lack previous links to seam cell development, a set of 9 TFs were selected to be tested. The screen was based on terminal seam cell number as the readout, which could capture seam cell developmental defects occurring across development. The rationale was that this approach would not just likely confirm expression in the seam cells, but could potentially also uncover developmentally important factors, expanding the seam cell regulatory network.

The tested factors were *efl-3, tbx-35, zag-1, nhr-59, nhr-74, nhr-270, nhr-7, egl-46* and *nhr-127*. The *nhr*-factors, members of the extensively expanded family of nuclear hormone receptors in *C. elegans* containing 284 members (Antebi, 2006), had not been previously associated with any developmental functions. EGL-46 is a zinc finger TF that has been shown to work with EGL-44 to supress the differentiation of FLP neurons to touch receptor neurons (Wu, Duggan & Chalfie, 2001), while ZAG-1 is a zinc finger/homeodomain transcriptional repressor that is known to act on terminal differentiation of neurons (Wacker *et al.*, 2003). These factors with neuronal development functions could have been identified due to the differentiation of certain seam cell lineages to neurons. The T-Box TF TBX-35 has been previously reported to specify the MS blastomere in the embryo (Broitman-Maduro *et al.*, 2006). Other, factors acting in embryogenesis are known to be re-utilised in postembryonic development for different functions (Chisholm & Hsiao, 2012). Lastly, EFL-3 is an E2F factor that has been shown to supress apoptosis in the ventral cord and has been hypothesised to prevent differentiation of somatic gonad precursor cells allowing them to maintain their multipotency (Winn *et al.*, 2011; Mathies *et al.*, 2019). The screening was performed by Mar Ferrando-Marco a Master's student I supervised.

Interestingly, *efl-3* and *tbx-35* knockdown by RNAi was found to have an effect on seam cell number, while the rest of the factors did not significantly change seam cell number (Figure 4.8A, C). More specifically, the *efl-3* knockdown caused a significant increase of the mean seam cell number from 16.07

in the control to 18 ($p<0.0001$ with a one-way ANOVA). The *tbx-35* treatment did not alter the mean but showed a significant increase in the variance of the seam cell number in the population (control: 16.01 ± 0.25 SD, *tbx-35*: 16.22 ± 0.49 SD, $p<0.05$ with a Levene's median test), with individuals frequently observed having 18 seam cells, an otherwise rare phenotype. Increases in variance could potentially indicate breakdown in the robustness of the seam cell number determination (Boukhibar & Barkoulas, 2016; Katsanos *et al.*, 2017).

For some of the treatments (i.e. *nhr-74, nhr-270, nhr-127*) which did not significantly alter the mean or variance of seam cell number in the population, individuals were observed more frequently with aberrant seam cell numbers that otherwise occur very rarely in wild-type conditions. The seam cell number has been proposed to be a robust phenotype; that is its development can mitigate or withstand perturbations producing an invariant phenotypic outcome (Boukhibar & Barkoulas, 2016; Katsanos *et al.*, 2017). Weaker or subtle phenotypes can be masked, thus in such cases it is common to employ sensitised backgrounds to allow for phenotypes to manifest on the populations average (Conte *et al.*, 2015). Therefore, as an additional, effort to uncover if any of them play a role in seam cells development, a sensitised *elt-1(ku491)* hypomorphic mutant background was treated against *nhr-74, nhr-270* and *nhr-127*, along with *efl-3* as a positive control (Figure 4.8B). None of the treatments except of *elf-3* altered the mutant *elt-1* phenotype. Intriguingly, the *efl-3* treatment had the opposite effect, leading to a significant reduction of the mean seam cell number in this context from 15.5 in the control to 13.5 ($p<0.01$ with a t-test). Considering that the *efl-3* mode of action in the seam cells is not yet known, this phenomenon is difficult to explain. However, effects like this have been previously reported in the seam cells for *pop-1* RNAi in the presence or absence of an *egl-18* mutation (Gorrepati, Thompson & Eisenmann, 2013). Overall, the identification of *efl-3* as a novel seam cell development regulator through this approach is proof of concept and underscores the potential value of the TaDa-identified seam cell genes.

Expanding on this finding, *efl-3* was also found to be amongst the putative targets of LIN-22 from TaDa experiments in chapter 3. Signal enrichment representing LIN-22 binding was seen near the TSS of *efl-3,* particularly in the L2 stage profile (Figure 4.8D), likely corresponding to genuine binding. To assess if *efl-3* is indeed a target of LIN-22, smFISH experiments looking at *efl-3* expression were performed at the late L1 stage, in WT and *lin-22(icb-49)* mutant animals. The smFISH results in WT, clearly demonstrated seam cell-specific expression of *efl-3* at L1, confirming the cell-type specificity predicted by TaDa (Figure 4.8E top). Importantly, the expression in LIN-22 mutants was found to be significantly increased in V1-V4 seam cells ($p<0.05$ with a t-test) (Figure 4.8E, F), suggesting that LIN-22 acts as a repressor of *efl-3* in the seam cells. Thus, through these experiments EFL-3 was both identified as a novel seam cell development regulator and its first link with another seam cell factor was identified.

4.2.8 An RNAi screen against TaDa-identified seam cell-expressed chromatin factors uncovers pleiotropic factors with roles in seam cell development

As introduced in the previous section the set of 1090 seam cell-only genes were also mined for chromatin factors. A previously predicted datasets of all the *C. elegans* chromatin factors includes 167 genes (Haerty *et al.*, 2008). However, for the identification of chromatin factors performed here, the more updated set represented by the Ahringer chromatin factor RNAi library (Source biosciences) containing 257 genes, was used. This led to the identification of 35 chromatin factors identified by TaDa to be expressed in the seam cells but not the hypodermis (Appendix B.18).

Amongst those factors was *bub-1*, a kinetochore binding, mitotic spindle checkpoint factor that regulates cell cycle progression and has been shown to act in multiple embryonic and post-embryonic lineages (Wang *et al.*, 2009). A reduction-of-function allele of *bub-1* shows severe reduction of seam cell number indicating a role for *bub-1* in seam cell development (Wang *et al.*, 2009). Similarly to TFs, a small scale RNAi screen was performed for a subset of those factors based on availability of clones. The readout as before was terminal seam cell number to capture developmental defects that accumulate throughout development.

The tested factors included the histone deacetylases HDA-1 and HDA-2, both of which have been shown to participate in epidermis related developmental events. They are both required for correct sensory ray development, a number of which arise from seam cells (Choy *et al.*, 2007) and HDA-1 in particular drives developmental phenomena like the acquisition of the invasive fate of the anchor cell (Matus *et al.*, 2015). In the embryo it has been shown to be recruited by POP-1 which is an important factor for seam cell development as well (Calvo *et al.*, 2001). Other genes from the screen were the high mobility group factors HMG-11, HMG-1.1 and HMG-4. Of these, only HMG-4 has some previously shown developmental functions relating to anterior pharynx and germline development (Suggs *et al.*, 2018). The chromodomain helicases CHD-1 and CHD-3 were also tested. Of them CHD-3 has been identified as a member of the NurD complex with an important role in supressing ectopic vulval development (Solari & Ahringer, 2000). The member of the SWI/SNF chromatin remodelling complex *swsn-7* was also in the screen and has been shown to be necessary for mitotic progression in embryonal lineages (Krüger *et al.*, 2015). The MYND-type zinc finger factor BRA-2, the bromodomain adjacent to zinc finger BAZ-2 and the GC-rich sequence DNA binding factor F43G9.12, were included in the screen but have not been previously involved in developmental events in *C. elegans*. *bub-1* was also included in the screen as a positive control. The screening was performed by Mar Ferrando-Marco a Master's student I supervised.

The screen was initially attempted with onset of treatment at the L4 stage of the previous generation to the one that is scored, allowing for likely depletion of maternal deposition of the targeted genes. As outlined above a lot of these factors are required for embryogenesis or are known to be required in multiple tissues. This manifested in the screen with treatments for *bub-1, hda-1, hmg-4, swsn-7* and *F43G9.12* exhibiting either severe embryonal lethality, early arrest of larval development or even very small broods

laid from the treated ancestors, precluding scoring of seam cells. To overcome this, treatments for these target genes were also performed postembryonically. For the rest, treatments were performed similarly to TFs, with none of them leading to any significant change in the mean seam cell number or the variance in the population (Figure 4.9A).

The treatments for *hda-2*, *hmg-11*, *baz-2* and *hmg-1.1* showed individuals with aberrant seam cell numbers perceived to occur more frequently than in WT. As in the case of TFs, these treatments were repeated using a sensitised *elt-1(ku491)* hypomorph mutant, to allow possible weak effects to appear (Figure 4.9B). None of the treatments altered the *elt-1(ku491)* phenotype either.

The postembryonic treatments for the genes listed above, all caused a seam cell phenotype with the exception of *swsn-7* (Figure 4.9C). More specifically, *bub-1* which was used as a positive control, showed a significant increase in seam cell number variance in the population (control: 16.11 ± 0.33 SD, *bub-1*: 16.16 ± 0.87 SD, $p<0.001$ with a Levene's median test). This is different to the previously described reduction in seam cell number in mutants producing truncated proteins (Wang *et al.*, 2009). Similar, significant increases in the seam cell number variance without a change in the mean, were also seen in treatments for the factors *F43G9.12* (control: 16.03 ± 0.38 SD, *F43G9.12*: 16.1 ± 0.88 SD, $p<0.01$ with a Levene's median test) and *hmg-4* (control: 16.03 ± 0.38 SD, *hmg-4*: 15.65 ± 1.62 SD, $p<0.01$ with a Levene's median test). The *hda-1* RNAi was the only treatment that significantly altered the mean seam cell number, causing a significant increase from 16.07 in the control to 16.77 ($p<0.0001$ with a t-test). Broadly, these results indicate that a lot of these factors, predicted by TaDa to be expressed in the seam cells, actually likely possess developmental relevant roles within them.

The tendency of these factors to perturb the seam cell number variance without altering the mean, producing individuals with both more and less seam cells than the stereotypical 16, is interesting. It is likely that these chromatin factors have a broader array of genes that they regulate and that could potentially participate in opposing developmental programs in the seam cells. In addition, these opposing effects on seam cell number could also be partly explained by the evident or previously described systemic/pleiotropic effects, which these factors seem to have on the animal's development. This is demonstrated both by the severe effect in embryos but also from other phenotypes that some of these treatments seem to cause. For example, qualitative observations showed severe molting defects for *F43G9.12*, protruding vuvla and multi-vulva phenotype for *hmg-4* and slow movement for *hda-1*. In combination with the embryonic effects and the fact that the phenotyping happens at the end of development, it is conceivable that the seam cell number is perturbed by defects in ancestral lineages or by non-cell autonomous effects, driven by other affected tissues, leading to a more stochastic seam cell outcome.

4.2.9 A novel versatile tool for cell-type-specific RNAi confirms the role of *hda-1* in seam cell development

To disentangle such systemic effects from the seam cell number phenotype and provide evidence of the cell autonomous function within the seam cells, a versatile platform for tissue or cell-type-specific RNAi, based on transgenesis, was built. The approach relies on transgenic expression of an mRNA molecule carrying inverted repeats of a target gene, interrupted by an intron, resulting in folding that creates a hairpin-RNA (hpRNA) structure. These hairpin RNAs are processed by the RNAi pathway and have been shown to cause silencing of target genes in multiple tissues (Tavernarakis *et al.*, 2000; Timmons *et al.*, 2003).

The perks of this approach as opposed to RNAi by feeding is that expression of the hairpin can in principle be targeted spatiotemporally, based on the promoter used to drive expression. For our applications, that is postembryonic expression in the seam cells. However, RNAi is known to be transmissible between tissues in most systems along with *C. elegans* (Jose & Hunter, 2007), which could hinder tissue-specificity. This was thought to require the RNA channel SID-1 (Winston, Molodowitch & Hunter, 2002), which based on sci-RNA-seq data from L2, is very lowly expressed in the seam cells and the hypodermis (14.7 TPM in the seam cells and 8.3 in the hypodermis with a working threshold of 10 TPM), likely suggesting that hpRNAs expressed within them might not be transmitted effectively. Nevertheless, other transmission pathways, independent of SID-1, have been shown to act in uptake and export of RNAi-inducing molecules (Jose, Smith & Hunter, 2009). Previous attempts using hairpin-RNAi have indicated that transmission for hpRNAs in particular, does not always follow the transmission dynamics of other dsRNA and it is highly context and promoter dependent (Timmons *et al.*, 2003; Jose & Hunter, 2007; Briese *et al.*, 2006).

To produce a platform for quick and efficient assembly of such hairpin-RNAi constructs, targeting any gene of interest and in any tissue, a versatile cloning platform was produced and its key features are presented in Figure 4.9D. Cloning is based on single-step Golden Gate (Engler, Kandzia & Marillonnet, 2008), directional assembly of inverted repeats, from a single PCR product. In more detail, the construct carries restriction sites flanking the promoter site, to allow for insertion of any promoter of interest. For applications here, those are the seam cell specific *srf-3i1* and hypodermis specific *dpy-7syn1*. Most importantly, it carries two entry sites (forward and reverse) for the insertion of the inverted repeats, flanking the intron 5 of *srf-3* isoform a, followed by a 3' UTR. The insertion sites contain inverted repeats of recognition sites for the Type IIS restriction enzymes BpiI in the forward entry site and Esp3I in the reverse. These enzymes cut asymmetrically 2 bp to 6 bp away from the recognition sequence creating 5' 4 bp long overhangs. In the design of this system the sequences that are cleaved for both enzymes in both sites create incompatible non-palindromic overhangs, once the recognition sites sequence has been removed. These overhangs are in reverse order in the two entry sites and a fragment with compatible overhangs will
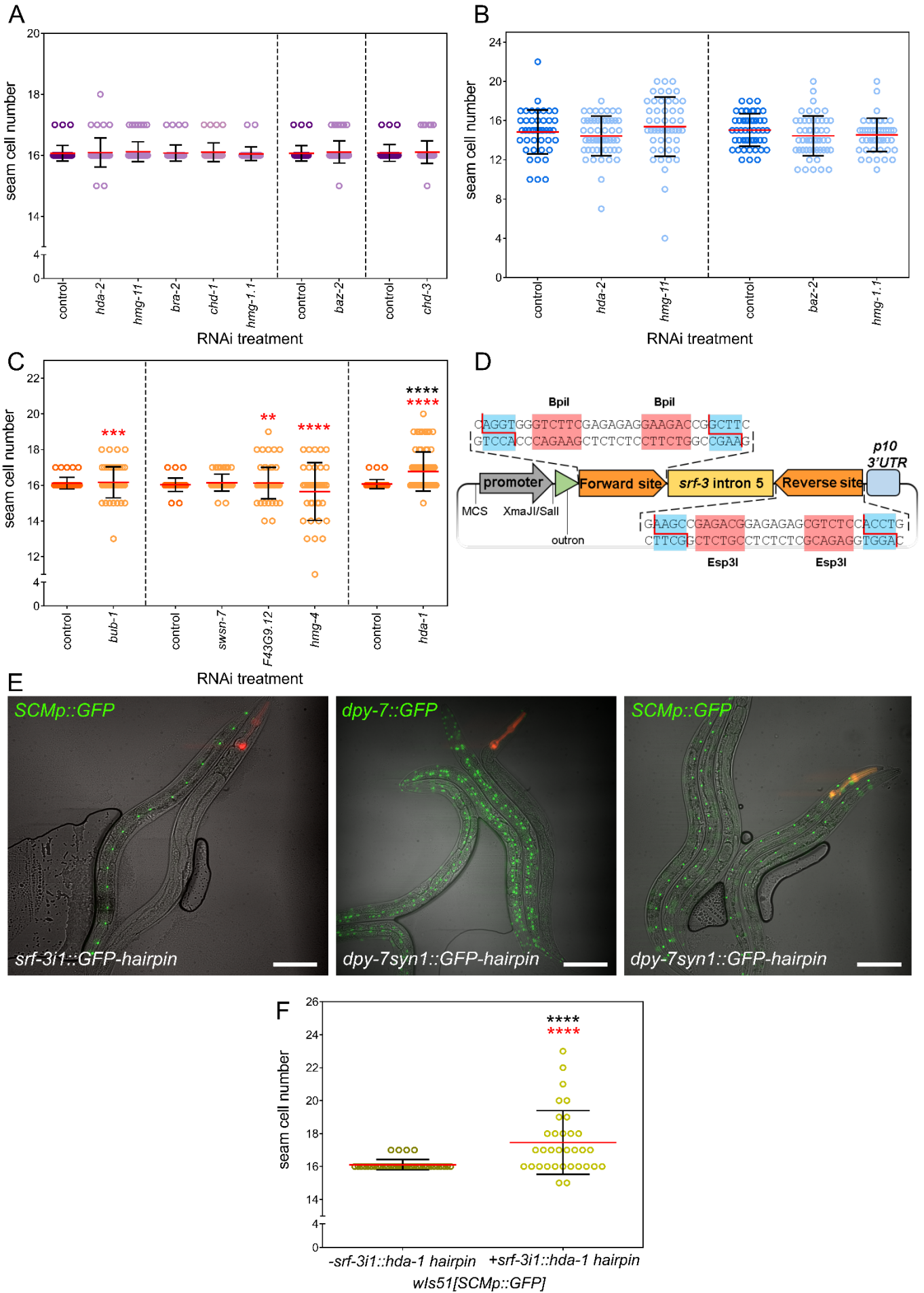
Figure 4. 9 The TaDa-identified seam cell expressed chromatin factors *F43G9.12, hmg-4* and *hda-1* are involved in seam cell development (A) Quantification of seam cell number at the late L4/Early adult (EA) stage of RNAi treated animals carrying the *SCMp::GFP* reporter in a WT background (39≤n≤63 animals per treatment) (A) or an *elt-1(ku491)* hypomorphic mutant background (44≤n≤52 animals per treatment) (B). RNAi treatments are for the chromatin factor genes indicated on the X-axis and the sets of treatments performed on the same day are grouped and separated from others by dashed lines having their respective control. No significant changes to the mean or the variance were observed in both backgrounds for the presented treatments. **(C)** Quantification of the seam cell number phenotype in post-embryonic RNAi treatments for the pleiotropic chromatin factors indicated on the X-axis at the late L4/Early adult stage of *SCMp::GFP* carrying animals of WT background. Treatments are grouped based on scoring day with a respective control. *hda-1* knockdown showed significant increase in the mean seam cell number while *bub-1, F43G9.12* and *hmg-4* showed a significant increase in seam cell number variance in their populations without a change in the mean. **(D)** Illustration of the key features of the versatile golden gate-based platform for transgenic expression of RNA hairpins for tissue or cell-type specific RNAi. From left to right the construct features: a promoter site flanked by restriction sites to allow replacement, an outron to promote expression, a forward insertion site with two tandem inverted BpiI restriction sites (shaded in red) digesting where the red line indicates and two non-palindromic non-complementary sequences at the cut sites (shaded in blue), the 5th intron of *srf-3* isoform a, the reverse insertion site with two tandem inverted restriction sites of Esp3I and the same cut sites as the forward site but flipped and inverted and a *p10* 3' UTR (Pfeiffer, Truman & Rubin, 2012). **(E)** Representative composite fluorescence images of transgenic animals carrying versions of the hairpin RNAi construct to knockdown GFP expression in the seam cells (*srf3i1* promoter) or the hypodermis (*dpy-7syn1* promoter) to assess the system. GFP expression in the seam cells is from the *SCMp:GFP* reporter and in the hypodermis from a *dpy-7p::GFP*. Note that driving the expression of the GFP-hairpin in the same tissue as the GFP expression domain, abolishes the GFP signal but not when GFP is expressed in the seam cells and the hairpin in the hypodermis. **(F)** Quantification of the seam cell number phenotype, by *SCMp:GFP* nuclei scoring, in L4/EA animals either carrying a multi-copy extrachromosomal array transgene expressing an *hda-1* hairpin in the seam cells by *srf-3i1* (n=33) or not carrying it (n=36). Array negative animals are progeny of the same transgenic mothers that have not inherited the array. A significant increase in the mean seam cell number is observed. In A, B, C and F the red line indicates the mean and error bars are ± SD. Black stars indicate statistically significant differences to the mean with a t-test. Red stars indicate statistically significant differences in variance with a Levene's median test. In both cases ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$. Scales bars are 100 μm in E.

be inserted in the two sites in opposite orientations. Such fragments can be produced easily by PCR with oligos that add enzymatic restriction sites for one of the two used enzymes, creating compatible overhangs.

Highly efficient insertion is performed in a one-tube, single step incubation at 37 °C for 30 minutes. The only requirement is that a fragment from the target gene that does not contain a BpiI or Esp3I site is selected for amplification. Overall, the cloning process is completed in 1 day and reactions are assembled with common lab enzymes without requirement for specialised kits. The pWormgate which is an existing platform (Briese *et al.*, 2006), is based on the Gateway cloning system and requires at least two cloning steps for the BP and LR reactions needed for assembly at least doubling the time and increasing the cost.

To confirm that the system is functional and explore the level of RNAi transmission driven by this system in the epidermis, control experiments targeting GFP expression were performed. Seam cell expression of a hairpin against GFP in animals expressing GFP from a multi-copy transgene in the seam cells, led to complete abolishment of GFP signal in 50% of the transgenic animals and strong observable reduction of signal in 36.2% (n=36) (Figure 4.9E left). A similar effect was observed in the case of hypodermal expression for both the hairpin and GFP (Figure 4.9E middle). However, hypodermal

expression of the hairpin and seam cell expression of GFP showed no reduction of seam cell signal (n=13) (Figure 4.9E right). Considering that seam cells fuse to the hypodermis, the opposite combination was not attempted. These findings demonstrate the functionality of the system and provide some evidence supporting the cell-type-specificity of the RNAi effect, at least between these two cell-types of the epidermis.

Using this platform, a seam cell expressed hairpin for *hda-1* was produced and multi-copy transgenic animals were generated. Scoring of the terminal seam cell number was performed by Mar Ferrando-Marco, a Master's student I supervised and revealed that the transgenic animals exhibited a significant increase in the mean seam cell number that was more pronounced than the RNAi by feeding treatment. Specifically, the mean seam cell number increased from 16.1 in non-transgenic animals to 17.5 in transgenics ($p<0.0001$ with a t-test) (Figure 4.9F). In addition, the seam cell specific *hda-1* knockdown in transgenic animals did not exhibit any of the systemic effects seen in RNAi by feeding. All animals were morphologically normal and capable of reproduction, which is different to knocking down *hda-1* using RNAi by feeding. Therefore, these findings likely confirm the cell-autonomous role of *hda-1* in seam cell development.

4.2.10 TaDa gene expression profiling by RPB-6 occupancy reveals epidermal miRNAs with roles in seam cell development

One of the advantages of gene expression profiling based on tracking the occupancy of RPB-6 on DNA is that all types of transcription can be captured. This can for example allow the discovery of expressed miRNAs in the tissue of interest within the same profiles used for assessment of protein-coding gene expression. In RNA-seq based approaches, discovery of small RNA molecules in the transcriptome requires alternative protocols for isolation and library preparation, as well as they lack the quantitative aspect that RNA-seq has for longer molecules (Ozsolak & Milos, 2011; Lu, Meyers & Green, 2007). Using RPB-6 that participates in all RNA complexes has the added advantage of profiling total transcription and can allow discovery, based on occupancy, of all types of small RNAs (miRNA, rRNA, tRNA, siRNA, piRNA, snRNA, snoRNA)(Lu *et al.*, 2005; Ghildiyal & Zamore, 2009) expressed in the tissue of interest. Here, miRNA expression in the seam cells and the hypodermis is investigated base on RPB-6 occupancy.

Out of the 256 annotated miRNAs of the *C. elegans* genome 64 showed expression in the seam cells and/or the hypodermis by TaDa. More specifically, 35 miRNA genes were found to be significantly occupied (FDR<0.05) in the *srf-3i1* expression domain at L2, 39 at L4, 28 in the *dpy-7syn1* expression domain at L2 and 39 at L4. Multiple intersections of those miRNA sets revealed that the majority of them for each set were shared across expression domains, likely indicating housekeeping or broad epidermal functions (Figure 4.10A). An equal number of miRNAs were shared between tissues or were uniquely expressed in seam cells or the hypodermis.

Examination of the intersecting sets found the major heterochronic pathway regulator *let-7* to be expressed only at L4 in both tissues, as it is expected for this miRNA (Slack & Ruvkun, 1997) (Figure 4.10A, B right). Moreover for the miRNA cluster *mir-42*, *mir-43* and *mir-44,* was found to be expressed only in the seam cells in both stages and *mir-47,* was found to be expressed only in the hypodermis in both stages (Figure 4.10A, B left and middle). Available reporters using their putative promoter sequences support the expression domains indicated by TaDa (Martinez *et al.*, 2008). These examples provide evidence for the tissue and temporal specificity of TaDa in identifying likely expressed miRNAs. It is conceivable that identified miRNAs with tissue-specific expression patterns have functions that specifically relate to the tissue's identity.

Potential roles of the above TaDa-discovered tissue-specific miRNAs in regulating fate determination in the epidermis, were investigated. Overexpression constructs were built to drive their expression either within their native epidermal domain, or ectopically, in the tissue they were excluded from. The seam cell number was used as a readout to permit identification of effects relating to seam cell development and scoring was performed by Mar Ferrand-Marco a Master's student that I supervised. The miRNAs *mir-42, mir-43, mir-44* form an operon on their genomic locus on chromosome II (Martinez *et al.*, 2008) and were therefore expressed as a single unit in these experiments.

Interestingly, both the *mir-42, mir-43, mir-44* cluster and *mir-47* produced significant increases in seam cell number variance (control: 16 ± 0.27 SD, *srf-3i1::mir-42;mir-43;mir44*: 16.06 ± 1.06 SD and control: 15.95 ± 0.22 SD, *dpy-7syn1::mir-47*: 15.9 ± 0.68 SD, $p<0.01$ with a Levene's median test for both) only when overexpressed within their endogenous domain (Figure 4.10C). It appears that both regulate targets that are more likely to be relevant and act within the miRNAs' expression domains, thus not producing phenotypes ectopically. These findings constitute an unambiguous proof of concept identification of tissue-specific miRNAs by TaDa, with evidence for functional roles in their expression domains. It also proposes functions related to epidermal development for these miRNA genes that did not have previous functional assignments (Miska *et al.*, 2007).

4.3 Discussion

4.3.1 TaDa as a powerful new tool for tissue-specific gene expression profiling in *C. elegans*

The transcriptional state of a cell largely determines its function, identity and developmental trajectory. Knowledge of the array of genes expressed in a tissue can elucidate how particular fates are acquired, developmental programs executed and differentiation driven. Our model of interest, the *C. elegans* epidermal seam cells, possess a stem cell-like fate capable of producing differentiated cell types, primarily of the hypodermis, while maintaining their identity (Brabin & Woollard, 2012; Chisholm & Hsiao, 2012; Joshi *et al.*, 2010). Identifying, how the batteries of genes utilised in those cell-types differ, can allow

A

srf-3i1 L2

srf-3i1 L4

dpy-7syn1 L2

dpy-7syn1 L4

mir-86
mir-72
mir-230

mir-2213
mir-5548
mir-4921
mir-4923.1
mir-8211
mir-4923.2

mir-239.1

mir-56
mir-1829.2
mir-237
mir-55
mir-5545
mir-61
mir-8201

mir-44
mir-43
mir-42
mir-8192

mir-2211
mir-45

mir-239.1

-

mir-4813
mir-239.2

mir-62
mir-8206
mir-47
mir-8210

mir-5551
mir-356.2
mir-235
mir-2212
mir-356.1
mir-250

mir-252
mir-233
mir-5546
mir-50
mir-38

mir-1829.3
mir-353
mir-66

mir-71
mir-58.1
mir-58.2   mir-65
mir-4937   mir-46
mir-1819   mir-65
mir-64
mir-79

mir-354

mir-229
mir-82
mir-241

let-7
mir-81
mir-1829.1
lin-58

mir-80 mir-38

-

mir-40
mir-41
mir-39

B

chrII

mir-42, mir-43, mir-44

chrX

mir-47

chrX

let-7

srf-3i1 L2

srf-3i1 L4

dpy-7syn1 L2

dpy-7syn1 L4

6
0
-4

C

seam cell number

20

18

16

14

12

4

0

**

*

control

srf-3i1
mir-42; mir-43; mir-44

control

dpy-7syn1
mir-42; mir-43; mir-44

control

srf-3i1::mir-47

control

dpy-7syn1::mir-47

Figure 4. 10 RPB-6 TaDa identifies expressed miRNAs with epidermis developmental functions. (A) Venn diagram listing all the miRNAs found to be expressed by TaDa in the *srf-3i1* and *dpy-7syn1* domains and all the overlaps between the sets found in each domain and stage. In bold, the miRNAs *mir-42, mir-43, mir-44, mir-47* and *let-7* are found in overlaps agreeing with their known spatiotemporal expression patterns. **(B)** TaDa signal profiles of RPB-6 occupancy for miRNAs showing enrichment only in the *srf-3i1* domain for *mir-42, mir-43, mir-44*, only in the *dpy-7syn1* domain for *mir-47* and only at the L4 stage for both domains for *let-7*. The Y-axes represent log$_2$*(dam:rpb-6/dam:NLS-GFP)* scores and the scale bar is 2 kb. **(C)** Quantification of the seam cell number phenotype by *SCMp:GFP* nuclei scoring at the late L4/EA stage, in animals with ectopic expression or overexpression of either *mir-47* or the *mir-42, mir-43, mir-44* cluster in the seam cells, using the *srf-3i1* promoter, or in the hypodermis, using the *dpy-7syn1* promoter. Transgenic animals carried multi-copy extrachromosomal arrays of the transgenes and their respective controls are progeny of the same transgenic mothers that have lost the array. Overexpression within the native expression domain resulted to an increase in seam cell number variance for the *srf-3i1* expressed *mir-42, mir-43, mir-44*, and an increase in seam cell number variance for the *dpy-7syn1* expressed *mir-47* (21≤n≤53 for controls, 31≤n≤47 for transgenics). Red lines indicate the mean and error bars are ± SD. Red stars indicate statistically significant differences in variance with a Levene's median test, * $p<0.05$, ** $p<0.01$.

us to pinpoint those factors crucial for cell fate determination, maintenance and differentiation. In particular genes that are expressed in the seam cells but not the differentiated hypodermis are likely to include such seam cell fate regulators, important for their stem-cell like character or development and could potentially be relevant across systems for stem cell fate specification and differentiation.

In achieving this cell-type-specific identification of expressed genes in the seam cells and hypodermis, current approaches pose limitations. Mechanical or laser dissection of the tissues to separate and sequence the transcriptomes is impractical in *C. elegans* and is exceptionally challenging for tissues that associate with the cuticle (Page, 2007; Schwarz, Kato & Sternberg, 2012). Chemical dissociation of the cuticle followed by isolation of the cell-type of interest, is based on methods like fluorescent activated cell sorting (FACS) that requires fixation, can perturb the transcriptional state of the cells, is dependent on fluorescent markers that can lead to mixed cell populations and requires very large numbers of starting material. The seam cells are only a small fraction of the total number of cells, which added to the difficulty in preparation of dissociated cells suspensions has resulted in minimal recoveries <0.85% in previous applications for the NSM neurons from 3 million larvae (Spencer *et al.*, 2014). The INTACT method (Deal & Henikoff, 2011) circumvents the challenging cell isolation but is technically demanding, requires large amounts of starting material and due to extraction of only nuclear RNA it misses the mature mRNA of the cytoplasm. Not requiring any isolation, the mRNA-tagging method based on transgenically expressed tagged poly(A)-binding protein (PABPC), allows purification of the mRNA from the tissue of interest (Von Stetina *et al.*, 2007). When followed by sequencing it is referred to as PAT-seq and has been used in *C. elegans* to acquire quantitative transcriptomes for various tissues (Blazie *et al.*, 2015, 2017). However, potential poly(A) length biases and reported toxicity of the PABPC in *Drosophila* (Yang, Edenberg & Davis, 2005) are amongst its disadvantages. Lastly, the powerful single-cell combinatorial indexing RNA-seq method has been used to create comprehensive maps of cell-type specific expression in *C. elegans* at the L2 stage (Cao *et al.*, 2017). However, assignment of transcriptomes to tissues is based on existing knowledge, leading to convoluted epidermal gene expression profiles. In addition, despite being cost-

effective for the wealth of information it can provide, it is considerably costly for more focused question, as well as requiring strong bioinformatics capabilities for data processing. Adding to this selection of methods, the first application of TaDa for gene expression profiling by assaying RNA polymerase occupancy is presented here, for the cell-type-specific identification of expressed genes in the seam cells and hypodermis of *C. elegans*.

TaDa does not require cell isolation and is based on transgenic tissue-specific expression of fusions between the protein of interest and Dam. To assists the construction of TaDa transgenes for the experiments presented in this chapter, as well as for all future applications, a versatile cloning platform was built. This universal TaDa vector allows for the cloning of a protein of interest either upstream or downstream of Dam, depending on the optimal configuration for that protein, as Dam has been shown to be robust in both N- or C-terminal fusions (Ramialison *et al.*, 2017). It also contains a docking site for insertion of any promoter by Gateway® cloning, which allows one to take advantage of the gateway design of the *C. elegans* promoterome project (Dupuy *et al.*, 2004), as well as being universal MoSCI-ready.

For the aim of identifying differences between gene expression profiles from the closely related seam cells and hypodermis, the selection of appropriate promoters was crucial. Firstly, promoters had to be highly specific to ensure that genes found in the profiles of one of the two targeted cell-types of the epidermis were not originating from expression occurring elsewhere. Secondly, due to the fusion of differentiating seam cells to hypodermis and the resulting Dam-fusion spread to it, some added artificial overlap was expected. Therefore, to identify unambiguously seam cell specific genes the hypodermal promoter had to be strictly excluded from the seam cells. These criteria were fulfilled here with the *de novo* discovery of a seam cell specific enhancer in the 1st intron of the isoform a of the sugar nucleotide transporter gene *srf-3* (Höflich *et al.*, 2004) and the targeted modification of the promoter of the *dpy-7* gene.

srf-3 had been previously described to be primarily expressed in the seam cells, as one of the strongest expressing genes, but also in other tissues (Höflich *et al.*, 2004; Cao *et al.*, 2017). A survey of the expression driving capacities of its upstream putative promoter sequences showed that the seam cell regulatory capacity was almost entirely exerted from an element in the first intron of the isoform a of the gene. This is a somewhat frequent occurrence in *C. elegans,* where enhancers with regulatory traits that do not overlap the respective promoters, localise in the first introns of genes (Fuxman Bass *et al.*, 2014). Isolation of that enhancer from the surrounding genomic context drove strong seam cell-specific expression and was thus used for the applications here. For the hypodermal expression, the popular promoter of *dpy-7* is shown here to drive expression in the seam cells as well, raising consideration for previous hypodermis transcriptomes acquired by PAT-seq based on its domain (Blazie *et al.*, 2017). Inspection of the promoter sequence highlighted two AGATAA sites previously associated with seam cell expression (Katsanos *et al.*, 2017). Conversion of those sites, on the synthetic version of the promoter *dpy-7syn1,* to TGATAA, the likely binding site for ELT-3*,* a major hypodermal fate regulator (Shao *et al.*, 2013; Gilleard & Mcghee, 2001), abolished the seam cell expression. So far there is no *C. elegans* GATA

factor known to bind the AGATAA motif but it is likely to correspond to one of the seam cell-specific GATA factors like *elt-1, egl-18* or *elt-6.*

Lastly, TaDa gene expression profiling is based on tracking the occupancy of genes by RNApol. Here, the major subunit of RNA pol II, AMA-1 failed to generate efficient methylation in the seam cells and hypodermis, with the sequencing output containing very few unique mappable reads. A similar fusion with the *Drosophila* AMA-1 homologue RpII215 had previously allowed the study of occupancy (Southall *et al.*, 2013). The subunit 6 of RNApol, RPB-6 was used as an alternative producing efficient methylation. It participates in all 3 RNApol complexes (Jones *et al.*, 2000) therefore probing total transcription and its homologue in *Drosophila* has been successfully used in previous DamID experiments (Filion *et al.*, 2010).

4.3.2 RPB-6 occupancy signatures found in the seam cells and hypodermis reveal genes with spatiotemporal resolution

TaDa identification of genome-wide RPB-6 occupancy by next-generation sequencing was found here to be highly efficient. Even for the seam cells that constitute only a small fraction of the total cells of the animal (32/~1000), from a moderately-sized population of as few as ~2000 individuals, an average of 15.5 million unique mappable reads were acquired, beyond previously reported thresholds (Askjaer, Ercan & Meister, 2014; Marshall & Brand, 2015). As a point of reference, for sci-RNA-seq at least 150000 larvae were used and PAT-seq was performed using liquid cultures for increased yield (Blazie *et al.*, 2017; Cao *et al.*, 2017).

Sequence-aligned read count maps for the samples showed very high correlation between the RPB-6 occupancy found in the seam cells and the hypodermis, likely reflecting the common epidermal character of both tissues. In contrast, lower correlation and reproducibility was observed between control samples, which was different to what was presented in chapter 3. Differences in the spatiotemporal aspects of the expression of the transgenes (*wrt-2p* vs *srf-3i* / *dpy-7* promoter) could contribute to the difference in correlation. The most likely explanation however is that in this chapter RPB-6 was fused downstream of Dam, thus to create appropriate controls NLS-GFP was also fused downstream. Dam is relatively robust to steric effects relating to N- or C- terminal fusions, maintaining its capacity to methylate but the configuration of the fusion can affect the fused protein (Ramialison *et al.*, 2017). In this case, the *dam:NLS-GFP* fusion contains an SV40 nucleolocalisation signal peptide (NLS) that as a result is internally positioned. Multiple TFs are known to possess fully-functional internal signals (Boulikas, 1994) but in this synthetic context it is conceivable that nuclear localisation does not occur efficiently, resulting to increased stochasticity in methylation, reducing correlation across samples. Nevertheless, some randomness in methylation is expected from the control fusions as they do not interact with DNA in a targeted manner. Since only a few experiments have so far been performed the source of this remains to be clarified.

Examination of the genome-wide relationship between protein-coding gene sequences and the acquired RPB-6 occupancy signal, revealed apparent preference of signal enrichment within genic

sequences, in comparison to intergenic. This was observed for the signal for both tissues and developmental stages, supporting the hypothesis that it reflects active expression.

Intriguingly the average RPB-6 signal within genes was depleted near the TSS and increased closer to the 3' end, a tendency for RNApol II occupancy that has not been previously observed either by ChIP-seq for AMA-1 or for RNA pol II TaDa in *Drosophila* (Araya *et al.*, 2014; Southall *et al.*, 2013). For both of the above reported cases the average signal peaks near the TSS and again at the TES of genes. The ChIP-seq pattern although informative is not directly comparable since it captures positions that RNApol II is more likely to occupy for longer, thus being more frequently identified creating peaks. In TaDa, methylation of GATC sites should be occurring throughout the sequence of an expressed gene, albeit likely occurring more robustly in locations were polymerase pauses. Pausing has been described to occur primarily on the 5' of genes in metazoans (Gilchrist *et al.*, 2010), therefore not agreeing with the depletion observed at the 5' for RPB-6.

Another potential explanation could be that RPB-6 does not participate in pausing complexes and is recruited to the RNApol II complex only post-initiation of transcription. This is unlikely as homologues of RPB-6 have been shown to promote assembly of the RNApol II complex and be present in the transcriptional initiation complex (Ishiguro *et al.*, 2000; Minakhin *et al.*, 2001). Therefore, a more likely explanation could relate to the specific structural conformation of the initiation complex, that could potentially obstruct DAM:RPB-6 from catalysing methylation of GATCs in the vicinity. As transcription starts, the RNApol II is known to disengage from various components of the initiation complex (Hahn, 2004) thus likely freeing DAM to methylate over the rest of the genic sequence. Lastly, the possibility that this 3' enrichment reflects alternative transcription initiation for isoforms, that on average occupy 3' regions of genes more frequently, was not found to contribute to the phenomenon.

Close inspection of the RPB-6 occupancy signal profiles over selected loci, confirmed that enrichment was occurring within genes, indicating that the occupancy was not random but likely reflected active transcription. In addition, the loci/genes were selected based on their known spatiotemporal expression characteristics as a means to primarily assess the tissue-specificity of the acquired profiles for the seam cells and the hypodermis. Accordingly, signal enrichment was observed only in the seam cell profiles within the known seam cell expressed genes *elt-1, egl-18, elt-6, srf-3, grd-10, grd-3, grd-13* (Katsanos *et al.*, 2017; Gorrepati, Thompson & Eisenmann, 2013; Koh & Rothman, 2001; Aspock *et al.*, 1999; Höflich *et al.*, 2004)*.* smFISH data for both *elt-1* and *egl-18* support the TaDa-identified exclusivity for seam cell expression (Katsanos *et al.*, 2017). Similarly, signal enrichment over the known hypodermis specific *osm-7* and *wrt-8* (Wheeler & Thomas, 2006; Aspock *et al.*, 1999) was observed only in the hypodermal profiles. Commonly expressed genes like *nhr-25* showed enrichment in both. Additionally, genes with expression onset at L4 like *aff-1* (Sapir *et al.*, 2007) for the seam cells and *col-19* (Liu, Kirch & Ambros, 1995) primarily for the hypodermis showed enrichment only in the respective L4 profiles. Such examples highlighted the achieved cell-type-specificity and overall quality and biological relevance of the acquired profiles.

4.3.3 TaDa-identified sets of expressed genes are relevant to their cell-type of origin and comparable to datasets from alternative methodologies

Genes that showed significant occupancy of RPB-6 were deemed to be expressed within the cell-type and at the stage they were detected in. The resulting lists of genes were the gene expression profiles that this chapter set out to identify. Encouragingly, these sets of genes showed large significant overlaps between the stages, indicating reproducible detection of certain genes that are reasonably expected to be expressed across stages. In addition, a majority of genes for all of the sets were found to be shared across cell-types. This was not surprising as both the seam cells and the hypodermis are cell-types of the epidermis. They both participate in very central functions of the epidermis, like synthesis and secretion of cuticular constituents and timing and execution of molting across larval development (Chisholm & Hsiao, 2012; Page, 2007). For example there are 173 known collagens that participate in cuticle formation (Teuscher *et al.*, 2019), 106 of their genes are found amongst the 4 profiles, with 50 of those shared across all of them. This is an indication of the similarity that is potentially driven by their common epidermal character and is reflected in their gene expression profiles. This was further supported by gene ontology and tissue enrichment analysis that identified the epithelial system, that the epidermis belongs to, as one of the most similar tissues in terms of expression, as well as the GO terms relating to molting and structural constituents of cuticle, which were enriched in all gene-sets.

Nevertheless, regardless of the high degrees of similarity, the discovered gene-sets also showed convincing evidence of cell-type-specificity. The seam cell gene-sets were enriched for neurogenesis and male mating organ morphogenesis-related ontology terms, which reflect known developmental functions of the seam cells (Sulston, Albertson & Thomson, 1980; Sulston & Horvitz, 1977; Chisholm & Hsiao, 2012). In addition, precursors of seam cells and the PVD neuron that arises from the V5 seam cell, were found to be amongst the significantly related tissues (Sulston & Horvitz, 1977). Likewise, organism growth and reproduction related terms, corresponding to functions carried out by the hypodermis (Chisholm & Hsiao, 2012; Lints & Hall, 2004), were found for hypodermal gene-sets. Amongst the significantly enriched tissues were cells of the anterior hypodermis.

Interestingly, in the promoters of the highly likely hypodermal genes, common between the two stages, a TGATAA motif, strikingly similar to that bound by ELT-3 (Shao *et al.*, 2013), was found to be significantly enriched. ELT-3 is a major regulator of the hypodermal fate and is required to drive differentiation to hypodermis (Gilleard & Mcghee, 2001). Therefore it is likely to control a plethora of hypodermal genes, further supporting the cell-type-specificity of the findings. In the case of the seam cells, similar motif discovery identified an enriched AGATAG motif, likely binding site for an unknown GATA factor. Comparisons with available databases showed significant similarity with the human homologue of *elt-1* or *elt-6*, both important seam cell fate regulators. AGATAG sites exist in the seam cell specific enhancers of the first intron of *srf-3* and of the distal upstream sequence of *lin-22* (Katsanos *et al.*, 2017).

It is thus likely that this motif drives seam cell expression and its identification advocates to the cell-type-specificity of the TaDa data presented here.

Different methods achieve tissue-specificity of expression profiling using different approaches. So far transcriptomes for the seam cells and hypodermis have been identified by PAT-seq (Blazie *et al.*, 2017) for mixed stage populations and sci-RNA-seq only at the L2 stage (Cao *et al.*, 2017). The TaDa identified datasets showed very significant overlaps with the available datasets for both seam cells and hypodermis. These overlaps were very extensive, in particular with the sci-RNAseq, representing ≥72.4% of the total genes identified by TaDa. This is a strong indication that the majority of genes found by TaDa are truly expressed within the cell-types where they were detected, since alternative methods identify them.

The number of total genes identified by TaDa is however smaller than most of the other datasets. In the case of PAT-seq the sampling happened in a mixed population, therefore genes that might not be expressed at the stages interrogated in TaDa are likely to contribute to some proportion of the difference. It is the hypodermal set of PAT-seq that is particularly inflated in comparison to the TaDa set. Aside of the mixed-stage aspect, this could be explained by the use of the *dpy-7* promoter to perform the mRNA-tagging in that experiment. As shown in the present study *dpy-7* drives expression in the seam cells as well, thus likely contaminating the hypodermal set with seam cell expressed genes. This is supported by the overlap between PAT-seq and sci-RNA-seq for the hypodermis that included only 61% of the PAT-seq genes.

In the case of sci-RNA-seq, the determination of the origin of a single cell transcriptome is based on expression of genes used us markers and can therefore be biased. In addition, clustering of the single-cell transcriptomes specifically for the epidermis did not create separate clusters for the seam cells and hypodermis, as it did for other tissues (Cao *et al.*, 2017). Therefore, a somewhat increased representation of genes not truly expressed by both is possible. TaDa-related weaknesses could undoubtedly also be the source of the size difference between sets. *rnt-1,* which is shown in chapter 3 by smFISH evidence to be seam cell-specific but very lowly expressed (on average 3.01 transcripts per cell), is not detected by TaDa. This could suggest lower sensitivity for low expressing genes in TaDa. However, *rnt-1* was not detected by PAT-seq either and was borderline above the working threshold in sci-RNA-seq (10.3 TPM). However, TaDa identified genes with similarly low expression levels as indicated by sci-RNA-seq (322 out of 1324 genes between 10 and 20 TPM), albeit with a smaller relative overlap in comparison to the overlap of the complete sets. This indicates that likely, lowly expressed genes can be identified by TaDa and are not broadly missed. Another contributor could be the availability of GATC sites across a gene. *lin-22* is expressed specifically in the seam cells but was not detected by TaDa in that tissue. Examination of its 2.5 kb genic sequence revealed that it only contains 2 GATC sites 2 kb apart. This is likely to severely impact detection capability since it necessitates the methylation of both sites for any occupancy to be recorded. The existence of more such occasions could partially explain differences in set sizes between methods, however it is not expected to be a pervasive phenomenon since the average length of GATC fragments within *C. elegans* protein-coding genes is 518 bp.

As a final poimt it is of note that common genes between TaDa and PAT-seq or sci-RNA-seq showed correlation between RPB-6 occupancy and expression levels as those are defined by each method. This is an encouraging observation highlighting that expression level information can be extracted from TaDa, such that expression profiles could be treated semi-quantitatively.

4.3.4 Functional confirmation for seam cell expressed transcription factors, chromatin factors and miRNAs revealed novel regulators of seam cell development

One of the main pursuits in acquiring the seam cell and hypodermis specific expression profiles, was the identification of genes expressed in the seam cells but not the hypodermis. These could include factors that specify the seam cell identity and allow them to maintain their stem-cell-like fate, while generating differentiated tissues. To achieve that, the identified subset of 1090 genes expressed in the seam cells but not the hypodermis was mined for transcription and chromatin factor genes, with the rationale being that they are more likely to participate and regulate developmental and fate specification events. 58 transcription factors and 35 chromatin factors were found based on comparisons with available datasets (Haerty *et al.*, 2008). These sets of genes already contained factors with known seam cell specific expression like *nhr-73* and *nhr-74* (Cassata *et al.*, 2005; Koh & Rothman, 2001; Miyabayashi *et al.*, 1999), as well as crucial for seam cell development regulators like the TFs *elt-1, egl-18, elt-6, ceh-16* (Brabin, Appleford & Woollard, 2011; Gorrepati, Thompson & Eisenmann, 2013; Koh & Rothman, 2001; Huang *et al.*, 2009; Cassata *et al.*, 2005) or the chromatin factor *bub-1* (Wang *et al.*, 2009).

Functional identification of seam cell development regulators within those sets was performed by a small-scale RNAi screen based on phenotyping for aberrant terminal seam cell number. This functional confirmation approach partly encompasses the confirmation of expression domain, which has been used before for validation of TaDa findings (Southall *et al.*, 2013). Those factors found to be implicated in seam cell development through the confirmation experiments above are also most likely truly expressed within the tissue as they were detected by TaDa. The RNAi screen identified two TFs and three chromatin factors that were for the first time linked to seam cell development, as their silencing led to aberrant seam cell numbers.

The transcription factors were *efl-3* and *tbx-35*. *tbx-35* knockdown caused a mild seam cell number variance phenotype. It encodes a T-box factor that is required for the specification of the MS blastomere fate in the embryo (Broitman-Maduro *et al.*, 2006). It is not found to be expressed in the seam cells by PAT-seq (Blazie *et al.*, 2017) or sci-RNA-seq (Cao *et al.*, 2017). The RNAi treatment that generated the phenotype was initiated at the L4 stage of the ancestors of the phenotyped animals. Taking the embryonal role into account it is possible that the seam cell phenotype might be a side-effect of abnormal embryogenesis, resulting from *tbx-35* knockdown in the embryo by maternal deposition. Postembryonic treatment or seam cell specific RNAi would allow us to better dissect the source of the phenotype.

In contrast, the *efl-3* treatment caused a substantial increase of the average seam cell number. It was also detected in the seam cells by sci-RNA-seq but not PAT-seq. EFL-3 is an E2F factor that was first described in *C. elegans* for its role in supressing cell-death in the VA and VB cells of the ventral cord (Winn *et al.*, 2011). It has also been found by mRNA profiling to be expressed in the multipotent somatic gonad precursors but not in their differentiated sisters, the head mesodermal cells and was hypothesised to suppress differentiation in that system (Mathies *et al.*, 2019). Here, it is confirmed by smFISH to be expressed in the seam cells, likely exerting its function cell-autonomously.

Its human homologue, E2F7, is an atypical E2F that acts most likely as a transcriptional repressor, to regulate and antagonise other E2F genes and cell cycle components, overall having a role as a negative regulator of cell proliferation (Di Stefano, Jensen & Helin, 2003; Lammens *et al.*, 2009). In achieving that role, it has also been shown to repress miRNAs involved in the cell cycle (Mitxelena *et al.*, 2016). Interestingly, E2F7 has been linked to the prevention of endoreduplication in mammalian cells (Lammens *et al.*, 2009). Endoreduplication takes place during differentiation of the seam cell daughters that will fuse to the hypodermis and has been linked to the acquisition of the hypodermal fate (Chisholm & Hsiao, 2012). However, loss of *efl-3* here leads to expansion of the non-endoreduplicated seam cell population. Further research in humans has implicated E2F7 with diverse functions in controlling cell proliferation, differentiation and apoptosis and has been proposed as a putative tumour suppressor gene, with its absence leading to poor prognosis (Di Stefano, Jensen & Helin, 2003; Lammens *et al.*, 2009; Endo-Munoz *et al.*, 2009). In the *C. elegans* seam cells context, the observed seam cell hyperplasia could result from failed differentiation of hypodermal-destined daughters, or excess symmetric divisions. More work will be required to determine the underlying mechanism. Nevertheless, in this study *efl-3* is shown to be a target for repression by LIN-22 in the seam cells creating its first link with a member of the seam cell regulatory network.

In the case of the chromatin factors, RNAi knockdown of *hmg-4, F43G9.12* and *hda-1* were the treatments that were shown for the first time to perturb seam cell number. All of these factors showed expression in the seam cells in sci-RNA-seq but not PAT-seq datasets (Blazie *et al.*, 2017; Cao *et al.*, 2017).

HMG-4 is a member of the histone chaperone FACT (facilitates chromatin transcription) complex, homologue of the human SSRP1, and has been shown to be expressed in multiple somatic tissues (Suggs *et al.*, 2018; Kolundzic *et al.*, 2018). The FACT complex has been shown to act as a barrier to alteration of cell fate both in *C. elegans* and mammalian systems (Kolundzic *et al.*, 2018). In *C. elegans* it has been mostly studied in the context of regulating cell cycle timing in the embryo, where HMG-4 acts redundantly with its paralog HMG-3 and in the intestine where it shows evidence of maintenance of inaccessible chromatin states that prevent activation of genes that can drive fate reprograming (Suggs *et al.*, 2018; Kolundzic *et al.*, 2018). Based on its functions in other somatic tissues, in the seam cells HMG-4 as a member of the FACT complex could conceivably oversee the differentiation decisions between seam cells and hypodermis, which are misregulated in its absence, leading to the observed phenotype.

Likewise, the *F43G9.12* RNAi treatment was found to cause a significant seam cell number variance phenotype. This factor has not been studied in *C. elegans* before. It is a homologue of the mammalian PAXBP1 (PAX3 and PAX7 binding protein 1). It has been shown in mice to act by binding the paired-box TFs Pax3 or Pax7 and recruit at the site the histone 3 lysine 4 (H3K4) methyltransferase (Diao *et al.*, 2012). H3K4 monomethylation or trimethylation are considered marks of active enhancers and promoters respectively, leading to chromatin state changes that can permit or promote gene expression (Heintzman *et al.*, 2007). In mice, Pax3/7BP (the *M. musculus* orthologue) has been shown to be required for the proliferation of myoblasts by leading to the activation via the above epigenetic mechanism of, amongst others, cell cycle genes (Diao *et al.*, 2012). This could be proposing a mode of action in the seam cells where its knockdown leads to the observed phenotype. It is notable that there is no known *C. elegans* homologue for PAX7, while the PAX3 homologue of *C. elegans pax-3,* is not expressed in the seam cells (Cao *et al.*, 2017; Thompson *et al.*, 2016). Importantly, it has been shown to be a critical regulator of the ventral hypodermis fate and its absence causes those cells to convert to a seam-like fate, thus proposing a seam cell fate suppressor role for *pax-3* (Thompson *et al.*, 2016). It is therefore an open question what recruits F43G9.12 to specific chromatin sites in the seam cells. Work in mice has suggested that it is also likely that it can link the histone methyltransferase to TFs other than Pax3 or Pax7 (Diao *et al.*, 2012), opening the way for other involved TFs in the seam cells.

Amongst chromatin factors, the *hda-1* RNAi treatment was the only one that led to a uni-directional shift towards higher seam cell numbers. It was seen to cause systemic abnormalities and to validate that its effect on seam cell number was cell-autonomous, it was also knocked-down with relative specificity in the seam cells by hairpin-RNAi, resulting to a similar, more pronounced phenotype. *hda-1* encodes for a histone deacetylase homologue of the human HDAC1 and HDAC2. In *C. elegans, hda-1* has been shown to instruct correct transcriptional programmes during embryogenesis, based on cell-type by associating with POP-1 (Calvo *et al.*, 2001). POP-1 is the *C. elegans* TCF, the convergence point for the crucial for seam cell development Wnt/β-catenin asymmetry pathway, which largely determines seam cell fate during asymmetric divisions (Sawa & Korswagen, 2013). Moreover, HDA-1 has been specifically shown to be required for the development of the male sensory rays (Choy *et al.*, 2007) that arise from seam cells, already suggesting a function within our tissue of interest. It also participates in vulval development, by promoting the invasive fate of the anchor cell through regulation of genes that drive the invasive behaviour and invadopodia formation (Matus *et al.*, 2015). It was hypothesised to act in parallel or as a response to a G1 arrest of cell cycle required for transition to the invasive fate (Matus *et al.*, 2015). If the acquisition of the invasive fate could be considered the differentiation event in this context, then HDA-1 could be similarly promoting differentiation in the seam cells which fails in its absence causing an increase in the seam cell number.

In mammalian embryonic stem cells the homologues of HDA-1, HDAC1 and HDAC2, are essential for cell proliferation at least partly by regulating G1 arrest, required for viable proliferation of the cells, indicating some mechanistic conservation across systems (Jamaladdin *et al.*, 2014). Additionally, loss of

their (HDAC1 and HDAC2) activity has been shown to cause downregulation of pluripotency factors (Jamaladdin *et al.*, 2014). Reduction of *hda-1* expression in the seam cells increases their numbers, therefore a mechanism based on the downregulation of pluripotency factors in the seam would be contradictory to the notion of seam cells being the stem-like cells in the epidermis. Further work on the roles of the mammalian *hda-1* homologues during hematopoiesis has shown that they possess diverse roles including stemness maintenance, cell proliferation, differentiation and fate determination that are context and co-factor dependent (Wang, Wang & Liu, 2020). Interactions have been found of HDAC1 or HDAC2 in complexes with homologues of major seam cell regulators like LIN-22 (Hes1) and RNT-1 (Runx1) (Wang, Wang & Liu, 2020). Consequently, understanding the mode of action in the seam cell of HDA-1 will require further work pursuing these directions.

TaDa gene expression profiling is based on marking transcribed genomic loci, which can reveal expression of small RNAs without the need for protocol alteration. In this study, a subunit of RNA polymerase that participates in all three complexes was used, assaying for total transcription and in principle permiting identification of all types of expressed small RNAs. This advantage was utilised here to identify a miRNA cluster containing *mir-42, mir-43* and *mir-44* that is found to be specifically expressed in the seam cells as an operon and *mir-47* found to be expressed in the hypodermis. The expression domains indicated by TaDa, were in agreement with published reporters for the promoters of these miRNAs (Martinez *et al.*, 2008).

Most importantly, these miRNAs were shown here to perturb seam cell numbers when overexpressed in their native spatial expression domain but not when expressed ectopically in the epidermis. A possible role in epidermal development is the first function that any of these miRNAs have been associated with (Miska *et al.*, 2007). Tracing of miRNAs expression across development using northern blots has indicated that *mir-42* and *mir-43* are predominantly embryonic, whereas *mir-44* expression persists in larvae (Lau *et al.*, 2001). Considering that they are expressed from an operon active in larvae, *mir-42* and *mir-43* might be subject to post-transcriptional regulation (Martinez *et al.*, 2008). The overexpression here could be overcoming these regulation mechanisms allowing for *mir-42* and *mir-43* to act in the larval context, possibly causing the phenotype. Overexpression of each miRNA of the cluster individually is certainly one of the first steps to pursue, in identifying the culprit amongst them.

In the case of *mir-47,* it is particularly interesting how expression in the hypodermis, but not the seam cells, causes a seam cell phenotype. This non-cell autonomous effect could be based on misregulation of miRNA targets that produce developmental signals from the hypodermis to the seam cells. *In silico* prediction of targets for these miRNAs (Lewis, Burge & Bartel, 2005; Jan *et al.*, 2011) identified amongst others the major seam cell fate regulator *elt-1* (Smith, McGarr & Gilleard, 2005) for *mir-43* and the *frizzled* receptor *lin-17*, which participates in seam cell-specifying Wnt signalling (Katsanos *et al.*, 2017), for *mir-47*. These avenues will have to be explored to identify the exact roles for these miRNAs in epidermal development.

Chapter 5

Identification of accessible chromatin in the *C. elegans* seam cells and hypodermis using Chromatin Accessibility targeted DamID (CATaDa)

5.1 Introduction

The complexity observed in multicellular organisms can be attributed to the plethora of expression programmes of specific arrays of genes, with their temporal and quantitative characteristics, which largely determine cell and tissue identity. Development is the process by which these tissue-specifying expression profiles are established by the selective decoding of the universal genomic information. These epigenomic instructions are often in the form of repressed or active chromatin states that determine which genes can be expressed. *cis*-regulatory elements, like promoters or enhancers, can activate expression of their associated genes in the presence of the right factors (Gaudet & McGhee, 2010; Tsompana & Buck, 2014). To recruit transcription factors and drive expression, their sequences need to reside in regions of accessible, low complexity chromatin (Spitz & Furlong, 2012; Gaudet & McGhee, 2010). Thus, chromatin accessibility is a crucial level of gene expression regulation for the establishment of tissue-specific expression patterns and can indicate genomic locations of active promoter or enhancer sequences (Tsompana & Buck, 2014).

Assessing chromatin accessibility can reveal such regulatory elements *en masse* and provide insights on the epigenomic regulation of transcriptional states relevant to cell-fate and differentiation. Such questions are relevant to our model of interest. The stem cell-like *C. elegans* seam cells maintain their "stemness" and cell identity throughout post-embryonic development, while producing differentiated, mostly hypodermal, daughters. Assessing the chromatin accessibility profiles of the seam cells and hypodermis could indicate if and how this aspect of the epigenome transitions from a precursor stem cell state to a terminally differentiated cell-fate state, mediating the differentiation that occurs. It could also reveal epidermis-specific regulatory elements.

"Open" or accessible chromatin is devoid of nucleosomes, freeing up those DNA sequences for regulatory interactions to take place. This feature has been utilised by most techniques aiming to profile chromatin accessibility. Traditionally, open chromatin was characterised by its increased sensitivity to nuclease digestion (Gross & Garrard, 1988). Coupled with next-generation sequencing technologies, DNase hypersensitivity sites can been mapped across the genome, indicating loci of increased accessibility (Song & Crawford, 2010). Applications in *C. elegans* have revealed sites of enhancer and promoter sequences but only on the whole-animal level, with mapped sites representing open chromatin from all tissues (Ho, Quintero-Cadena & Sternberg, 2017; Jänes *et al.*, 2018). On the same premise the FAIRE-seq approach is based on chemical cross-linking and shearing of DNA by sonication to isolate the non-nucleosome bound sequences (Giresi *et al.*, 2007). FAIRE-seq-discovered open sites coincide with DNase hypersensitivity sites but this method has not seen any applications in *C. elegans*. ATAC-seq is the simplest, fastest and most sensitive method in comparison to the above (Tsompana & Buck, 2014; Buenrostro *et al.*, 2013). It is based on *in vitro* tagmentation of open chromatin with sequencing adaptors by the Tn5 Transposase (Buenrostro *et al.*, 2013). It has been used in *C. elegans* to identify whole-animal chromatin accessibility changes and dynamics across development and aging and to identify regulatory

elements (Daugherty *et al.*, 2017; Jänes *et al.*, 2018). In the study by Jänes *et al.*, 2018 the accessible chromatin sites were also extensively annotated for the type of regulatory element they harbour.

Despite the undeniable wealth of information produced in the above studies, the crucial dimension of tissue-specificity is absent and the acquired chromatin accessibility profiles are an amalgamation of all the open chromatin found in the animal. To be able to pursue questions, like those in our interests, about seam cell to hypodermis differentiation, epigenomic regulation of seam cell fate and to discover epidermis specific *cis*-regulatory elements, a cell-type-specific approach is essential.

To achieve this using the above methods, cell isolation is required, which as previously described is particularly laborious and challenging especially for the *C. elegans* epidermis. Overcoming this limitation, the chromatin accessibility targeted DamID (CATaDa) method is based on transgenic expression of Dam that is not specifically recruited to any locus of the genome and therefore will interact and methylate accessible chromatin DNA more frequently, than DNA bound and protected by nucleosomes (Aughey *et al.*, 2018). Notably, with CATaDa marking of the accessible chromatin occurs *in vivo* in contrast to the *in vitro* processing which is essential for the above methods and can introduce artefacts. As another member of the TaDa method "family", CATaDa utilises the same protocols and reagents and chromatin accessibility profiles can be acquired from control Dam-fusion sequencing samples from other applications.

In this chapter, cell-type-specific assessment of genome-wide chromatin accessibility is performed for the first time in *C. elegans,* with the first application of CATaDa in this model organism. Chromatin accessibility profiles were acquired for the seam cells and hypodermis at the L2 and L4 stage. Overlaps of the discovered sites between profiles and their preference for specific genomic locations were assessed. To examine the degree to which they reflect sites with regulatory capacity, their association with marks of active chromatin states, transcription factor binding and conservation was investigated. Moreover, their agreement with existing profiles from alternative methods was explored, along with the association of sites to nearby genes that are expressed in the interrogated cell-types. Lastly, a selection of intergenic accessible chromatin sites were tested for epidermis specific enhancers by assessing their ability to drive expression in transgenic animals. Overall, this chapter demonstrates the feasibility of employing CATaDa for cell-type-specific assessment of chromatin accessibility in *C. elegans* and provides the first such maps for the seam cells and hypodermis, along with a selection of novel confirmed epidermis-specific *cis*-regulatory elements.

5.2 Results

5.2.1 Accessible chromatin sites in the seam cells and hypodermis by CATaDa show preference for TSS and non-coding regions

The same sequencing results of the control samples from the TaDa experiment of chapter 4 were reanalysed here to look into chromatin accessibility. These DAM:NLS-GFP fusions have been shown to

reflect chromatin accessibility which is the premise of CATaDa (Aughey *et al.*, 2018). They are therefore used here to acquire cell-type-specific chromatin accessibility profiles for the seam cells, based on the *srf-3i1* expression domain, and for the hypodermis, based on the *dpy-7syn1* promoter, at the L2 and L4 stage.

Principal component analysis of the sequence alignment read-count maps indicated similarity across all samples, which did not cluster separately along any of the relevant sample variables, like expression domain or developmental stage (Figure 5.1A). In addition, biological replicates particularly for the *dpy-7syn1* expression domain were positioned apart on the PCA space, which is consistent with the moderate correlation between samples, as previously presented in chapter 4. The *srf-3i1* samples showed a higher tendency for grouped mapping of their read-count maps on the PCA but the overall lack of separation can have multiple potential explanations. Given that the read-count maps are expected to reflect methylation based on chromatin accessibility across two related cell-types it is conceivable that they are broadly similar with most of the variation explained by other parameters. Such an example could be the potentially more stochastic methylation by the *dam:NLS-GFP* fusions which is not targeted but occurs serendipitously more often in open chromatin regions. As discussed in chapter 4 these stochasticites could be exacerbated by discrepancies in nucleolocalisation efficiency.

The pipeline described in Aughey *et al.*, 2018 was used to convert the sequencing results to signal tracks representing chromatin accessibility profiles across the genome for each replicate, in each expression domain, at each developmental stage (complete genome-wide signal tracks for each replicates are available in Appendix C.13, C.14). The example presented in figure 5.1B illustrates the CATaDa signal enrichment across chromosome I, with peaks expected to represent sites of increased chromatin accessibility. It is worth noting that qualitatively the peaking observed across all profiles shows high degree of similarity as hypothesised. Throughout this chapter, signal or peaks/sites expected to reflect chromatin accessibility are often abbreviated to CA using the promoter to specify the expression domain.

In relationship to protein-coding genes the signal for all expression domains and stages was found to show on average preference for areas around the transcriptional start sites (Figure 5.1C). This preference has been seen previously by CATaDa in *Drosophila* and suggests that the signal is likely to represent chromatin accessibility regions because the TSS are sites most often surrounded by open chromatin (Ho, Quintero-Cadena & Sternberg, 2017; Aughey *et al.*, 2018). Some profiles also showed a tendency for above average signal occurring more frequently in upstream to the TSS regions rather than within the gene or downstream (Figure 5.1C). This is likely to reflect the generally upstream to genes positioning of promoter and enhancer elements, which are known to associate with accessible chromatin when they are active (Chen *et al.*, 2013; Klemm, Shipony & Greenleaf, 2019). Biological replicates showed almost entirely overlapping average signal across genes highlighting their similarity and thus potential biological meaningfulness.

As in Aughey *et al.*, 2018, peak-calling for the identification of regions with statistically significant signal enrichment (FDR<0.05) was performed independently for each replicate. The resulting peak profiles were then merged by averaging overlapping peaks to produce a single profile for each expression domain

at each developmental stage. These selections of peaks are expected to represent the accessible chromatin regions of the genome found by CATaDa in the respective cell-types. Specifically, 2398 chromatin accessibility peaks/sites were found in the *srf-3i1* seam cell domain at L2 and 2410 at L4, while in the *dpy-7syn1* hypodermis domain 2299 were identified at L2 and 2303 at L4. In total 4449 unique accessible chromatin sites were identified by CATaDa for the seam cells and the hypodermis combined.

The positional layout of the CATaDa accessible sites across the genome, in relationship to protein coding genes, was examined (Figure 5.1D) and showed localisation patterns strongly resembling those previously reported for accessible chromatin sites identified by whole-animal DNase-seq (Ho, Quintero-Cadena & Sternberg, 2017). Since TaDa peaks are in general broad, due to their dependence on GATC fragment size, the centre was used as a reference, based on which identification of positioning within the genome was carried out. In more detail, between 24.3% and 35.7% of peaks had their centres in exons, which often occurs in active genes, similar to observations in human cell lines (Mercer *et al.*, 2013). Between 12.4% and 18.3% were overlapping start sites and between 22.7% and 42.2% were fully intergenic. These numbers agree with the DNase-seq observations in L1arrest of 13% for promoters up to 300 bp from the start site and 28% for intergenic localisation (Ho, Quintero-Cadena & Sternberg, 2017). Between 7% and 9.5% of CATaDa peaks were also found to overlap the end site and between 13.2% and 17.5% were found to have their centres in introns. Overall, CATaDa sites showed preference for non-coding regions as previously described (Ho, Quintero-Cadena & Sternberg, 2017), where regulatory elements that associate with open chromatin are more likely to be found (Klemm, Shipony & Greenleaf, 2019; Spitz & Furlong, 2012).

5.2.2 Chromatin accessibility profiles acquired for the seam cells and hypodermis show extensive similarity

The signal profiles of chromatin accesibility acquired for the seam cell and hypodermis showed qualitatively similar patterns of enrichment across chromosomes. Closer inspection indicated a strong overall agreement (Figure 5.2A) for many of the observed genomic loci. At the genome-wide level, all pairwise tests of aggregated peak signal for a profile, over the accessible sites of another, showed definitive increase of the averaged signal for the position of the CATaDa sites, in comparison to adjacent regions (Figure 5.2B).This illustrated that accessible sites found for each expression domain and stage are very likely to show extensive overlaps across the genome.

To further investigate this, all the pairwise intersections of genomic positions for the identified CATaDa sites between profiles, were performed (Figure 5.2C). This analysis confirmed that the majority of identified CATaDa sites for all tissues and stages overlap with each other. Between 51% and 65% of the total accessible chromatin peaks for each profile occupied the same genomic loci as those of another. All the pairwise intersections were found to be highly significant and non-random by Monte Carlo simulations (p<1e-320). In addition, a total of 933 accessible chromatin sites were identified by all profiles,

A

PC2 (17.2% of var. explained)

PC1 (68.6% of var. explained)

- *srf-3i1-NLS-GFP* L2 rep 1
- *srf-3i1-NLS-GFP* L2 rep 2
- *srf-3i1-NLS-GFP* L4 rep 1
- *srf-3i1-NLS-GFP* L4 rep 2
- *dpy-7syn1-NLS-GFP* L2 rep 1
- *dpy-7syn1-NLS-GFP* L2 rep 2
- *dpy-7syn1-NLS-GFP* L4 rep 1
- *dpy-7syn1-NLS-GFP* L4 rep 2

B   chrI

reads per million

C

| *srf-3i1* CA L2 rep 2 / *srf-3i1* CA L2 rep 1 | *srf-3i1* CA L4 rep 1 / *srf-3i1* CA L4 rep 2 | *dpy-7syn1* CA L2 rep 1 / *dpy-7syn1* CA L2 rep 2 | *dpy-7syn1* CA L4 rep 1 / *dpy-7syn1* CA L4 rep 2 |

D

srf-3i1 L2
2398

srf-3i1 L4
2410

dpy-7syn1 L2
2290

dpy-7syn1 L4
2303

	srf-3i1 L2	*srf-3i1* L4	*dpy-7syn1* L2	*dpy-7syn1* L4
Overlapping TSS	14.3%	14.4%	12.4%	18.3%
Overlapping TES	9.0%	9.0%	7.0%	9.5%
Gene inside peak	0.8%	1.2%	0.8%	1.1%
Peak in intergenic region	24.2%	22.7%	42.2%	27.9%
Centre of peak in intron	16.6%	17.0%	13.2%	17.5%
Centre of peak in exon	35.0%	35.7%	24.3%	25.6%

Overlaping TSS   Overlaping TES   Gene inside peak   Peak in intergenic region   Centre of peak in intron   Centre of peak in exon

Figure 5. 1 CATaDa found increased chromatin accessibility around TSSs with preference for non-coding regions (A) Principal component analysis of the normalised read-count maps, from the sequencing results of both replicates of the *srf-3i1* and *dpy-7syn1* expressed *dam:NLS-GFP* fusions at L2 and L4, did not show distinct groupings of samples based on tissue or developmental stage. Some replicates showed higher variability between them than with samples from other tissues or stages. **(B)** Examples of the CATaDa (CA) reads per million-nomalised signal across chromosome I for all promoters, developmental stages and replicates showing somewhat reproducible enrichment patters. Signal tracks are in the order listed in the key of A, Y-axes are normalised reads per million values and the scale bar is 1 Mb. **(C)** Aggregation plots showing average CATaDa signal across a regions extending 5 kb upstream of the TSS to 5 kb downstream of the TES of all protein-coding genes, which are fitted to a pseudo-length of 5 kb. Average signal for all promoters and stages shows strong preference for the TSS region and proximal upstream sequences. Replicates plotted as different lines show substantial agreement. Y-axes are z-scores for the plotted sequence length and shaded areas represent 95% confidence intervals. **(D)** Pie charts of the proportions of the total statistically significant CATaDa peaks (FDR<0.05), indicated above each chart, that localise at different positions in relationship to protein-coding genes, for each expression domain and stage indicated by the aggregation plots above. The breakdown of the genomic locations are listed below the charts.

constituting between 38.7% and 40.6% of the total peaks of each profile. These commonly identified accessible chromatin sites across compared datasets and in particular across tissues, could be corresponding to universal accessible chromatin locations. However, the extent of the overlap could be also potentially attributed to the common epidermal identity of seam cells and hypodermis, which is reflected to a certain extent by similar chromatin accessibility patterns across the genome. Acquisition of more CATaDa profiles for other tissues would be required to further assess this similarity.

5.2.3 CATaDa sites associate with marks related to active regulatory elements

Accessible chromatin across the genome is well established to reflect sites of permissible interactions between regulatory elements, like promoters or enhancers and various types of factors that in concert regulate gene expression (Klemm, Shipony & Greenleaf, 2019). Chromatin accessibility assays are widely used to identify sites of such regulatory elements (Tsompana & Buck, 2014; Gaudet & McGhee, 2010; Ho, Quintero-Cadena & Sternberg, 2017) and similarly here the discovery of epidermis specific regulatory elements is one of the main objectives.

A preliminary qualitative assessment of the potential of CATaDa signal to reflect such elements, showed significant CATaDa signal enrichment at the sites of certain previously described enhancers. Specific examples include an element proximally upstream to *hlh-11,* that was identified based on its high occupancy target (HOT) status and was shown to drive expression amongst other tissues in the hypodermis (Chen *et al.*, 2014) (Figure 5.3A left). Significant enrichment was also seen for all profiles at the site of known *cis-regulatory* elements that were identified in the intergenic region between the Hox genes *ceh-13* and *lin-39*, based on multigenome conservation comparisons (Kuntz *et al.*, 2008) (Figure 5.3A middle). Interestingly, at the site of an ATAC-seq identified enhancer, upstream of the gene *bed-3*, that was reported to drive expression only in the hypodermis (Jänes *et al.*, 2018), CATaDa showed significant enrichment only in the *dpy-7syn1* profiles (Figure 5.3A right). This underscores the cell-type-
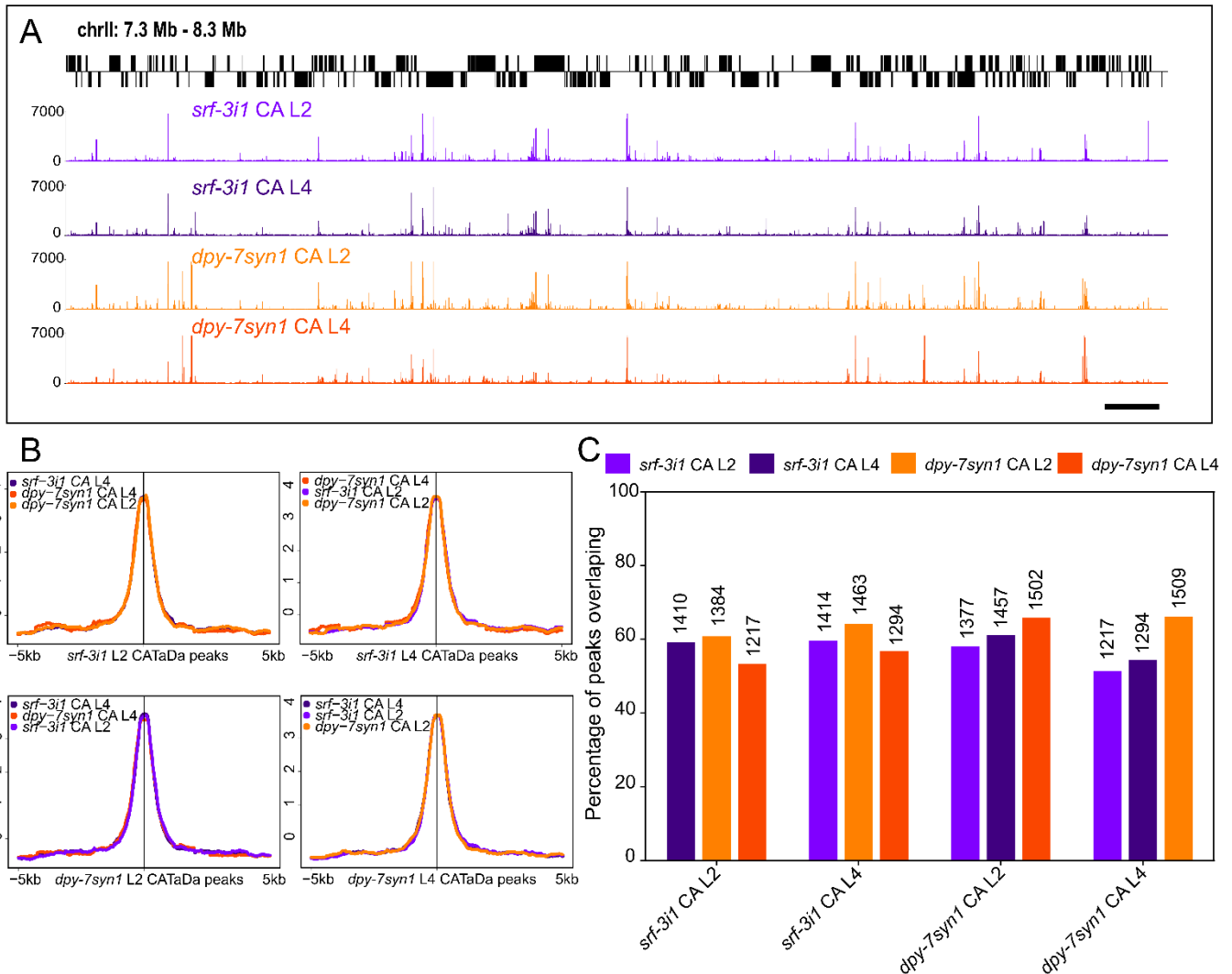
Figure 5. 2 CATaDa-identified chromatin accessibility peaks show significant overlaps across cell-types and developmental stages. (A) Qualitative example of CATaDa signal profiles for all expression domains and stages across 1 Mb of chromosome II, showing strong agreement between all profiles with multiple peaks reproduced in all of them. One replicate is shown for each promoter and stage for clarity. The scale bar is 50 kb. Y-axes are normalised reads per million. **(B)** Aggregation plots of average CATaDa peak signal for each expression domain and developmental stage profile, over a region ± 5 kb from the centres of all peaks of the sample indicated on the X-axis, show strong preference for the sites of peaks in comparison to adjacent regions for all pairwise comparisons. Y-axes are z-scores for the plotted sequence length **(C)** Barplot of the proportion of peaks from each sample, indicated by colour, that overlap the peaks of the sample indicated on the X-axis. The exact number of overlapping peaks is printed on top of the bar. The majority of peaks in all pairwise comparisons overlap.

specificity of the chromatin accessibility identification performed here. These examples are based on confirmed enhancers identified by diverse methods and constitute promising indications for the potential of regulatory element identification based on CATaDa.

To assess the general association of CATaDa accessible chromatin sites with potentially regulatory regions across the genome, their juxtaposition with known marks of regulatory activity was examined. Various histone modifications are responsible for transitions between alternative chromatin states in terms of compactness and nucleosome occupation of DNA, exerting a regulatory control over gene expression

(Bannister & Kouzarides, 2011). Different histone modifications are known to associate with active open chromatin as opposed to closed repressed chromatin or heterochromatin and are primarily based on methylation or acetylation of different histone lysine residues (Black, Van Rechem & Whetstine, 2012; Bannister & Kouzarides, 2011). Trimethylation of lysine 4 of histone H3 (H3K4me3) has been associated with active chromatin, often representing promoters and to some extent enhancers. In contrast, trimethylation of lysine 27 of histone H3 (H3K27me3) has been associated with repressed chromatin regions, often including closed promoters or enhancers (Bannister & Kouzarides, 2011; Black, Van Rechem & Whetstine, 2012; Heintzman *et al.*, 2007).

Interestingly, CATaDa sites for all expression domains were found to show preference for genomic regions with H3K4me3 marks (from L3 animals: modEncode_3576) in comparison to adjacent sequences and were neutral against stage-matched regions of H3K27me3 (Jänes *et al.*, 2018) (Figure 5.3B). This is in line with the active open chromatin state at the H3K4me3 sites and illustrates that CATaDa broadly captured open chromatin.

To expand on this observation, the relationship between CATaDa sites and different predicted chromatin state loci was investigated. Specifically, available ChromHMM-predicted chromatin states for the L3 stage (Daugherty *et al.*, 2017), were utilised to assess the global localisation of CATaDa sites in relationship to those chromatin states. ChromHMM, a hidden Markov-model chromatin state prediction algorithm (Ernst & Kellis, 2012), categorised chromatin based on a selection of histone modification marks into: heterochromatin, H3K27me3 repressed, repressed enhancer, active enhancer, TSS/promoter and transcribed gene body chromatin states (Daugherty *et al.*, 2017). This analysis revealed very significant enrichment for CATaDa sites, from all cell-types and developmental stages, in active chromatin regions like active enhancers, TSS/ promoters and transcribed gene bodies ($p<6.9e-75$, $p<1.65e-72$, $p<4.99e-15$ respectively by Monte Carlo simulations) (Figure 5.3C). In addition, heterochromatin and H3K27me3 repressed regions were significantly depleted for CATaDa sites ($p<2.47e-23$, $p<3.92e-41$ respectively by Monte Carlo simulations). Repressed enhancers showed some enrichment only for *dpy-7syn1* CATaDa sites. Overall, these findings definitively indicate that the CATaDa-identified sites largely reflect genuine accessible chromatin regions, agreeing with genome-wide chromatin state predictions for active regulatory functions based on histone modification marks.

This general propensity for CATaDa sites to correspond to sequences with putative regulatory function was further demonstrated by the significant overlaps they showed with other markers of regulatory functions. More specifically, between 9.9% and 12.3% of CATaDa sites for both cell-types and developmental stages co-localised with stage-matched high occupancy target (HOT) sites (Araya *et al.*, 2014) ($p\leq1.8e-70$ for L2 profiles, $p\leq1.3e-56$ for L4 profiles by Monte Carlo simulations) and between 14.4% and 37.6% with occupancy sites by ChIP-seq for AMA-1 the *C. elegans* major RNA pol II subunit ($p\leq1.6e-55$ for L2 profiles, $p\leq1.1e-97$ for L4 profiles by Monte Carlo simulations) (Figure 5.3D). The proportions of overlapping peaks are similar if not higher than previously reported for accessible chromatin by DNase-seq (Ho, Quintero-Cadena & Sternberg, 2017). Such significant overlaps with regions highly bound by
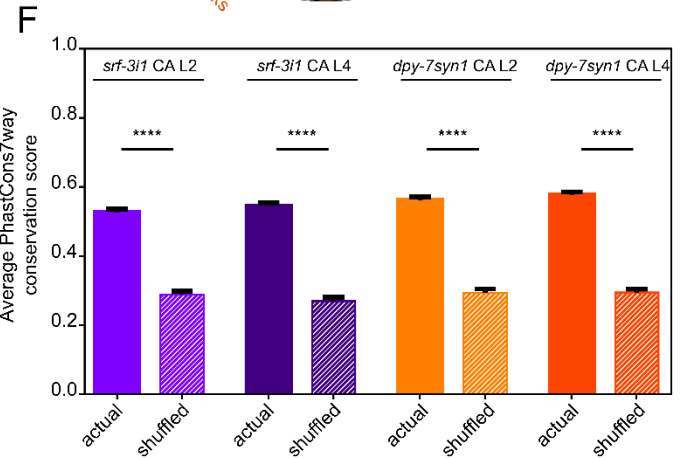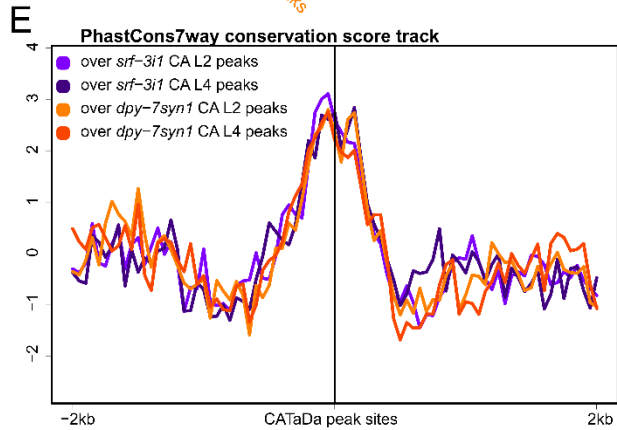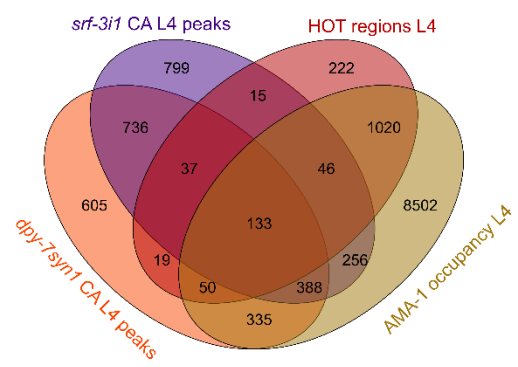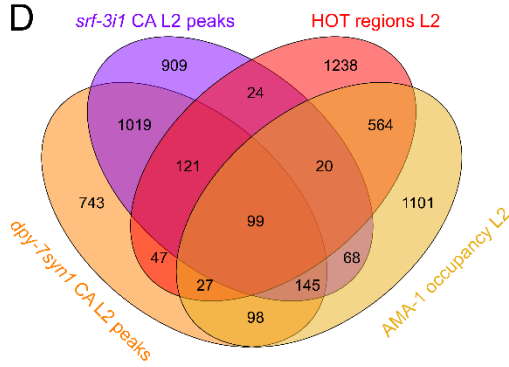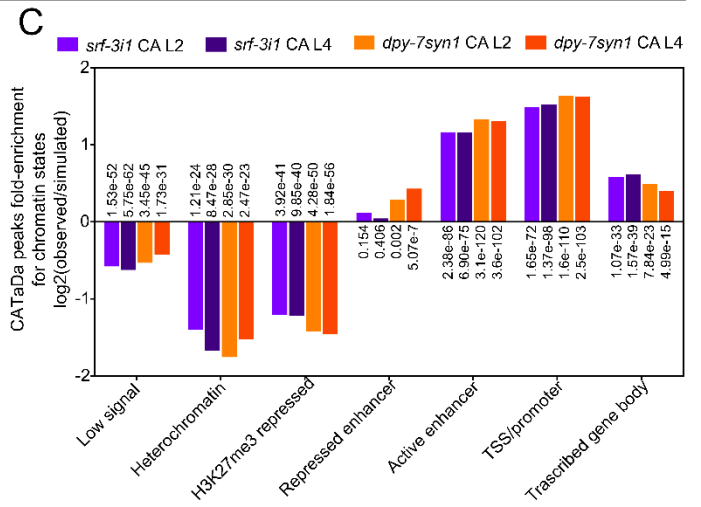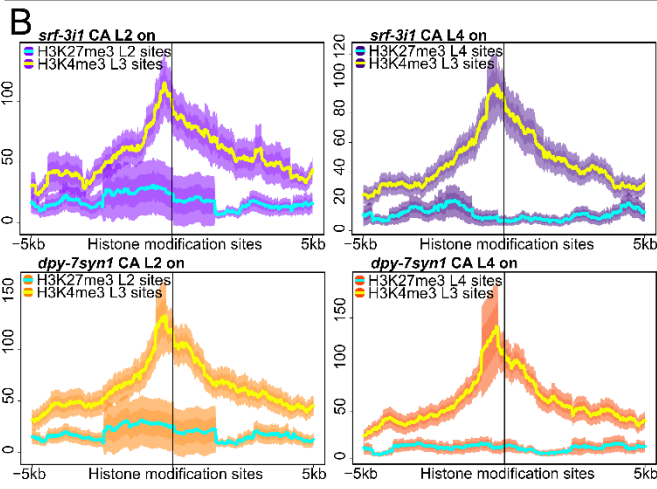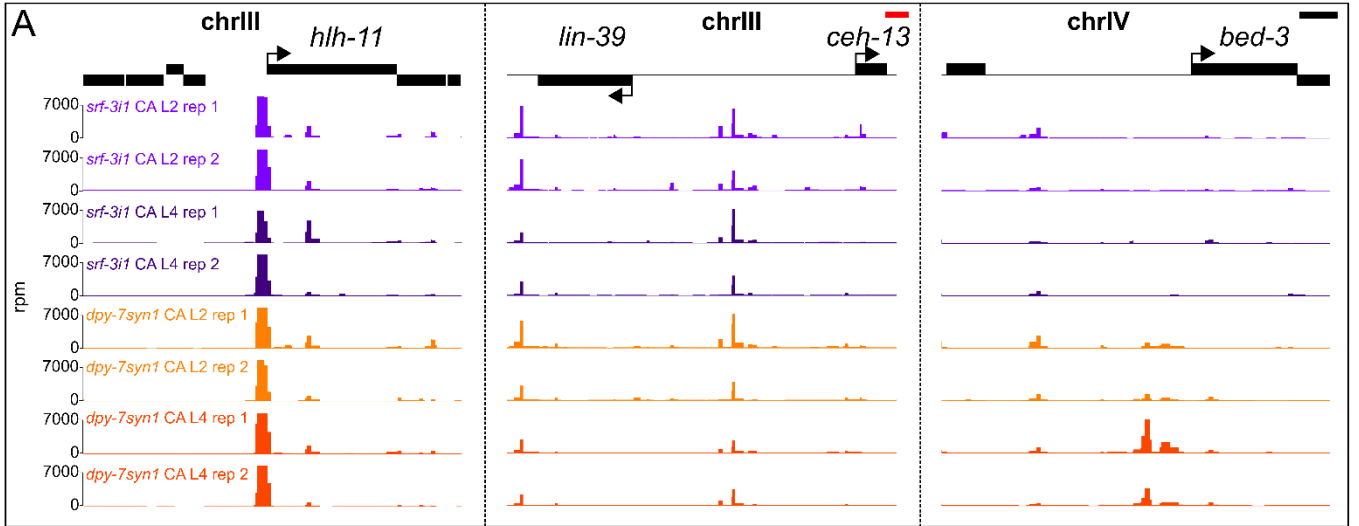
Figure 5. 3 CATaDa-identified sites associate with marks of active regulatory function and show increased conservation
(A) CATaDa signal tracks for all samples and replicates showing enrichment in the locations of previously described enhancers: upstream of the *hlh-11* gene (Chen *et al.*, 2014), between *lin-39* and *ceh-13* (Kuntz *et al.*, 2008) and upstream of *bed-3* (Jänes *et al.*, 2018)*.* For *bed-3* only *dpy-7syn1* samples form peaks at the locus. Y-axes are normalised reads per million and scale bars are 2 kb. Red scale bar corresponds only to the middle panel. **(B)** Aggregation plots showing average CATaDa peak signal for each expression domain and stage over a region ± 5 kb from the centre of either L3 H3K4me3 or stage-matched H3K27me3 histone modification sites. All plots show increased average CATaDa peak signal at H3K4me3 sites. **(C)** Barplot of enrichment of CATaDa peaks occurrence, for each sample, in different ChromHMM-predicted (Daugherty *et al.*, 2017) chromatin states, relative to simulated overlaps occurring at random across the genome, based on Monte Carlo simulations. The *p*-value of the statistical significance for each enrichment or depletions is printed opposite to the respective bar. **(D)** Venn diagrams depicting the number of overlaps of genomic locations of L2 (left) or L4 (right) CATaDa peaks for *srf-3i1* and *dpy-7syn1* expression domains between them and with stage matched HOT regions and AMA-1 (RNApol) ChIP-seq peaks. All pairwise overlaps are significant based on Monte Carlo simulations (*p*≤1.6e-55). **(E)** Aggregation plot of the average PhastCons7way score for the region ±2 kb from CATaDa peak centres for each expression domain and developmental stage peak profile. CATaDa peak locations show increased conservation compared to adjacent regions. **(F)** Barplot of the average PhastCons7way conservation score for the genomic locations of the intergenic CATaDa peaks, for each expression domain and stage peak profile, in comparison to the average for the locations of the peaks after chromosome-specific, non-overlapping and non-genic shuffling of the peaks. In all cases the actual peak positions showed significantly higher average conservation scores than the shuffled, *p*<0.0001. In B the Y-axis are CATaDa average rpm scores and in E z-scores for the plotted sequence length. Shaded areas in B are 95% confidence intervals. Error bars in F indicate the SEM. Black stars indicate statistically significant differences to the mean with a t-test, **** *p*<0.0001.

TFs, or with marks of active transcription, strengthen the hypothesis that CATaDa sites largely harbour regulatory elements.

Moreover, *cis*-regulatory elements are known to be sequences of high conservation, amongst the non-coding areas of the genome, a characteristic that has been utilised elsewhere for enhancer identification (Pennacchio *et al.*, 2006; Kuntz *et al.*, 2008). Assessment of the conservation based on nematode PhastCons7way (Spieth, Hillier & Wilson, 2005) in regions around CATaDa sites, showed that on average conservation is notably increased at the loci of the CATaDa sites in comparison to neighbouring regions for all cell-types and stages (Figure 5.3E).

Considering that a proportion of CATaDa sites were reported to be in exons, that are in general more conserved than non-coding sequences, the observed increase in conservation could have been artificially inflated. Therefore, CATaDa sites within genic sequences were excluded and the average conservation score for the intergenic CATaDa sites, for each cell-type and stage, was calculated for the actual sequences they occupied, as well as for shuffled, chromosome-specific, non-intergenic and non-overlapping positions. Remarkably, the original loci of the CATaDa sites for all were significantly more conserved (*p*<0.0001 with a t-test), approximately twice as much as the shuffled sequences (between 1.83x to 2.02x) (Figure 5.3F). This is a compelling indication that intergenic CATaDa sites are very likely to correspond to conserved *cis*-regulatory elements.

5.2.4 CATaDa chromatin accessibility profiles are comparable to whole-animal ATAC-seq and DNase-seq datasets

So far other methodologies have been used to acquire genome-wide chromatin accessibility maps in *C. elegans*, based on ATAC-seq and DNase-seq (Jänes *et al.*, 2018; Daugherty *et al.*, 2017; Ho, Quintero-Cadena & Sternberg, 2017) as already mentioned. These are at the level of the whole-animal, assaying chromatin accessibility in all tissues, as opposed to just the epidermis that is interrogated here by CATaDa. However, since they represent accessible chromatin sites, some of which have been followed up by discovery of enhancers, their comparability to CATaDa would further ensure the genuine chromatin accessibility status of CATaDa sites.

Of all the previous published work, the most recent by Jänes *et al.*, 2018 is the most comprehensive, looking at chromatin accessibility across all of development. This allowed for stage-matching of the CATaDa and ATAC-seq datasets for more relevant comparisons. Initial qualitative assessment of the signal enrichment profiles of ATAC-seq and CATaDa showed evident agreement in the localisation of multiple peaks across loci of the genome (Figure 5.4A). At the genome-wide level, average CATaDa signal for both tissues, stages and replicates, across regions of ATAC-seq peaks, exhibited strong increase specifically for the position of the stage-matched ATAC-seq open chromatin sites in comparison to adjacent regions, indicating broad agreement between the datasets (Figure 5.4B). Breakdown of the exact overlaps between datasets indicated that between 56.4% and 71.9% of the CATaDa sites were also identified by ATAC-seq at the same stage, with the co-localisation of sites being highly significant for all comparisons ($p \leq 5.1e-229$ by Monte Carlo simulations) (Figure 5.4C).

Additionally, Jänes *et al.*, 2018 had incorporated histone modification information (by ChIP-seq) and transcription initiation, as well as productive transcript elongation information (by RNA-seq), to annotate and assign regulatory roles to the discovered accessible chromatin elements. Here, this annotation is utilised to categorise the CATaDa sites that overlap ATAC-seq for their putative regulatory function. Interestingly, the majority of sites for all tissue and stages were found to overlap ATAC-seq-defined coding promoters (between 23.6% and 32.9% of the total CATaDa sites) and putative enhancers (between 34.5% and 45% of the total CATaDa sites), while very few belonged to the classification of unassigned promoters, non-coding RNAs, pseudogene promoters and other elements (Figure 5.4C). This encouraging finding underscores the value of the CATaDa-identified sites, as putative locations of active regulatory elements. Importantly, due to the cell-type-specificity of the method the putative promoters or enhancers identified are also most likely capable of driving expression in the corresponding cell-type of the epidermis where they were discovered. The identification of the tissue where an open chromatin element is active had not been possible without functional confirmation experiments.

In addition, CATaDa sites were also tested for overlaps against the other available accessible chromatin maps for a more well-rounded assessment of the comparability across methods. From Daugherty *et al.*, 2017 comparisons were made against the L3 ATAC-seq peaks that are developmentally

the closest to all CATaDa datasets. The intersection confirmed very significant overlaps with CATaDa sites for all cell-types and developmental stages ($p \leq 6.3e\text{-}227$) (Figure 5.4D left). Between 37.1% and 42.5% of CATaDa sites were also identified by whole-animal ATAC-seq at L3. Likewise, intersections with DNase-seq-identified accessible chromatin sites from arrested L1 animals, showed very significant overlaps for all CATaDa profiles ($p \leq 2.8e\text{-}79$) (Figure 5.4D right). More specifically, between 36.8% and 39.4% of CATaDa sites from each profile were also found by DNase-seq.

The above overlaps as well as the one found against stage-matched ATAC-seq peaks from Jänes *et al*., 2018, are comparable to previously reported intersections between chromatin accessibility datasets, even when acquired using the same method of identification (Jänes *et al*., 2018). The fact that CATaDa sites only correspond to a small proportion of the ATAC-seq or DNase-seq sites is very likely to be an outcome of the tissue-specificity as previously discussed. However, even when performed in the same tissue CATaDa, ATAC-seq and FAIRE-seq sites have been reported to show overlaps for approximately 50% of the sites in all pairwise comparisons (Aughey *et al*., 2018).

Interestingly, for the commonly identified accessible chromatin sites in CATaDa and stage-matched ATAC-seq (from Jänes *et al*., 2018), the intensity of the signal peaks was found to show significant correlation between the two methods ($p < 0.0001$ with a Pearson's correlation test) despite the low R^2 values capturing the increased data scattering indicating varation between the methods (Figure 5.4E). This was observed both for the seam cells and the hypodermis at both stages. Peak intensity or height in both methods is expected to somewhat reflect the degree of openness of the chromatin, both within individuals and in terms of frequency in the population. The fact that the two methods agree to a certain extent, as had been previously reported (Aughey *et al*., 2018), is an additional encouraging sign that the identified sites reflect genuine chromatin accessibility that displays similar characteristics even when detected by different methods. This is also an indication that CATaDa scores can potentially be used to identify sites that likely show higher regulatory activity.

5.2.5 LIN-22 and NHR-25 TaDa-predicted binding sites overlap significantly with seam cell and hypodermal accessible chromatin sites

In chapter 3, binding sites for LIN-22 and NHR-25 in the epidermis were identified by TaDa and were shown to significantly overlap with ATAC-seq-identified accessible chromatin sites. Transcription factor binding primarily occurs in *cis*-regulatory elements, like enhancers or promoters, that most often need to be in a permissible chromatin state for such interactions to take place (Spitz & Furlong, 2012; Klemm, Shipony & Greenleaf, 2019). Therefore, a large proportion of the binding sites (between 72.4% and 89%) were shown to overlap accessible chromatin.

For LIN-22 and NHR-25 that act in the epidermis, their binding could occur within the epidermal accessible chromatin sites. The acquisition of cell-type specific CATaDa profiles allowed for testing of this hypothesis. Initial comparative inspection of the signal profiles between CATaDa and LIN-22 revealed that
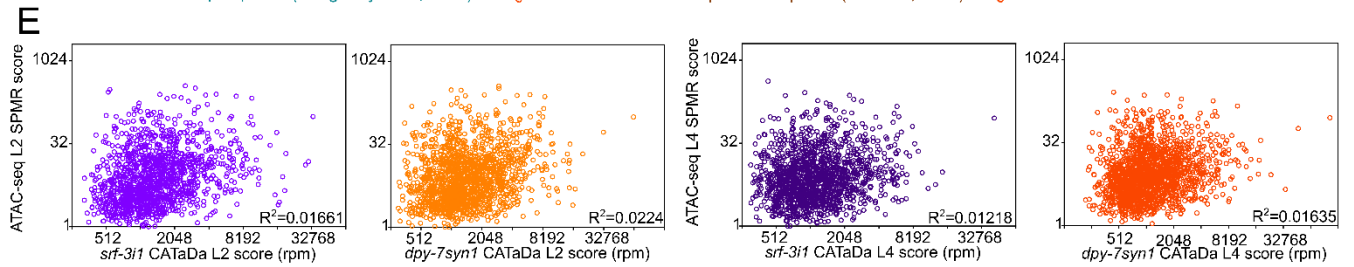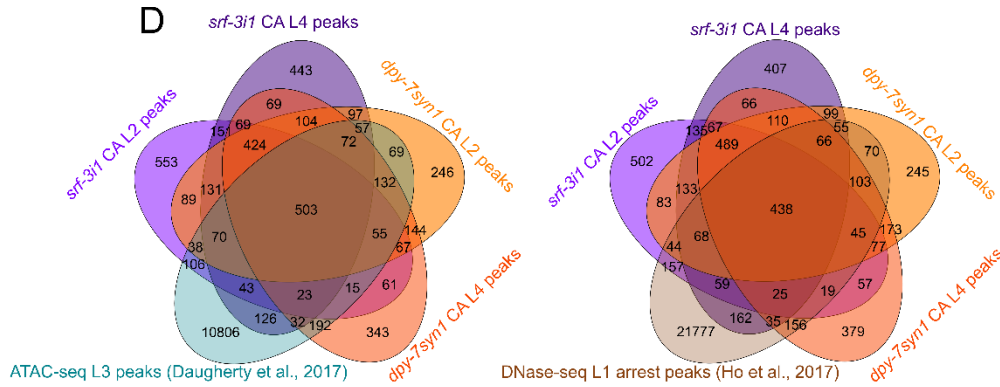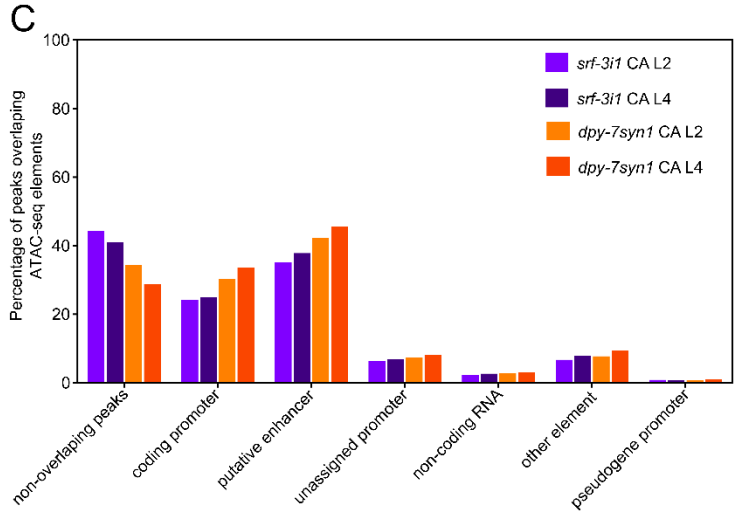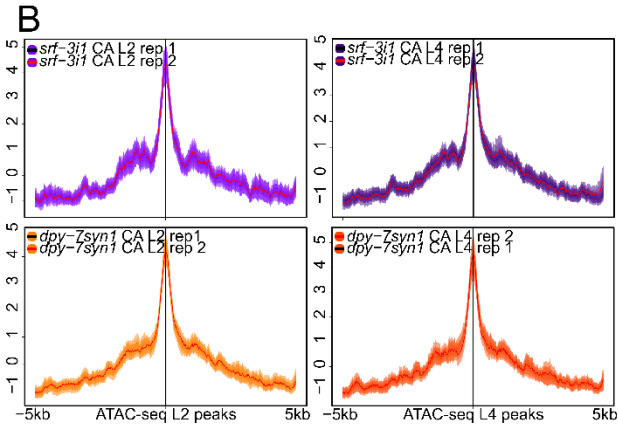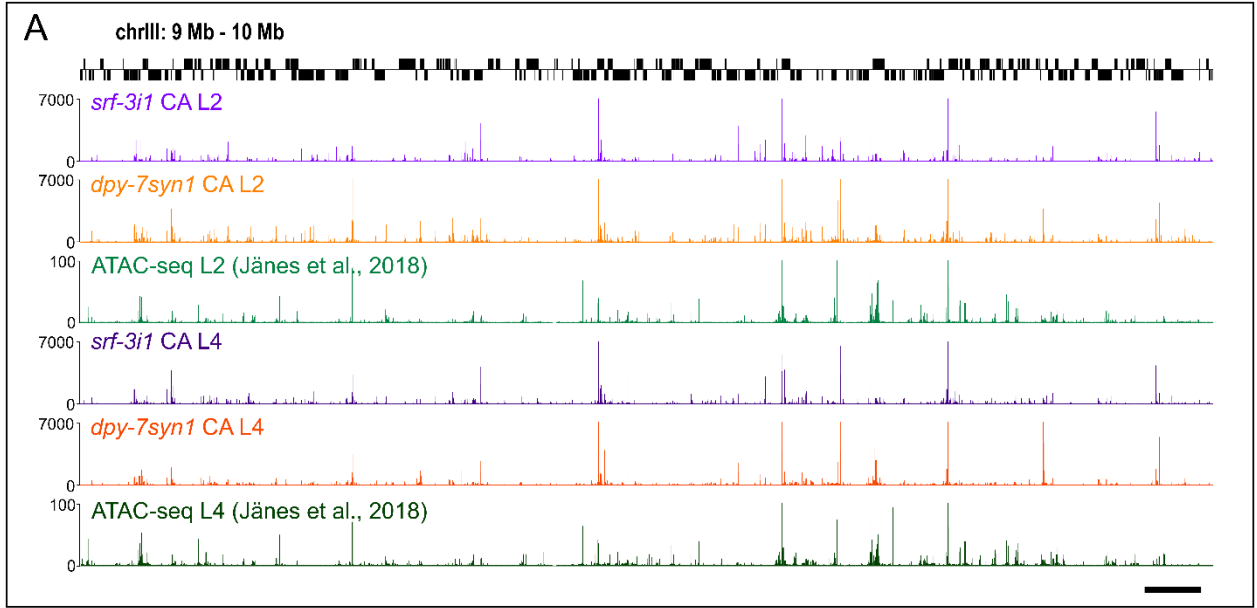
Figure 5. 4 CATaDa signal and peak profiles show significant agreement with published ATAC-seq and DNase-seq profiles. (A) Qualitative comparison example of the agreement in accessible chromatin signal across 1 Mb of chromosome II as captured by CATaDa for both expression domains and stages and whole-animal ATAC-seq for the same stages (Jänes *et al.*, 2018). Note that multiple peaks are seen across all profiles. The Y-axes are reads per million for CATaDa tracks and signal per million reads (SPMR) for ATAC-seq. One replicate per sample is plotted for clarity. **(B)** Aggregation plots showing average CATaDa signal for *srf-3i1* and *dpy-7syn1* L2 and L4 across regions ±5 kb from the centres of stage-matched ATAC-seq peaks. Average CATaDa signal is enriched at the loci of ATAC-seq peaks for all samples. Replicates have been plotted independently but show substantial agreement. **(C)** Proportions of the total CATaDa peaks for each expression domain and stage, indicated by colour, that are overlapping with different categories of regulatory annotated accessible chromatin elements from ATAC-seq (Jänes *et al.*, 2018). A CATaDa peak can overlap more than one element. **(D)** Venn diagrams depicting the number of overlaps between the CATaDa peaks discovered here and accessible chromatin sites determined in previous ATAC-seq at L3 (Daugherty *et al.*, 2017) (left) or DNase-seq at L1 arrest (Ho, Quintero-Cadena & Sternberg, 2017) (right) experiments. The overlaps with all CATaDa peak profiles are significant with Monte Carlo simulations for both datasets ($p \leq 6.3e-227$ for ATAC-seq L3 and $p \leq 2.8e-79$ for DNase-seq). **(E)** Scatterplots for correlation of peak-intensity levels for common peaks between CATaDa for *srf-3i1* and *dpy-7syn1* a t L2 and L4 and stage-matched ATAC-seq (Jänes *et al.*, 2018). All correlations were found to be significant ($p < 0.0001$ Pearson's Correlation test) and the R^2 value is reported on each plot.

for the confirmed targets *lin-17* and *cki-1,* the LIN-22 binding was overlapping the accessibility peaks indicated by CATaDa in all cell-types and stages (Figure 5.5A). In particular for *lin-17*, the chromatin accessibility signal peaks at the precise location of the CRE1 element shown in chapter 3 to drive expression in the epidermis. Likewise for NHR-25 the binding signal overlapped the accessible chromatin site upstream of the confirmed NHR-25 target *idh-1* (Shao *et al.*, 2013) (Figure 5.5B). In addition, multiple putative binding sites of NHR-25 around its genomic locus overlapped with peaks of accessible chromatin. The downstream to *nhr-25* peak of chromatin accessibility, which occurs only in the hypodermal CATaDa profiles, has been identified before as an accessible regulatory element that can drive expression in the hypodermis (Daugherty *et al.*, 2017) (Figure 5.5B).

A potentially interesting side note for the biology of TaDa identification, is that in these examples but also more broadly, the chromatin accesibility signal-enrichment appears to be more focal. The chromatin accessibility peaks often involve a single GATC fragment, while the TF peaks are typically spread across multiple GATC fragments. This is likely an outcome of the more stable binding of TFs that allows DAM more opportunity to methylate surrounding sites. Moreover, changes in chromatin structure likely accompany the TF binding causing further spreading of methylation. In contrast, in the case of CATaDa the non-targeted *dam:NLS-GFP* fusion is expected to serendipitously interact and methylate more frequently those regions that already possess stable, fully accessible open states. Those are likely to be truly narrower prior to binding of factors that induce or maintain open chromatin states.

At the genome-wide level, average peak signal for both TFs was found to be evidently more increased at the sites of accessible chromatin in comparison to adjacent regions, from both seam cell and hypodermis CATaDa at the same stage (Figure 5.5C). A more detailed breakdown revealed that the TF binding sites were significantly occurring at stage-matched CATaDa sites from both tissues ($p \leq 4e-79$ by Monte Carlo simulations) (Figure 5.5D). Similar overlap sizes were found for LIN-22 at seam and
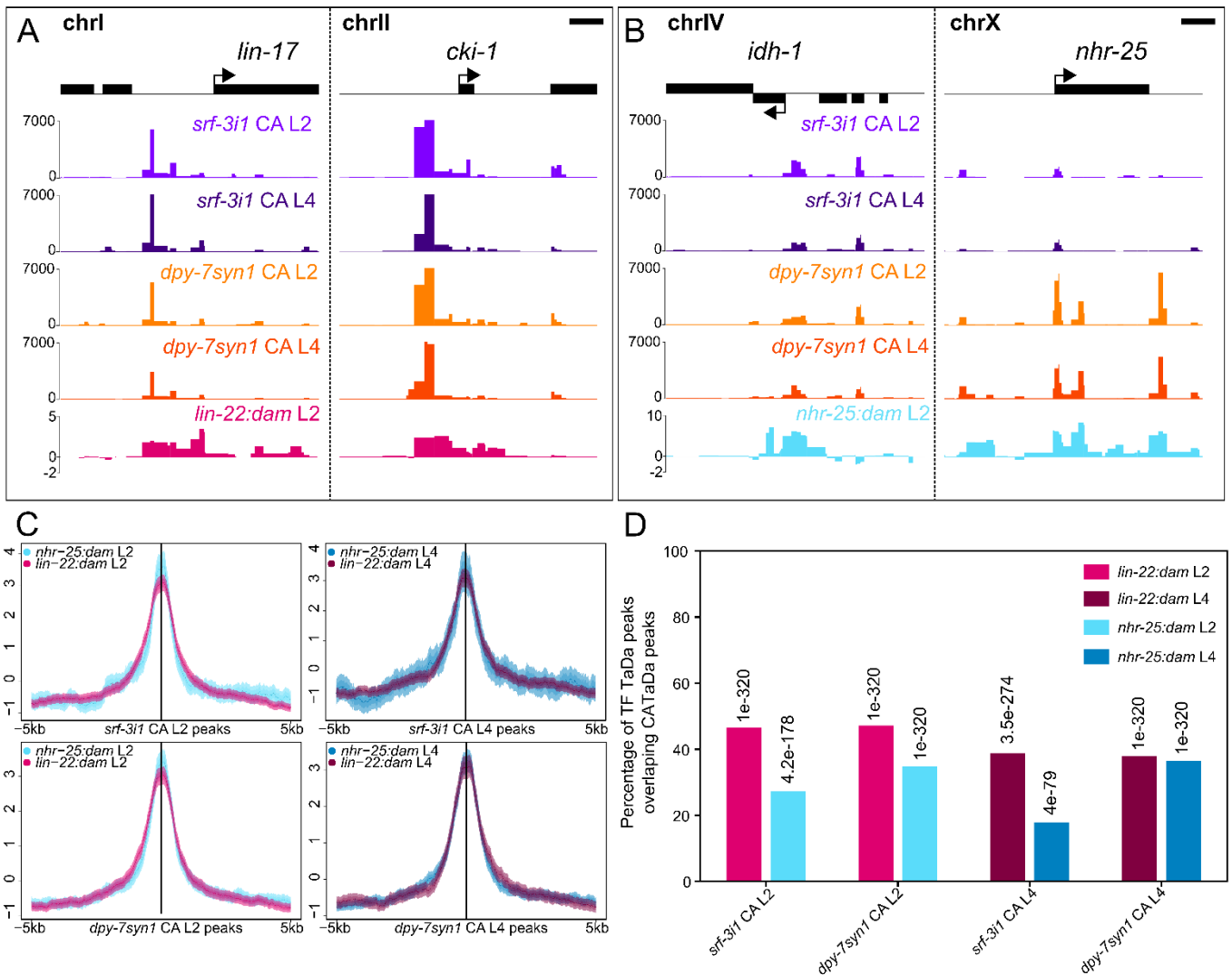
Figure 5. 5 TaDa-identified binding of LIN-22 and NHR-25 significantly overlaps with CATaDa sites. (A-B) Examples of CATaDa signal enrichment that forms significant peaks, overlapping loci with LIN-22 (A) or NHR-25 (B) TaDa peaks of putative binding. Accessible chromatin sites are found in both cell-types on the promoters of the LIN-22 targets *lin-17* and *cki-1* (A) and the promoter of the NHR-25 target gene *idh-1*(B). CATaDa peaks downstream of *nhr-25* are only found in *dpy-7syn1* profiles and overlap an NHR-25 binding peak (B). The Y-axes are reads per million for CATaDa tracks and $\log_2$*(TF:dam/NLS-GFP:dam)* scores for LIN-22 and NHR-25 TaDa tracks. Scale bars are 2 kb. **(C)** Aggregation plots showing average *lin-22:dam* and *nhr-25:dam* L2 and L4 peak signal across a region ±5 kb from the centres of all stage-matched CATaDa peaks from the *srf-3i1* or *dpy-7syn1* expression domains, indicated on the X-axis. Both TFs show preference for the CATaDa peaks loci compared to adjacent regions. **(D)** Barplot of the proportions of the total *lin-22:dam* and *nhr-25:dam* L2 and L4 peaks that overlap stage-matched CATaDa peaks from the *srf-3i1* or *dpy-7syn1* expression domains, indicated on the X-axis. The *p*-value of the statistical significance of each set of overlaps from Monte Carlo simulations is printed on top of the bars.

hypodermal CATaDa sites with approximately 46% of TF peaks overlapping at L2 and 37% at L4 stage. NHR-25 binding showed a preference for *dpy-7syn1* accessible chromatin sites, with 33.9% of the TF peaks overlapping with *dpy-7syn1* and 26.4% with *srf-3i1* CATaDa sites at L2, while 35.7% of the peaks were overlapping with *dpy-7syn1* and 17% with *srf-3i1* CATaDa sites at L4. *lin-22* is expressed specifically in the seam cells while *nhr-25* is also expressed in the hypodermis and glia (Katsanos *et al.*, 2017; Cao *et*

al., 2017; Gissendanner & Sluder, 2000). The preference for CATaDa sites of the hypodermis might reflect hypodermis specific targets but more work is required to strengthen that claim.

The number of TF peaks found to overlap between LIN-22 or NHR-25 and CATaDa sites are smaller than with ATAC-seq (between 72.4% and 89% for ATAC-seq and between 17% and 46% for CATaDa). A parsimonious explanation is that some sites have been missed in CATaDa. On the other hand, the overlaps with ATAC-seq sites are likely to be somewhat inflated because they represent all open chromatin, including from tissues outside the expression domains of the TFs. Acquisition of other profiles and systematic comparison are required to explain this. Overall, these findings illustrate that TFs that are expressed and act in the epidermis are likely to bind in epidermal accessible chromatin regions. Since 16% to 39% of CATaDa sites from each tissue and stage were bound by at least one of the two TFs, this further advocates to the regulatory role of the identified regions.

5.2.6 TaDa-identified seam cell and hypodermis expressed genes largely associate with nearby CATaDa-identified accessible chromatin sites

Since a significant proportion of CATaDa peaks were found to overlap with TF binding sites and were shown to co-localise with putative *cis*-regulatory elements and their marks, the CATaDa sites were assigned to nearby genes that they could potentially regulate. As for assignment of TF TaDa peaks in chapter 3, the CATaDa sites were assigned here to genes when their centre coordinate was positioned within 6 kb upstream of the start or 1 kb downstream of the end of the gene. A peak could be assigned to more than one gene if it fulfilled those criteria.

In total, 3988 genes were assigned with nearby CATaDa sites in the *srf-3i1* domain at L2 and 3836 genes at L4, while 3703 genes with *dpy-7syn1* CATaDa sites at L2 and 3601 at L4. In line with the extensive overlaps of CATaDa sites shown in section 5.2.2, between 42.6% and 47.2% of genes associating with sites from each spatial domain and stage were common in all sets (1698 genes) (Figure 5.6A). Moreover, the vast majority of genes corresponding to a set of CATaDa sites for a cell-type were also found for the set of the other cell-type. Only between 20.5% and 29.1% of genes associating with CATaDa sites for each expression domain and stage were not found for the CATaDa sites of at least one stage of another expression domain. All the pairwise intersections of gene-sets were highly statistically significant with a Fisher's exact test (*p*<1e-320) (Figure 5.6B).

Considering that the assigned genes are potentially regulated by elements in the accessible chromatin sites, the agreement in gene-set contents between domains could be reflecting the epidermal fate of seam cells and hypodermis. In addition, based on the fact that CATaDa sites are enriched for active regulatory marks and overlap with previously identified promoters and enhancers, it is likely that genes assigned with CATaDa sites are regulated by them and show expression in those same cell-types. To assess this, the TaDa-identified sets of expressed genes found for the same expression domains (chapter 4) were compared to the sets of genes assigned for the identified CATaDa sites. Between 32.5% and
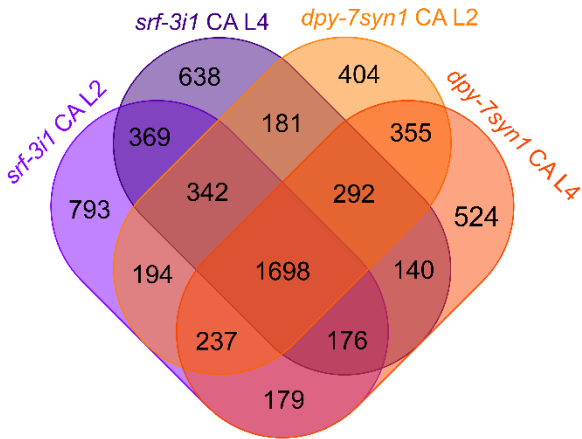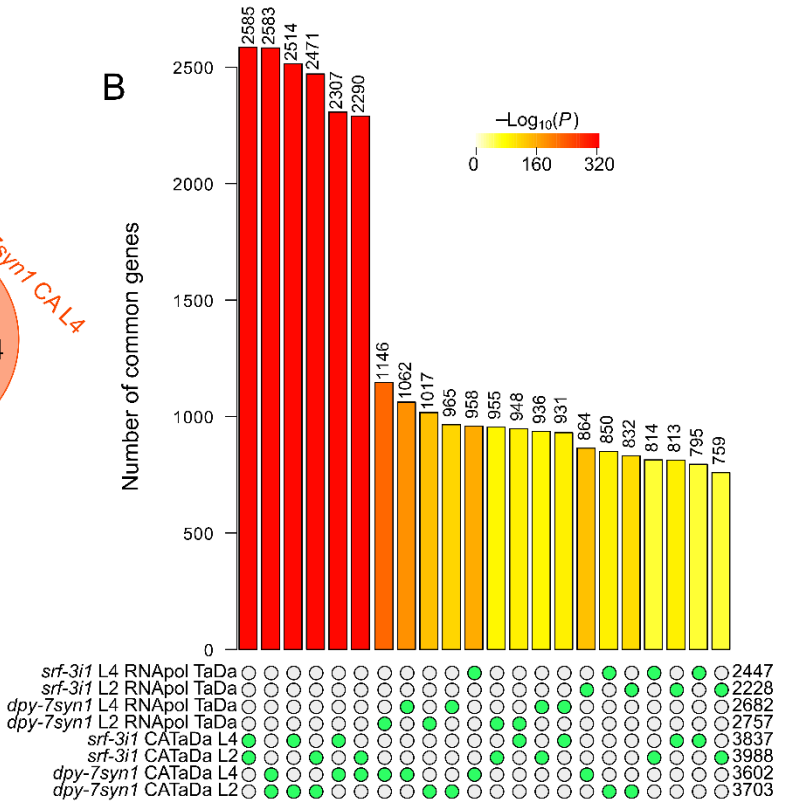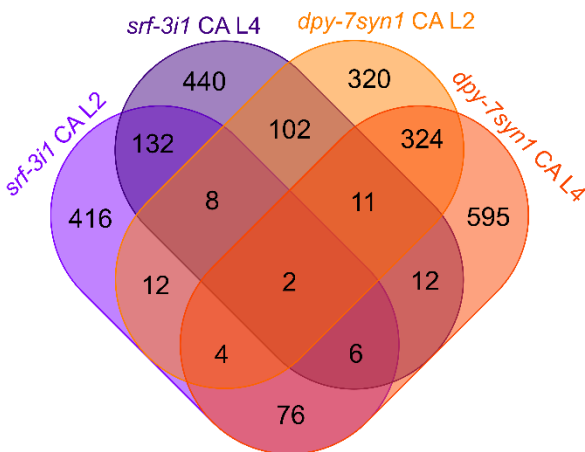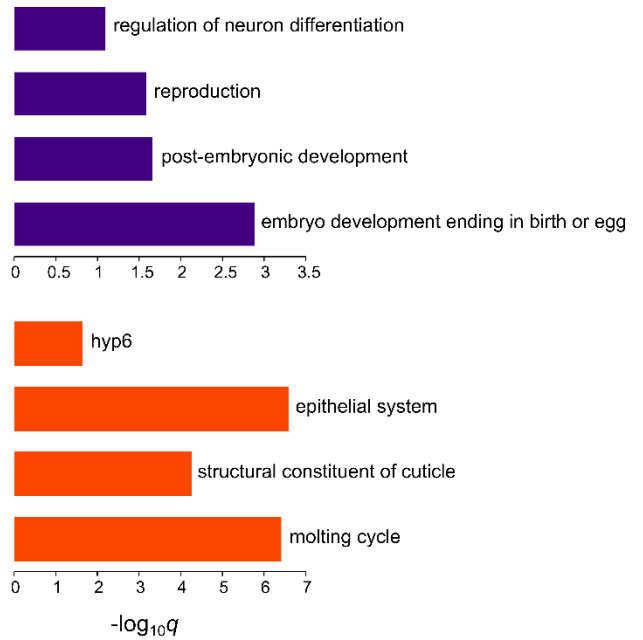
Figure 5. 6 CATaDa peak-associated genes are enriched for relevant ontology terms and over-represent epidermal expression (A) Venn diagram depicting the numbers of genes assigned to CATaDa peaks when their centres were within 6 kb upstream of the TSS to 1 kb downstream of the TES of the protein-coding gene. All the possible intersections of those gene sets for all the CATaDa samples are shown. **(B)** Barplot of all the pairwise intersections between the CATaDa gene-sets and RNApol TaDa-identified expressed genes (Chapter 4) in the *srf-3i1* and *dpy-7syn1* expression domains at L2 and L4. The exact number of common genes is printed on top of the bars. All overlaps were significant with a Fisher's exact test (*p*≤2.33e-64) with the level indicated by the colour hue of the bar. **(C)** Venn diagram of the gene-sets, assigned to peaks as described in A, for CATaDa peaks unique for each expression domain at each stage and with intergenic positioning. Very few genes are found in the overlaps as expected. **(D)** Barplots of enrichment analysis for gene ontologies and associated tissues depicting selected significantly enriched terms for the L4 gene-sets of the *srf-3i1* and *dpy-7syn1* domain-unique intergenic CATaDa peaks.

39.6% of genes found to be expressed in each expression domain and at each stage by TaDa were associated with at least one CATaDa site identified in the same cell-type and at the same stage (Figure 5.6B). These extensive overlaps of the gene-sets were highly statistically significant with a Fisher's exact test (1.57e-233≤*p*≤2.33e-64) (Figure 5.6B), indicating that the association of CATaDa sites with expressed genes is potentially linked to the active expression in those tissues.

CATaDa sites that are associated with expressed genes might harbour *cis*-regulatory elements responsible for driving the expression of those genes in the respective cell-type and could potentially lead to discovery of epidermis specific enhancers. Such putative enhancers if localised in regions of differential accessibility between the two cell-types might possess cell-type-specific expression activation attributes. Of the CATaDa sites for each domain and stage a minority, between 34% and 41% of total sites, did not show any overlap across cell-types within the same stage and could belong to that category. In addition, *cis*-regulatory elements controlling genes are most often within regions of accessible chromatin in non-coding sequences of the genome (Ho, Quintero-Cadena & Sternberg, 2017; Tsompana & Buck, 2014; Gaudet & McGhee, 2010).

Based on the above, from the selection of unique CATaDa sites for each cell-type those within genes were subtracted to generate a set of non-coding, intergenic CATaDa sites that were unique for each cell-type at each stage. These sets included 375 and 549 sites for *dpy-7syn1* at L2 and L4 respectively, as well as 328 and 358 sites for *srf-3i1* at L2 and L4 respectively. To assess if these sites likely controlled different genes, they were assigned to nearby genes as before. Encouragingly very few assigned genes were common between sets of CATaDa sites across the two cell-types (Figure 5.6C). This indicated that the vast majority of the identified cell-type-specific CATaDa sites were likely controlling different sets of genes rather than associating with different regions neighbouring the same gene.

Moreover, gene-set enrichment analysis performed on the 572 genes of the L4 *srf-3i1* unique intergenic CATaDa sites showed significant enrichment for gene ontology terms relating to seam cell roles (Figure 5.6D top). Specifically, terms relating to neurogenesis, development and reproduction all pertain to seam cell functions. Similarly, the *dpy-7syn1* counterpart set of genes was enriched for GO terms relating to molting and cuticle formation (Figure 5.6D bottom), both of which are hypodermis related functions. In addition, the set was enriched for genes relating to epithelial cells and the related hyp6 (Figure

5.6D) (all gene-set enrichment results for the above are available in Appendix B.19, B.20). This supported the hypothesis that the intergenic CATaDa sites that were unique for each of the cell-type were likely to harbour *cis*-regulatory elements controlling genes that perform cell-type-specific functions. Therefore, enhancers localising in those open chromatin regions might drive expression with specific spatial attributes within the epidermis.

5.2.7 Intergenic CATaDa sites harbour regulatory elements that can drive expression of transgenes in the *C. elegans* epidermis

To assess if these CATaDa sites harboured epidermal enhancers, a subset of 8 intergenic accessible chromatin regions that were significantly open either in the hypodermis or the seam cells, were selected to examine their capacity to drive expression and the spatial aspects of it. To achieve this, the selected sequences were cloned upstream of a minimal promoter from *pes-10* driving nucleolocalised GFP and multi-copy extrachromosomal array transgenes were generated to test any potential expression.

All sites were selected from the L4 sets of open chromatin regions to improve the chances that any potential expression would occur in late L4 staged animals where observation of transgenic expression is more easily detectable and robust. The selected CATaDa sites were all near genes that had been found to be expressed in the seam cells or hypodermis by TaDa and the names of the genes were used to also specify the elements. They ranged in size between 936 bp to 1737 bp. Two of the regions were found accessible only in the seam cells while the rest were found in the hypodermis. All but one were also identified in stage-matched ATAC-seq (Jänes *et al.*, 2018). Table 1 summarises all these key features for each of the selected sequences. Microscopy and imaging of the transgenics was performed by Mar Ferrando-Marco a Master's student I supervised.

Table 1 Key features of the selected CATaDa sites tested for the capacity to drive expression in transgenic animals

CATaDa element found in	likely ascociated gene	relative to gene position	distance from peak centre (bp)	RNApol TaDa identified expression in	chr	start	end	Jänes *et al*., 2018 annotation
srf-3i1	F22B7.3	upstream	937	*srf-3i1*	chrIII	8647415	8648488	-
srf-3i1	rps-25	upstream	5545	*srf-3i1 and dpy-7syn1*	chrIV	5952654	5953650	coding promoter
dpy-7syn1	K02A2.5	upstream	1650	*dpy-7syn1*	chrII	7413065	7414296	coding promoter
dpy-7syn1	nhr-4	distal upstream	11566	*srf-3i1 and dpy-7syn1*	chrIV	9867222	9868303	putative enhancer
dpy-7syn1	nhr-4	proximal upstream	3621	*srf-3i1 and dpy-7syn1*	chrIV	9875003	9876414	other element
dpy-7syn1	Y38F1A.8	upstream	2061	*srf-3i1*	chrII	13002045	13003782	putative enhancer, coding promoter
dpy-7syn1	nhr-25	upstream	5127	*srf-3i1 and dpy-7syn1*	chrX	13002897	13003835	other element
dpy-7syn1	nhr-25	downstream	507	*srf-3i1 and dpy-7syn1*	chrX	13013940	13014999	putative enhancer

For the two seam cell identified CATaDa sites upstream of the genes *F22B7.3* and *rps-25,* no epidermal expression was observed (Figure 5.7A, B). More specifically, for the *F22B7.3* upstream element only background expression was observed in the pharynx (Figure 5.7A right), due to the pharyngeal-specific co-injection marker (*myo-2::dsRed*) that can recombine with GFP on the extrachromosomal array

to produce pharyngeal GFP expression. This is seen in transgenic animals for other CATaDa sites as well. This element was also the only one not identified by ATAC-seq as open, likely suggesting that it could be a CATaDa false positive. The *rps-25* upstream element also did not drive expression in the seam cells or other tissues within the epidermis. However, it could drive expression in neurons of the nerve ring (Figure 5.6B), suggesting that it does harbour a tissue-specific enhancer.

The sequences from hypodermis-specific CATaDa sites all drove expression in the epidermis, indicating that they contained epidermis-specific enhancers as predicted. The element upstream of the gene *K02A2.5* showed hypodermis-specific expression that was more prominent in the anterior of the transgenic animals (Figure 5.7C right). The site, which was proximal to the TSS of the gene, was previously annotated as a coding promoter by ATAC-seq (Jänes *et al.*, 2018) and CATaDa here correctly predicts the domain of expression driven by the element. Similarly, two hypodermis-specific CATaDa sites were tested from the upstream region of the *nhr-4* gene. Both were somewhat distant from the TSS of the gene, one more proximal (3.6 kb away) and the other more distal (11.5 kb away) (Figure 5.7D left), with the proximal previously annotated as a putative enhancer. Both showed strong expression throughout the hypodermis, but not in the seam cells, with expression also observed, particularly for the distal element, in the Pn.p cells of the ventral hypodermis (Figure 5.7D right).

In contrast, the CATaDa site upstream of the gene *Y38F1A.8* drove expression primarily in the hypodermis but also to a lesser extent in the seam cells (Figure 5.7E right). This site, like the above, was found open only in the hypodermis (Figure 5.7E left) and according to ATAC-seq included both a coding promoter element and a putative enhancer. RNApol TaDa found *Y38F1A.8* expressed only in the seam cells (Table 1) but sci-RNA-seq found expression that is in agreement both spatially and in terms of levels with the observed expression here (Cao *et al.*, 2017).

Lastly, two sites around the epidermal regulator *nhr-25* were tested for enhancers: an upstream and a proximal downstream element (Figure 5.7F left). Both sites were found accessible only within the hypodermis. CATaDa sites overlapping the gene were excluded as only intergenic elements were selected. Interestingly, the upstream element showed expression both in the seam cells and hypodermis, while the downstream element showed strong expression primarily in the hypodermis and weak expression in the seam cells (Figure 5.7F right). Specifically, the downstream site overlaps a previously identified element by ATAC-seq (Daugherty *et al.*, 2017) that had been shown to drive expression in a few hypodermal cells, whereas the CATaDa-determined element is observed to drive expression throughout the hypodermis and in the seam cells.

In summary, all CATaDa sites but one harboured *cis*-regulatory elements that could drive expression. Most importantly, 6 out of the 8 tested putative regulatory elements drove expression specifically in the epidermis, confirming the tissue-specificity allowed by CATaDa in discovering enhancers by probing accessible chromatin.

Notably, as previously mentioned, the accessible chromatin around *nhr-25* was co-localising with putative NHR-25 binding (Figure 5.7F left), which has been hypothesised both here (chapter 3) and

A chrIII *F22B7.3*
srf-3i1 CA L4
dpy-7syn1 CA L4

B chrIV *rps-25*
srf-3i1 CA L4
dpy-7syn1 CA L4

C chrII *K02A2.5*
srf-3i1 CA L4
dpy-7syn1 CA L4

D chrIV *nhr-4*
srf-3i1 CA L4
dpy-7syn1 CA L4

E chrII *Y38F1A.8*
srf-3i1 CA L4
dpy-7syn1 CA L4

F chrX *nhr-25*
srf-3i1 CA L4
dpy-7syn1 CA L4
nhr-25:dam L4

CATaDa F22B7.3-upstream element:pes10::GFP

CATaDa rps-25-upstream element:pes10::GFP

CATaDa K02A2.5-upstream element:pes10::GFP

CATaDa nhr-4-distal-upstream element:pes10::GFP

CATaDa nhr-4-proximal-upstream element:pes10::GFP

CATaDa Y38F1A.8-upstream element:pes10::GFP

CATaDa nhr-25-upstream element:pes10::GFP (icbEx276)

CATaDa nhr-25-downstream element:pes10::GFP (icbEx278)

G

% of animals with GFP signal

control RNAi
nhr-25 RNAi

**

icbEx278
(*nhr-25* downstream element)

icbEx276
(*nhr-25* upstream element)

Figure 5. 7 CATaDa-identified accessible chromatin sites harbor *cis*-regulatory elements driving epidermal expression
(A-F) Experiments for identification of functional epidermal regulatory elements at the sites of L4 CATaDa peaks by assessment of the capacity to drive GFP expression in transgenic animals. CATaDa signal enrichment profiles for *srf-3i1* and *dpy-7syn1* at the L4 stage over the positions of significant peaks, indicated by teal bars for *srf-3i1* and gray for *dpy-7syn1,* that were used to build the tested reporters, are shown on the left. Representative fluorescence and brightfield images of transgenic animals at the late L4/EA stage are presented on the right, showing GFP expression driven by the respective CATaDa putative regulatory element indicated above and shown on the connected panel on the left. The *F22B7.3* upstream element (A) shows background expression only in the pharynx, the *rps-25* upstream element (B) in neurons of the nerve ring, the *K02A2.5* upstream element (C) in hypodermal nuclei mainly of the anterior body, the *nhr-4* distal and proximal upstream elements (D) in hypodermal nuclei and Pn.p cells for the distal element, the *Y38F1A.8* upstream element (E) in hypodermal nuclei and at lower levels seam cell nuclei, the *nhr-25* upstream and downstream elements (F) in seam cells and hypodermal nuclei mainly of the head region for the upstream element and throughout for the downstream. For signal tracks snapshots the scale bars are 2 kb and the Y-axes are reads per million for CATaDa tracks and $\log_2$*(nhr-25:dam*/*NLS-GFP:dam)* scores for NHR-25 TaDa tracks in F. For images scale bars are 100 μm, pink arrowheads indicate examples of non-specific background fluorescence, yellow arrowhead indicates expression in neurons of the nerve ring, white arrowheads examples of expression in hypodermal nuclei, green arrowheads in the seam cells and narrow white arrowheads in the Pn.p cells. All images are position with the ventral side down and the anterior to the left. **(G)** Quantification of the proportion of transgenic animals showing GFP expression from the *nhr-25* upstream and *nhr-25* downstream element reporters, when animals were treated with *nhr-25* or control RNAi. n numbers in the order they appear on the graph are n=31, n=13, n=24, n=37. Error bars represent the standard error of the proportion and black stars indicate statistically significant differences in the proportions between treatments with a Fisher's exact test (** $p<0.01$).

elsewhere (Shao *et al.*, 2013) to control its own expression. The overlap of open chromatin and binding could point to self-regulation occurring through the use of these accessible sites. To test this hypothesis, transgenic animals for the two discovered elements were treated either with control or *nhr-25* RNAi and the frequency in which transgenic animals showed GFP expression in the epidermis was measured. Both elements showed an increase in the frequency of transgenic animals that were expressing in the epidermis (Figure 5.7G), which was statistically significant with a Fisher's exact test only for the downstream element (downstream: *p*=0.0026, upstream: *p*=0.1133). The percentage of transgenic animals expressing for the downstream element increased from 25.8% in control treatment to 76.9% in *nhr-25* RNAi. This result provides further evidence supporting that NHR-25 likely self-regulates by supressing its expression by binding at least the proximal downstream enhancer found here.

5.3 Discussion

5.3.1 Chromatin accessibility probing by CATaDa in *C. elegans* is congruent with other established methodologies

The accessibility of chromatin provides an epigenomic level of control of the genomic information that determines transcriptional programmes (Klemm, Shipony & Greenleaf, 2019). Identifying accessible chromatin sites of the genome can indicate the location of regions with active regulatory function, often harbouring *cis*-regulatory elements, like promoters or enhancers, that control the expression of genes

(Klemm, Shipony & Greenleaf, 2019; Tsompana & Buck, 2014). Chromatin accessibility has been investigated in *C. elegans* before using DNase-seq (Ho, Quintero-Cadena & Sternberg, 2017) and ATAC-seq (Jänes *et al.*, 2018; Daugherty *et al.*, 2017) approaches and has been shown to change throughout development to mediate changes in relevant transcriptional states. However, all previous studies have been performed at the whole-animal level capturing open chromatin states from the totality of tissues. This precluded the discovery of open chromatin with tissue-specific resolution that could elucidate the contribution of chromatin accessibility in determining tissue differences and cell-fate decisions. In this chapter, the first example of acquisition of *C. elegans* cell-type-specific chromatin accessibility maps is performed using CATaDa for the first time in this model organism, to identify accessible chromatin regions in the seam cells and hypodermis.

To assess how accurately CATaDa reveals open chromatin, the profiles along with their various localisation characteristics were compared here to the published datasets from DNase-seq and ATAC-seq methodologies. The CATaDa profiles showed extensive, highly significant overlaps with the published datasets, with 36.8% to 71.9% of sites also identified in at least one of the previous datasets. The reproducibility of such sites across methods highlighted the most likely genuine openness of the CATaDa identified sites, providing support for the functionality of the method. The proportions of the overlapping sites were similar to those previously reported between ATAC-seq and DNase-seq, even between datasets employing the same method (Jänes *et al.*, 2018), indicating comparable levels of agreement across all methods.

However, the overlapping peaks were only a small proportion of the ATAC-seq or DNase-seq datasets. In total, 4449 unique accessible chromatin sites were identified for the two epidermal cell-type combined, by CATaDa. This number is substantially smaller than the combined 30828 found by ATAC-seq in embryos, L3 and adult animals in Daugherty *et al.*, 2017, the 42245 found across all development in Jänes *et al.*, 2018 or the 41825 found in embryos and arrested L1 animals by DNase-seq in Ho *et al.*, 2017. Since the seam cells and the hypodermis constitute only a subset of the cell-types of the animal, the vast discrepancy in size could be down to the tissue-specificity of CATaDa in contrast to the whole-animal-scale of the previous applications. ATAC-seq in whole-animal or only in *pax6* expressing tissues in *Xenopus* tadpoles has shown fewer peaks on average for the tissue-specific samples supporting this explanation (Kakebeen *et al.*, 2020).

Nonetheless, missing of particular sites due to CATaDa-related traits is also a potential explanation, with GATC availability biases being an obvious reason. The absence of whole-animal CATaDa or tissue-specific ATAC or DNase-seq hinder direct comparisons that could elucidate the cause of the discrepancy. Nevertheless, comparisons of accessible chromatin maps from tissue-specific ATAC-seq, FAIRE-seq and CATaDa in *Drosophila* imaginal eye discs have shown that approximately 50% of sites overlap in all pairwise comparisons (Aughey *et al.*, 2018), indicating that both tissue-specificity and method/experiment-related variability are contributors to the discrepancy in dataset sizes.

Further to the overlaps between datasets the CATaDa-discovered sites show similar positional characteristics in relationship to genes as those reported for the DNase-seq datasets (Ho, Quintero-Cadena & Sternberg, 2017). Only a minority of sites (between 24.3% and 35.7%) were within coding sequences, with the rest primarily associating with non-coding regions of the genome. Even though the sizes of the datasets differed, the proportions of sites in the different positions largely agreed showing that CATaDa also broadly captures sets of accessible sites with similar positional attributes, likely reflecting the true localisation preferences of accessible sites.

Aside of the positional agreement, sites common between CATaDa and ATAC-seq exhibited correlating peak intensities, suggesting that the methods also somewhat agreed at the levels of the openness that certain sites exhibited, as this is determined by each technique. The correlation points to the potential of using CATaDa semi-quantitatively to allow ranking of accessible sites based on "openness", which most likely reflects the frequency with which sites are found accessible and could indicate strong promoters or enhancers. The overall agreement between CATaDa and the more established methods for probing chromatin accessibility, provides evidence towards its use to identify regions of open chromatin with tissue and cell-type-specificity. It also constitutes proof-of-concept evidence for the ease of acquiring comparable accessible chromatin maps from a small subset of cells of as few as 2000 individual *C. elegans*.

5.3.2 CATaDa profiles capture chromatin regions with active regulatory roles

Chromatin accessibility has been generally shown to accurately reflect the level of regulatory activity exhibited by different areas of the genome (Klemm, Shipony & Greenleaf, 2019; Tsompana & Buck, 2014). Active promoters or enhancers that recruit TFs and the RNA polymerase are mostly free of nucleosomes or associate with more relaxed chromatin states, often dictated by specific combinations of histone modifications (Klemm, Shipony & Greenleaf, 2019; Bannister & Kouzarides, 2011). The seam cell and hypodermis CATaDa sites found here, showed preference for histone modifications associated with active enhancers like H3K4me3 but not with repressed chromatin H3K27me3 marks (Heintzman *et al.*, 2007, 2009; Bannister & Kouzarides, 2011). Comparisons to more comprehensive genome-wide annotations of chromatin states, generated by ChromHMM from another study (Daugherty *et al.*, 2017; Ernst & Kellis, 2012) based on multiple histone modification marks, revealed strong enrichment of CATaDa sites primarily for chromatin states associated with active enhancers and promoters. Therefore, CATaDa identification of accessible chromatin is very likely capturing the consequential aspect of epigenomic regulation, potentially pinpointing active *cis*-regulatory elements within the cell-type of interest.

Proving this hypothesis, open chromatin sites found by CATaDa coincide with loci of previously described enhancers with epidermal expression (Kuntz *et al.*, 2008; Chen *et al.*, 2014). Interestingly, a previously described hypodermis-specific enhancer, upstream of the gene *bed-3,* was found open only in hypodermis CATaDa profiles, highlighting the cell-type-specificity of the method and encouraging the potential to identify cell-type-specific enhancers.

More broadly, the identified CATaDa sites showed preference for the TSS of genes and proximal upstream regions, as has been previously shown in *Drosophila* (Aughey *et al.*, 2018) and *C. elegans* by different methods (Jänes *et al.*, 2018; Ho, Quintero-Cadena & Sternberg, 2017). These intergenic regions uncovered by CATaDa were also found to be highly conserved, a characteristic of regions with regulatory function, which has been previously reported for sites of accessible chromatin in *C. elegans* (Daugherty *et al.*, 2017; Ho, Quintero-Cadena & Sternberg, 2017) and has been used before as the sole indicator of the position of enhancer sequences (Kuntz *et al.*, 2008; Gaudet & McGhee, 2010).

Based both on position relative to genes and conservation, these regions are known to often contain promoters and enhancers regulating nearby genes by recruiting TFs and the RNA polymerase (Spitz & Furlong, 2012). Indeed, CATaDa sites were also found to significantly overlap loci of RNApol occupation and regions with high occupancy by TFs. Moreover, the binding sites of the epidermis-specific TFs LIN-22 and NHR-25 were significantly overlapping the CATaDa sites for the seam cells and hypodermis; constituting specific examples wherein CATaDa reveals areas where known interactions of *cis*-regulatory elements and TFs take place, demonstrating the capacity to identify regions with tissue-specific regulatory activity.

The evidence presented in this chapter associated CATaDa profiles for the seam cells and hypodermis with multiple marks and traits of genomic locations actively involved in regulation of gene expression. Consequently, chromatin accessibility identified by CATaDa is very likely indicative of the epigenomic regulatory state within the tissue or cell-type of interest and can potentially pinpoint the location of tissue or cell-type-specific *cis*-regulatory elements.

5.3.3 Epidermal CATaDa sites harbour enhancers that drive epidermis-specific expression

The application of CATaDa presented here is the first attempt in acquiring a tissue-specific map of accessible chromatin in *C. elegans*. The seam cell and the hypodermis that constitute the targeted cell-types in this case both belong to the epidermis. The produced evidence indicated that CATaDa sites very likely harboured *cis*-regulatory elements, thus the potential that they contain epidermis-specific enhancers was investigated. The rationale was that enhancers positioned in chromatin accessible only in specific cell-types could be potentially driving expression in a similar cell-type-specific manner, as has been previously shown (Aughey *et al.*, 2018).

To increase the chances of identifying such sequences, candidates were picked out of a curated set of only intergenic CATaDa sites that associated with genes that had showed expression in the seam cells or hypodermis by TaDa (Chapter 4). The assigned genes of the selected sites had not been previously characterised or associated with epidermal functions with the exception of *nhr-25*. Strikingly, from the final set of the 8 candidates that were tested in transgenic animals only 1 did not drive expression in any tissue. The remaining 7 drove expression in specific tissues demonstrating that all contained at least one

enhancer element sufficient to specify expression in a subset of the cells of the animal. Remarkably, 6 of the 7 sites drove expression specifically in cells of the epidermis.

Therefore, accessible chromatin sites identified by CATaDa in the epidermis were not only shown to contain enhancers but a majority of the enhancers they harboured were sufficient to drive tissue-specific expression in the epidermis. This constitutes a notable proof-of-concept application of CATaDa towards identifying tissue-specific enhancers in *C. elegans*. In addition, these confirmation experiments added to the thorough annotation of elements performed in Jänes *et al.*, 2018 showing that the proximal *nhr-4* element and the upstream *nhr-25* element contain enhancers.

The site upstream of the gene *rps-25* was the only one that drove expression but not in the epidermis. Even though *rps-25* was detected by TaDa in both seam cells and hypodermis, the enhancer in the CATaDa site drove specific expression in a subset of neurons of the nerve ring. Enhancers are known to often regulate genes that are very distant to their location (Noonan & McCallion, 2010) allowing for the possibility that the identified enhancer did not control *rps-25,* but a neighbouring neuronal-specific gene. However, *rps-25* is ubiquitously expressed (Cao *et al.*, 2017), thus it is possible that the identified enhancer controls its expression only in the neurons with other elements controlling it within other tissues. The localisation of this enhancer in a seam cell accessible chromatin site might reflect the potential of seam cells to generate neurons, possibly possessing a more related epigenomic state. Despite residing in open chromatin found in the seam cells, they most likely lack the TFs that can initiate expression by binding it. Nevertheless, this is a clear example where the tissue-specificity of chromatin accessibility does not agree with the specificity of expression dictated by the contained regulatory element in synthetic gene constructs. Further examination is required to assess how pervasive is the agreement between the two methodologies.

Regarding the two interrogated cell-types of the epidermis, even though the selection of the sites was made based on accessibility in only one of the two, certain hypodermal-specific accessible sites were also found to drive expression in the seam cells where they did not show accessibility by CATaDa. Specifically, the *Y38F1A.8* and the *nhr-25* elements. This could be an artefact of the transgenic expression caused by the multi-copy nature of the transgenes that can lead to some ectopic expression (Katsanos *et al.*, 2017). This explanation is more likely for the *Y38F1A.8* and *nhr-25* downstream elements where only minimal expression is observed in the seam cells. Another potential explanation is that the CATaDa signal in the seam cells is not as robust as in the hypodermis, possibly due to the vastly fewer cells, resulting to loss of certain genuinely accessible sites. Lastly, the seam cells might rely less on chromatin state for regulation, or possess chromatin that transitions between states more frequently and does not allow the same level of accessibility as in the hypodermis.

It is notable to mention that both of the accessible *nhr-25* elements colocalised with putative binding sites for NHR-25. Since the epidermal specific NHR-25, has been thought to self-regulate (in chapter 3 and (Shao *et al.*, 2013)), the hypothesis that it might do so through these epidermis-specific accessible *cis*-regulatory elements was tested. The findings showed that the frequency of expression from the downstream element increased in the absence of NHR-25. Assuming that this element is indeed controlling

the expression of *nhr-25,* this provides a potential underlying mechanism for the self-regulation, suggesting that NHR-25 acts to suppress its own expression. Deletion or alteration of the element in the endogenous context will be required to formally confirm this hypothesis.

Overall, these data demonstrate the feasibility of CATaDa in acquiring genuine tissue-specific chromatin accessibility information that truly reflects regulatory activity. It has been compellingly used here to identify epidermis-specific enhancers, providing proof that it can capture active chromatin regions within the tissue of interest and opening the way to dissect differences in chromatin states between tissues or cell-types without cell-isolation in *C. elegans.*

Chapter 6

General Discussion

6.1 TaDa as a powerful tool to discover transcription factor targets in *C. elegans*

Transcription factors are largely responsible for determining which parts of the genomic information will be decoded, thus controlling a plethora of biological processes (Spitz & Furlong, 2012). Development, which is a particularly complex process, utilises networks of transcription factors that along with signalling pathways and influences from other environmental cues carry out morphogenesis (Davidson, 2010). Such a complex developmental process is the pattering of the *C. elegans* epidermis. The seam cells divide both symmetrically to increase their numbers, and asymmetrically to produce differentiated daughters while maintaining the stem cell pool (Joshi *et al.*, 2010). Currently there are only a few transcription factors that are known to participate in these fate determination events (Chisholm & Hsiao, 2012) and little is known about how they bring about regulated developmental outcomes.

So far attempts to decipher how these TFs are connected to a network and which are their targets that act on cell fate decisions, have been mostly focused on genetic analysis. Networks of genetic interactions that have been discovered this way (summaries of such networks in (Thompson *et al.*, 2016; Chisholm & Hsiao, 2012)) lack the resolution required to identify direct regulatory relationships and are most often limited to known participants. The ability to assay for TF targets *en masse* can both inform the structure of existing networks, revealing specific direct interactions and can allow expansion of gene networks with previously unknown members of developmental programs of interest.

In this study, the first application of targeted DamID is performed in *C. elegans* to identify targets of the transcription factors LIN-22 and NHR-25. The aim was to use these two factors to both adapt previously established protocols and assess whether the method can be used in this model organism to begin dissecting gene networks with direct component interactions. LIN-22 was chosen based on previous work of the lab, which recovered the *hes*-related factor from a genetic screen implicating it in seam cell development (Katsanos *et al.*, 2017). It has been little studied so far and mostly in the context of neurogenesis from the seam cells (Wrischnik & Kenyon, 1997; Yip & Heiman, 2016). Our previous work had linked LIN-22 to the Wnt signalling in achieving its role of establishing the correct division symmetry or asymmetry (Katsanos *et al.*, 2017). NHR-25 on the other hand is a better studied epidermal factor, with the added benefit of available ChIP-seq data that allowed for comparisons between techniques.

ChIP-seq is a more established method for identifying TF targets than the more recent TaDa (or conventional DamID), with more available resources and datasets from large-scale projects (Kudron *et al.*, 2018; Celniker *et al.*, 2009). Evidence presented in this thesis, however, suggests that TaDa can be comparable or superior in certain aspects to ChIP-seq. Because ChIP-seq is based on capturing instances of TF-DNA interactions by cross-linking very large amounts of *C. elegans* larvae, typically between 1-2 ml of "worm pellet", is needed to have enough material for the identification (Askjaer, Ercan & Meister, 2014). In the case of TaDa as few as ~2000 individuals were found here to be enough to acquire above 10 million

unique mappable reads from methylated DNA coming from as few as 32 seam cells per individual. Similarly in mammalian cell lines, 333x more cells were used to identify targets for the same TF by ChIP-seq than TaDa (Cheetham *et al.*, 2018). As an added advantage, the TaDa protocol is significantly shorter and requires fewer, more broadly available reagents. Coupled with the substantially smaller populations of animals required for efficient identification of targets, it allows for parallel execution of multiple TF target identification experiments and considerable multiplexing of sequencing. Overall, this can significantly reduce the cost of both small and large-scale experiments.

One of the possible disadvantages of TaDa is that labelled DNA turns hemi-methylated after every division, making detection of GATC sites by DpnI about 60x less efficient (reported by the manufacturer New England Biolabs https://international.neb.com/faqs/2012/08/24/will-dpni-cleave-hemimethylated-dna). Nevertheless, in the developmentally active seam cells that undergo multiple divisions, even during the L2 stage where two rapid divisions occur, robust identification of targets was performed with approximately 2000 possible binding sites identified for both TFs. As a specific example, the newly confirmed LIN-22 target *rnt-1,* was found to have significant LIN-22 binding on its putative promoter at the L2 stage, indicating that cell division does not inhibit identification of targets by TaDa.

Most importantly, the identified sets of putative targets by TaDa were extensively comparable to those identified by ChIP-seq. In the case of NHR-25, for which direct comparisons across methods could be made, the vast majority of TaDa putative target genes (>61%) were also identified by stage-matched ChIP-seq. This indicated that there is extensive agreement between the methodologies with certain targets that were reproducibly identified by both, being very likely genuinely under NHR-25 regulation.

However, the ChIP-seq L2 dataset was approximately twice as large, both in terms of peaks and potential target genes. Understanding the source of this difference is important in assessing the effectiveness and resolution of target identification by TaDa, therefore it was pursued in this study. A parsimonious explanation relates to biases and limitations of TaDa, having to do with availability of GATC sites and amplification preferences during the PCR, which could have produced some loss of targets. However, the average length of GATC fragments in *C. elegans* is 368 bp and is thus unlikely that the depleted availability of GATC sites is pervasive enough to significantly reduce detection of targets. The detection of bound targets in ChIP happens *in vitro* and after chemical crosslinking, which has been hypothesised to produce artefacts. Such false positives have been claimed to constitute a proportion of the high-occupancy target (HOT) regions that have been determined by ChIP-seq experiments to be bound by multiple transcription factors (Teytelman *et al.*, 2013; Araya *et al.*, 2014). Interestingly, comparisons of genomic locations presented here indicated that only 13% of L2 HOT regions were represented in NHR-25 TaDa, in stark contrast to the 83% in ChIP-seq, corresponding to 38% of the size difference between the targets identified by the two methods. This could partially explain the discrepancy in the size of the total putative target sites discovered. Nonetheless, the principal contributor for this difference is most likely the tissue specificity that TaDa can achieve.

In most ChIP-seq experiments the entirety of the expression domain of a TF is usually assayed for binding targets. This means that the discovered set of targets is an amalgamation of binding from the TF in question across all tissues where it is expressed. One of TaDa's major advantages is that it can perform this only within a tissue or cell-type of interest, potentially identifying only targets of a TF that are relevant to the functions of that given tissue (Aughey, Cheetham & Southall, 2019; Aughey & Southall, 2016). Evidence in this study suggests that this is likely to be true. TaDa for NHR-25 was performed in a subset of its expression domain (*wrt-2* expression domain) including the seam cells and the hypodermis but excluding glial cells (Cao *et al.*, 2017). Genome-wide binding localisation co-association analysis here indicated that the TaDa binding profile for NHR-25 was more similar to that of other epidermal TFs than neuronal TFs, which the ChIP-seq profile was better associated with, likely due to its targets in the glia. Furthermore, tissue enrichment analysis on the set of NHR-25 ChIP-seq putative target genes that were not identified by TaDa showed enrichment for neuronal related terms. The narrower domain of identification of TaDa can therefore readily provide a biological explanation of the difference in numbers of targets between methods.

The above data are also preliminary indications of the tissue-specificity that can be achieved using TaDa and has the potential to dissect otherwise complex TF behaviours related to tissue-specific binding. For example, the much expanded family of HLH factors in *C. elegans* are known to bind different selections of targets depending on their dimerising partner, which could be tissue-specifically expressed (Grove *et al.*, 2009). TaDa could dissect such tissue-specific preferences. In addition, tissue-specificity of target identification, especially for more broadly expressed TFs, can permit links between regulation of a target and function within a tissue to be made more confidently.

It should be mentioned that tissue-specificity can also be potentially achieved for ChIP-seq with cell-isolation or by expressing the TF of interest fused to an epitope for immunoprecipitation from a tissue-specific promoter. However, this may result in fate-altering outcomes, especially for potent developmental factors, while evidence shown in this thesis highlights that TaDa avoids any such effects due to its construct configuration and mode of identification.

In this study, TaDa data were also used to identify DNA binding motifs for LIN-22 and NHR-25. Even though TaDa peaks have the drawback of being broader in size than ChIP-seq peaks, which are most commonly used for identification, available algorithms were capable of detecting appropriate motifs. Interestingly, peak score was the most effective parameter while filtering the regions to be used for motif identification. In addition, in NHR-25 TaDa, GATC fragments with the highest local score within a significant peak were often found to overlap with ChIP-seq summits for the same factor. Specifically, 74.8% of overlaps occured in this manner, likely indicating the site where binding takes place. Two such GATC fragments were overlapping confirmed binding sites for LIN-22 on the *lin-17* promoters further supporting this hypothesis. These findings point towards the possibility of obtaining some quantitative information from TaDa experiments and the potential of extracting focal peak information similar to ChIP-seq summits.

Overall, data presented in chapter 3 indicate that TaDa is a method that can successfully and descriptively capture the genome-wide binding of transcription factors in *C. elegans.* It does so comparably well to the more established ChIP-seq methodology, while being currently the only approach that can readily perform TF target identification tissue-specifically. It is cheaper and easier to perform and has been used in this study to produce a plethora of valuable information regarding the binding of the TFs LIN-22 and NHR-25 in the epidermis, in an effort to begin deciphering the seam cell gene regulatory network.

6.2 Probing gene expression and chromatin accessibility in the *C. elegans* epidermis using TaDa

Fate-determining transcriptional states emerge largely through the regulation of transcription by TFs and at the epigenomic level by chromatin states with different degrees of permissibility of expression. TaDa can allow us to dissect all layers of this regulation and also assay the gene expression outcome (Aughey & Southall, 2016). It is within our broader interests to understand how the seam cell acquire their identity, what sets them apart from their differentiated hypodermal daughters and how they maintain their fate. Distinct gene expression programs, established by epigenomic control and TF regulation, are reasonably expected to characterise the seam cells as well as their differentiated progenitors. In the previous section, the use of TaDa to dissect regulatory interactions of TFs, known to control seam cell fate, in the *C. elegans* epidermis, was discussed. However, to expand the pool of factors that participate in seam cell patterning and fate decisions and to better understand the batteries of genes utilised that describe the seam cell fate, TaDa was also used here for identification of epidermal gene expression profiles, for the first time in *C. elegans*.

The genome-wide occupancy of RPB-6, a subunit that participates in all RNApol complexes, was successfully used to identify expressed genes. A crucial component in appropriately separating the seam cell gene expression profile from that of the related hypodermis was the use of promoters with strict expression domains. It was achieved with the *de novo* discovery of a seam cell specific enhancer in the 1st intron of *srf-3* and the alteration of the promoter of *dpy-7*. The achieved specificity, even between two closely related cell-types, was indicated here for the discovered gene-sets, which were enriched for cell-type-appropriate ontologies. Most importantly, a plethora of previously described examples of cell-type-specifically expressed factors were identified by TaDa to be expressed with the correct specificity. Cases like the known major seam cell factors *elt-1, egl-18, ceh-16*, all of which had been previously shown to be specifically expressed in the seam cells (Katsanos *et al.*, 2017), were found to be expressed in the equivalent profiles by TaDa. In combination to similar proof for the specificity of the hypodermal profiles, the TaDa findings created the potential to identify other such unknown factors that based on their discovered seam cell-specificity could be implicated in seam cell development.

1090 genes were found in this study to be expressed in the seam cells but not the hypodermis, representing the set that could include such factors. Remarkably, small scale RNAi screens for 9 TFs and

11 chromatin factors mined from the above set confirmed this working hypothesis, with the identification of the TFs *efl-3* and *tbx-35* and the chromatin factors *F43G9.12, hmg-4* and *hda-1* as novel seam cell development regulators. This evidence stands as proof to the feasibility of cell-type-specific gene expression profiling in *C. elegans* using TaDa, which is confirmed here to allow discovery of biologically meaningful differences between cell-types of the epidermis.

Nevertheless, an important question to address pertains to the comparability of the TaDa findings to other currently available resources for tissue-specific transcriptome elucidation. Currently two other methods have produced such information for the seam cells and the hypodermis, PAT-seq and sci-RNA-seq (Blazie *et al.*, 2017; Cao *et al.*, 2017). Comparisons between the sets of genes identified with each method and TaDa showed very significant overlaps and even correlation in the levels of expression as they are captured by each method. The agreement between methods provided further evidence in support of the identification of biologically accurate cell-type-specific gene expression profiles by TaDa.

Some notable differences between these datasets were also identified. The hypodermal PAT-seq was based on expression from the *dpy-7* promoter, which is shown here in its original form to also drive expression in the seam cells and is therefore not exclusively hypodermal. In terms of the seam cell dataset, PAT-seq had missed important seam cell expressed genes like *elt-1*, *ceh-16* or even the strongly expressed *srf-3* (Page *et al.*, 1997; Cassata *et al.*, 2005; Höflich *et al.*, 2004). In addition, out of the 5 genes that TaDa identified and were shown to cause a seam cell phenotype when knocked down, none were identified to be expressed in the seam cells by PAT-seq. Considering that PAT-seq requires very large amounts of material (liquid cultures were used for the datasets compared here), often involves toxic transgenic expression of poly(A)-binding protein and relies on a more complex experimental protocol (Yang, Edenberg & Davis, 2005; Blazie *et al.*, 2015, 2017), it is fair to conclude that TaDa identification is a robust alternative.

On the other hand, the sci-RNA-seq datasets were significantly more extensive than the TaDa for both tissues. It is currently hard to assess how much of the difference between the expression profiles from the two methods represents truly expressed genes that TaDa has missed. Specific examples like the known seam cell expressed factors *lin-22* (Katsanos *et al.*, 2017) and *rnt-1,* that have been identified by sci-RNA-seq but not TaDa, illustrate the limitations of TaDa, presumably when it comes to GATC availability biases (2 GATC sites within the *lin-22* sequence) and low expression levels respectively. TaDa identified a smaller proportion of the lowly expressed genes determined by sci-RNA-seq than compared to the total, possibly indicating lower sensitivity for low expression in comparison. Possible biases may exist in sci-RNA-seq datasets as well, inflating the size of the transcriptomes. For example, attribution of single cell transcriptomes to a specific cell-type happens based on specific gene markers and could have conceivably led to some miss-attributions, especially in the case of the epidermis where transcriptomes for seam and hypodermis did not cluster apart sufficiently (Cao *et al.*, 2017). However, higher sensitivity of sci-RNA-seq compared to TaDa is more likely to be the source of the phenomenon. The extent to which

each method captures the true transcriptome of a cell, remains to be clarified and will require acquisition of TaDa gene expression profiles for other tissues as well.

A key feature of sci-RNA-seq is that a single application has allowed the elucidation of transcriptomes for all cell-types of *C. elegans,* which would require a separate identification experiment for each tissue if they were performed by TaDa. This makes sci-RNA-seq cost effective relative to the wealth of information that it creates, but also severely costly for more focused questions. Another global identification of all transcriptomes would have to be performed, for example to identify the gene expression profile of seam cells at L4 (sci-RNA-seq in (Cao *et al.*, 2017) was performed at L2) which was readily done by TaDa on the same experiment. Most importantly though, TaDa has the capacity to generate additional information for a cell's state as part of the same experimental data. Firstly, TaDa can capture the tissue-specific expression of small-RNAs in the same expression profiles, as seen here with the identification of the microRNAs *mir-42, mir-43, mir-44* and *mir-47,* which were also confirmed to have potential seam cell developmental roles. This cannot be achieved by RNA-seq based approaches without the alteration of the protocol (Lu, Meyers & Green, 2007). Secondly, the control samples of TaDa have been shown to capture chromatin accessibility within the same tissue, providing a description of the existing epigenomic regulation (Aughey *et al.*, 2018; Klemm, Shipony & Greenleaf, 2019). Based on the above, TaDa is comparable and not redundant to other gene expression identification methods and could be the method of choice in approaching certain biological questions.

As mentioned above, in another first for this study, the chromatin accessibility in the epidermis was also assayed using TaDa. This version of the method termed CATaDa utilised the *dam:NLS-GFP* control data to assess chromatin openness in the seam cell and hypodermis. The identified regions showed genome-wide localisation characteristics that were similar to previously reported for open chromatin (Aughey *et al.*, 2018; Ho, Quintero-Cadena & Sternberg, 2017), as well as showed association with various histone modification marks linked to active chromatin states, as is expected for accessible sequences (Heintzman *et al.*, 2007; Bannister & Kouzarides, 2011).

The CATaDa chromatin accessibility profiles were significantly comparable to those acquired in *C. elegans* by whole-animal ATAC-seq and DNase-seq methodologies (Jänes *et al.*, 2018; Ho, Quintero-Cadena & Sternberg, 2017; Daugherty *et al.*, 2017). Differences in the detected accessible regions were observed but cannot lead to definitive conclusions for comparisons between techniques, as CATaDa is tissue-specific. The chromatin accessibility probing performed in this study in the seam and hypodermis is the first example of cell-type-specific assessment of open chromatin in *C. elegans*. Cell or nuclear isolation would be required to attempt the same using ATAC-seq or DNase-seq, with the previously mentioned drawbacks of the process (Zhang & Kuhn, 2013).

The tissue-specificity achieved by CATaDa was proven here with the efficient identification of epidermis-specific enhancers in 6 out of the 8 tested accessible regions. Half of those regions could drive expression that precisely matched the hypodermal specificity predicted by CATaDa. Accessible chromatin is known to harbour *cis*-regulatory elements (Tsompana & Buck, 2014; Klemm, Shipony & Greenleaf,

2019) that are more conserved than other non-coding regions of the genome, something that was broadly observed for the data produced here, possibly indicating that the locations of multiple other such enhancers could be identified within the acquired datasets.

The fact that all this wealth of information about the transcriptional and regulatory state of a tissue or cell-type can be produced with the single methodology of TaDa strongly underscores the value of this method for use in *C. elegans*. Findings in this study have illustrated how unknown participants of the seam cells development can be elucidated using TaDa, expanding our current description of the mechanisms underlying the patterning of the tissue. Such findings discussed below, constitute a paradigm of how a single method can be used to expand and dissect gene networks in any context of interest.

6.3 An expanded seam cell developmental gene network based on TaDa findings

A principal aim of this study was to use TaDa to begin deciphering the gene regulatory network that controls epidermal development and primarily seam cell patterning. Experiments performed in all chapters of this study contributed information that allowed the expansion of the gene network and the determination of the precise nature of newly identified regulatory interactions that occur within it.

Findings from chapter 3 revealed a selection of direct targets for the known TFs of the seam cell network LIN-22 and NHR-25, results from chapter 4 added new previously unknown factors to the network by assaying seam cell-specific RNApol occupancy and in chapter 5 a selection of epidermal enhancers were identified by CATaDa. These findings were taken together along with previous summaries of the core seam cell development regulatory network (Thompson *et al.*, 2016; Koh & Rothman, 2001; Chisholm & Hsiao, 2012) and were combined with other literature-derived data (Katsanos *et al.*, 2017; Cassata *et al.*, 2005; Brabin, Appleford & Woollard, 2011; van der Horst *et al.*, 2019), to propose here an updated network underlying the development of the epidermis and patterning of seam cells and the fate decisions involved. This network is presented in figure 6.1 and describes the previously known, newly identified and inferred regulatory relationships between the different factors, at once, without depicting temporal or spatial information for the occurrence of these interactions.

This study has substantially refined the positions of *lin-22* and *nhr-25* in the seam cell network and has created significant insight into their functions and mode of action. *lin-22* was originally studied for its function in supressing neurogenesis in V1-V4 seam cells, which was thought to at least partially happen through repression of the Hox genes like *lin-39* and *mab-5* (Wrischnik & Kenyon, 1997). We also recently recovered *lin-22* from a genetic screen for its role in establishing division symmetry or asymmetry in the seam cells. Based on genetics and smFISH evidence, we had proposed that it acts by antagonising Wnt signalling by repressing the receptor gene *lin-17* and the target *egl-18* (Katsanos *et al.*, 2017). Here LIN-22 was confirmed by TaDa and smFISH to directly repress *mab-5* in V1-V4 seam cells but not *lin-39*. It was also shown to directly repress *lin-17* by binding two conserved elements of its promoter, but no direct

interaction with *egl-18* was detected. Amongst the list of putative targets there were other Wnt components (such as *mom-1, bar-1, lit-1, mig-5, mom-5, gsk-3, mig-14, pop-1*) that were not tested here but could point to a more broad suppression of Wnt signalling as previously proposed (Katsanos *et al.*, 2017; Gorrepati, Thompson & Eisenmann, 2013). In addition, a feedback on its own expression was also not found to be direct by TaDa.
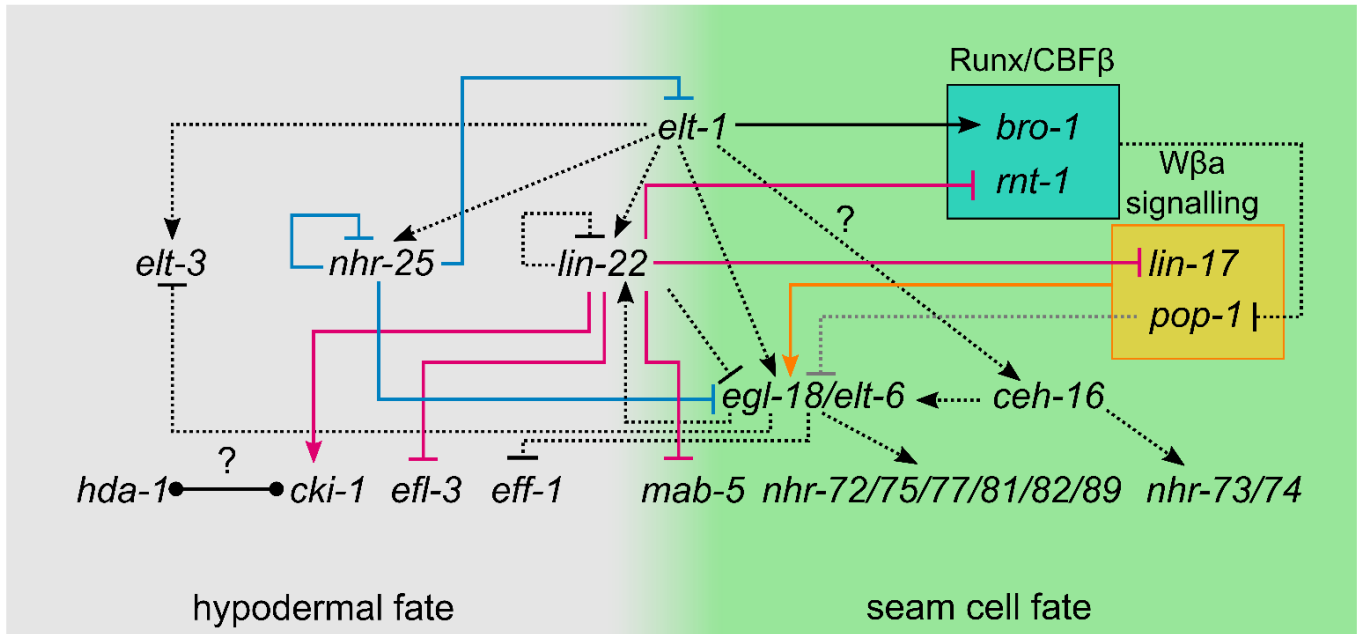
Figure 6. 1 Consolidated gene regulatory network controlling epidermal development. A combined network of interactions between factors from previous published research and findings of this study. Solid lines correspond to direct regulatory interactions, with arrowheads indicating activation and flatheads repression. Dashed lines indicate interactions that are yet unknown whether they are direct or involve more intermediate factors and are mostly genetic interactions. Black lines indicate literature-described interactions, magenta lines LIN-22 targets and blue lines NHR-25 targets found in this study. The orange line indicates activation by the Wβa signalling cascade and the grey dashed line suppression by POP-1 in the absence of activation of the Wβa pathway. Question marks denote possible interactions that currently lack definitive evidence. The proposed interaction between *hda-1* and *cki-1* is unknown in direction or mode and is indicated by the symmetric circle-ended line. The genes have been arranged between left and right depending on the fate they are either known or hypothesised to promote. The separation does not describe a regulatory state for the acquisition of a specific fate This is an adapted and updated version of the networks presented in (Thompson *et al.*, 2016; Koh & Rothman, 2001; Chisholm & Hsiao, 2012), including data also from (Katsanos *et al.*, 2017; Cassata *et al.*, 2005; Brabin, Appleford & Woollard, 2011; van der Horst *et al.*, 2019) and new links identified in this study.

Aside of previously hypothesised targets, LIN-22 was found to repress *rnt-1* the Runx homologue of *C. elegans*, which in complex with BRO-1 promote seam cells fate and symmetric divisions by supressing POP-1 (van der Horst *et al.*, 2019). Other newly identified targets included *efl-3,* a novel seam cell regulator found by RPB-6 TaDa to be expressed in the seam cells. This E2F TF was also found to be repressed by LIN-22 TaDa, demonstrating how data from TF target identification and gene expression profiling within the same tissue, can converge to uncover such candidates. *efl-3* knockdown was shown here to increase terminal seam cell number, potentially suggesting a role in mediating differentiation of seam cell daughters to the hypodermal fate in wild-type conditions, although other mechanisms via regulation of proliferation are also possible based on knowledge about the human homologue (Di Stefano, Jensen & Helin, 2003;

Lammens *et al.*, 2009; Endo-Munoz *et al.*, 2009). Lastly, LIN-22 was found here to activate the cell cycle inhibitor *cki-1*. Knockdown increases seam cell number and has been shown in other developing tissues to act by instructing G1 arrest in order to permit differentiation to progress (Hong, Roy & Ambros, 1998; Matus *et al.*, 2015). Therefore, it has been assigned to a hypodermal fate promoting role here. Interestingly, the newly identified by RNApol TaDa, seam cell regulator *hda-1,* could possibly act in the seam cells in collaboration with *cki-1*. In vulva development, the G1 arrest mediated by *cki-1* has been hypothesised to promote *hda-1* activity that leads to adoption of a differentiated anchor cell fate (Matus *et al.*, 2015). *hda-1* knockdown is also shown here to increase seam cell number, thus it could act in collaboration with *cki-1* in the hypodermal differentiation process. In light of the new links along with previous data, *lin-22* seems likely to possess a role in mediating correct hypodermal differentiation in the seam cells, apart from suppressing neurogenesis.

In the case of *nhr-25,* two new targets were identified by TaDa creating novel connections within the seam cell gene network. Specifically, NHR-25 was found to directly repress the major seam cell fate regulators *egl-18* and *elt-1* by TaDa and smFISH. *elt-1* in particular was repressed within differentiating daughters. These finding suggest a more prominent than previously thought role for *nhr-25* in the determination of the hypodermal fate and correct execution of the differentiation program, by repressing important seam cell fate factors. The discovery of this direct suppression of *elt-1* constitutes proof in the larval context, for previously hypothesised feedbacks between *elt-1* and some of its putative targets like *nhr-25,* at the level of embryonal epidermal morphogenesis (Chisholm & Hsiao, 2012).

On a different front, previous ChIP-seq studies had proposed self-regulation of *nhr-25* expression based on the identification of NHR-25 binding in its proximity (Shao *et al.*, 2013), and TaDa reproduced this binding. Additionally, CATaDa identified regions of accessible chromatin around *nhr-25* that contained enhancers driving expression that matches the *nhr-25* expression domain and were thus proposed to be the *cis*-regulatory elements controlling its expression. These elements overlapped the NHR-25 binding sites and the expression from one of them (downstream) was significantly increased by *nhr-25* silencing. Based on this evidence self-regulation of *nhr-25* by direct suppression is proposed in this study. This also illustrates how information on TF binding and chromatin accessibility, both of which can be generated by a single TaDa experiment, can further our knowledge on regulatory interactions.

6.4 Towards a quantitative regulatory network describing seam cell development

Results in this study have presented a new working model towards the deciphering of the seam cell gene regulatory network, relying on the broad capabilities of TaDa as a methodology. Targets of known factors and new participants of the network were identified by TaDa. When TaDa is used in combination with other methods, this allows confirmation and assessment of the type of regulatory interactions. A key experimental combination used here is TaDa and smFISH, where interactions predicted by TaDa can

readily be confirmed by smFISH in mutant or silenced backgrounds. It should be noted that this combination does not provide biochemical confirmation of the interaction between the TF and a given target site, but given the reproducibility of DamID and the ability of smFISH to quantify individual transcripts with single-cell resolution, this can be considered strong evidence for a direct interaction.

smFISH generates invaluable information essential for the assembly of the network, regarding the type of regulatory relationships that are indicated by TaDa (i.e. positive or negative), as well as the tissues and the stage or developmental timepoint at which an interaction likely takes place. This is further mediated by the tissue-specificity of both methods. Putative target identification by TaDa already happens within a specific tissue of interest, allowing for the effect of the regulation to be observed in the correct tissue by smFISH, while also capturing how that expression may change elsewhere. Examples here like the identification of *elt-1* repression by NHR-25 only in anterior hypodermal-destined daughters of seam cell divisions, illustrate how the discovered regulatory interactions can lead to more detailed descriptions, with spatiotemporal dimensions better explaining the developmental outcome. Employing this combined methodology the regulatory network can be largely worked out in great detail and it will be further pursued in the future.

Additionally, another critical advantage conferred by smFISH is that the discovered connections can also possess a quantitative character. Providing a measure of the effect on expression of a gene from a regulatory interaction, the quantitative information along with the directionality of the links, can permit modelling of sub-circuits of the network, providing a mathematical understanding of the genetics of seam cell patterning (Piano *et al.*, 2006). Tissue-specific chromatin accessibility that can be simply acquired by any TaDa experiment, can further enrich such networks with chromatin state information for the identified sites. Furthermore, the elucidation of tissue-specifically expressed miRNAs by standard RNApol TaDa protocols and their action on their targets, which can be easily detected by smFISH, could facilitate their introduction to future expanded versions of the network. For these reasons, TaDa and smFISH is a powerful combination, which can be used to address similar questions in *C. elegans* and other systems. The updated seam cell gene regulatory network represents a framework for future experiments to build upon, expand, study existing interactions and identify new connections by incorporating more detailed spatiotemporal information. The conserved nature of the majority of the participating factors means that this network has the potential to be used as reference to inform us about the logic of developmental decisions in other stem cell contexts, as well as describe interactions that more broadly underlie stem cell behaviour.

Bibliography

Altun, Z.F. & Hall, D.H. (2009) Introduction. *WormAtlas*. [Online] Available from: doi:10.3908/wormatlas.1.1.

Altun, Z.F. & Hall, D.H. (2002) WormAtlas Hermaphrodite Handbook - Epithelial System - Hypodermis. *WormAtlas*. [Online] Available from: doi:10.3908/wormatlas.1.13.

Altun, Z.F., Herndon, L.A., Wolkow, C.A., Crocker, C., et al. (2020) *WormAtlas*. [Online]. Available from: http://www.wormatlas.org.

Ambros, V. & Horvitz, H.R. (1984) Heterochronic mutants of the nematode *Caenorhabditis elegans*. *Science*. [Online] 226 (4673), 409–416. Available from: doi:10.1126/science.6494891.

Angeles-Albores, D., N. Lee, R.Y., Chan, J. & Sternberg, P.W. (2016) Tissue enrichment analysis for *C. elegans* genomics. *BMC Bioinformatics*. [Online] 17 (1), 366. Available from: doi:10.1186/s12859-016-1229-9.

Antebi, A. (2006) Nuclear hormone receptors in *C. elegans*. *WormBook : the online review of C. elegans biology*. [Online] 1–13. Available from: doi:10.1895/wormbook.1.64.1.

Apfeld, J. & Alper, S. (2018) What can we learn about human disease from the nematode *C. elegans*? In: *Methods in Molecular Biology*. [Online]. pp. 53–75. Available from: doi:10.1007/978-1-4939-7471-9_4.

Araya, C.L., Kawli, T., Kundaje, A., Jiang, L., et al. (2014) Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. *Nature*. [Online] 512 (7515), 400–405. Available from: doi:10.1038/nature13497.

Arendt, D. (2008) The evolution of cell types in animals: Emerging principles from molecular studies. *Nature Reviews Genetics*. [Online] 9 (11), 868–882. Available from: doi:10.1038/nrg2416.

Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., et al. (2016) The origin and evolution of cell types. *Nature Reviews Genetics*. [Online] 17 (12), 744–757. Available from: doi:10.1038/nrg.2016.127.

Askjaer, P., Ercan, S. & Meister, P. (2014) Modern techniques for the analysis of chromatin and nuclear organization in *C. elegans*. *WormBook : the online review of C. elegans biology*. [Online] 1–35. Available from: doi:10.1895/wormbook.1.169.1.

Aspock, G., Kagoshima, H., Niklaus, G. & Burglin, T.R. (1999) Caenorhabditis elegans has scores of hedgehog- related genes : Sequence and expression analysis. *Genome Research*. [Online] 9,

909–923. Available from: doi:10.1101/gr.9.10.909.

Aughey, G.N., Cheetham, S.W. & Southall, T.D. (2019) DamID as a versatile tool for understanding gene regulation. *Development (Cambridge)*. [Online] 146 (6), 1–8. Available from: doi:10.1242/dev.173666.

Aughey, G.N., Estacio Gomez, A., Thomson, J., Yin, H., et al. (2018) CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. *eLife*. [Online] 7, 1–22. Available from: doi:10.7554/eLife.32341.

Aughey, G.N. & Southall, T.D. (2016) Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdisciplinary Reviews: Developmental Biology*. [Online] 5 (1), 25–37. Available from: doi:10.1002/wdev.205.

Austin, J. & Kenyon, C. (1994) Cell contact regulates neuroblast formation in the Caenorhabditis elegans lateral epidermis. *Development*. 120 (2), 313–323.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., et al. (2009) MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*. [Online] 37 (SUPPL. 2), 202–208. Available from: doi:10.1093/nar/gkp335.

Bannister, A.J. & Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Research*. [Online] 21 (3), 381–395. Available from: doi:10.1038/cr.2011.22.

Barkoulas, M., Vargas Velazquez, A.M., Peluffo, A.E. & Félix, M.A. (2016) Evolution of New cis-Regulatory Motifs Required for Cell-Specific Gene Expression in *Caenorhabditis*. *PLoS Genetics*. [Online] 12 (9), 1–23. Available from: doi:10.1371/journal.pgen.1006278.

Barkoulas, M., van Zon, J.S., Milloz, J., van Oudenaarden, A., et al. (2013) Robustness and Epistasis in the *C. elegans* Vulval Signaling Network Revealed by Pathway Dosage Modulation. *Developmental Cell*. [Online] 24 (1), 64–75. Available from: doi:10.1016/j.devcel.2012.12.001.

Ben-Tabou De-Leon, S. & Davidson, E.H. (2007) Gene regulation: Gene control network in development. *Annual Review of Biophysics and Biomolecular Structure*. [Online] 36, 191–212. Available from: doi:10.1146/annurev.biophys.35.040405.102002.

Betschinger, J. & Knoblich, J.A. (2004) Dare to be different: Asymmetric cell division in *Drosophila, C. elegans* and vertebrates. *Current Biology*. [Online] 14 (16), 674–685. Available from: doi:10.1016/j.cub.2004.08.017.

Bhambhani, C., Ravindranath, A.J., Mentink, R.A., Chang, M. V., et al. (2014) Distinct DNA Binding Sites Contribute to the TCF Transcriptional Switch in *C. elegans* and *Drosophila*. *PLoS Genetics*. [Online] 10 (2). Available from: doi:10.1371/journal.pgen.1004133.

Black, J.C., Van Rechem, C. & Whetstine, J.R. (2012) Histone Lysine Methylation Dynamics: Establishment, Regulation, and Biological Impact. *Molecular Cell*. [Online] 48 (4), 491–507. Available from: doi:10.1016/j.molcel.2012.11.006.

Blazie, S.M., Babb, C., Wilky, H., Rawls, A., et al. (2015) Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in *Caenorhabditis elegans* intestine and muscles. *BMC Biology*. [Online] 13 (1). Available from: doi:10.1186/s12915-015-0116-6.

Blazie, S.M., Geissel, H.C., Wilky, H., Joshi, R., et al. (2017) Alternative polyadenylation directs tissue-specific miRNA targeting in *Caenorhabditis elegans* somatic tissues. *Genetics*. [Online] 206 (2), 757–774. Available from: doi:10.1534/genetics.116.196774.

Block, D.H. & Shapira, M. (2015) GATA transcription factors as tissue-specific master regulators for induced responses. *Worm*. [Online] 4 (4), e1118607. Available from: doi:10.1080/21624054.2015.1118607.

Boukhibar, L.M. & Barkoulas, M. (2016) The developmental genetics of biological robustness. *Annals of Botany*. [Online]. Available from: doi:10.1093/aob/mcv128.

Boulikas, T. (1994) Putative nuclear localization signals (NLS) in protein transcription factors. *Journal of Cellular Biochemistry*. [Online] 55 (1), 32–58. Available from: doi:10.1002/jcb.240550106.

Brabin, C., Appleford, P.J. & 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*. [Online] Available from: doi:10.1371/journal.pgen.1002200.

Brabin, C. & Woollard, A. (2012) Finding a niche for seam cells? *Worm*. [Online] 1 (2), 107–111. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3670224/pdf/worm-1-107.pdf.

Branda, C.S. & Dymecki, S.M. (2004) Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. *Developmental Cell*. [Online] 6 (1), 7–28. Available from: doi:10.1016/S1534-5807(03)00399-X.

Brenner, S. (1973) The genetics of behaviour. *British Medical Bulletin*. [Online] 29, 269–271. Available from: doi:https://doi.org/10.1093/oxfordjournals.bmb.a071019.

Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics*. [Online] 77 (1), 71–94. Available from: doi:10.1002/cbic.200300625.

Briese, M., Esmaeili, B., Johnson, N.M. & Sattelle, D.B. (2006) pWormgatePro enables promoter-driven knockdown by hairpin RNA interference of muscle and neuronal gene products in *Caenorhabditis elegans*. *Invertebrate Neuroscience*. [Online] 6 (1), 5–12. Available from: doi:10.1007/s10158-005-0011-x.

Broitman-Maduro, G., Lin, K.T.H., Hung, W.W.K. & Maduro, M.F. (2006) Specification of the *C. elegans* MS blastomere by the T-box factor TBX-35. *Development*. [Online] 133 (16), 3097–3106. Available from: doi:10.1242/dev.02475.

Buck, S.H., Chiu, D. & Saito, R.M. (2009) The cyclin-dependent kinase inhibitors, *cki-1* and *cki-2*, act in overlapping but distinct pathways to control cell cycle quiescence during *C. elegans* development. *Cell Cycle*. [Online] 8 (16), 2613–2620. Available from: doi:10.4161/cc.8.16.9354.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., et al. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*. [Online] 10 (12), 1213–1218. Available from: doi:10.1038/nmeth.2688.

Callinan, P.A. & Feinberg, A.P. (2006) The emerging science of epigenomics. *Human molecular genetics*. [Online] 15 Spec No (1), 95–101. Available from: doi:10.1093/hmg/ddl095.

Calvo, D., Victor, M., Gay, F., Sui, G., et al. (2001) A POP-1 repressor complex restricts inappropriate cell type-specific gene transcription during *Caenorhabditis elegans* embryogenesis. *EMBO Journal*. [Online] 20 (24), 7197–7208. Available from: doi:10.1093/emboj/20.24.7197.

Cameron, E.R. & Neil, J.C. (2004) The Runx genes: Lineage-specific oncogenes and tumor suppressors. *Oncogene*. [Online] 23 (24), 4308–4314. Available from: doi:10.1038/sj.onc.1207130.

Cañestro, C., Yokoi, H. & Postlethwait, J.H. (2007) Evolutionary developmental biology and genomics. *Nature Reviews Genetics*. [Online] 8 (12), 932–942. Available from: doi:10.1038/nrg2226.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., et al. (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. [Online] 357 (6352), 661–667. Available from: doi:10.1126/science.aam8940.

Cassata, G., Shemer, G., Morandi, P., Donhauser, R., et al. (2005) ceh-16/engrailed patterns the embryonic epidermis of *Caenorhabditis elegans*. *Development*. [Online] 132 (4), 739–749. Available from: doi:10.1242/dev.01638.

Celniker, S.E., Dillon, L.A.L., Gerstein, M.B., Gunsalus, K.C., et al. (2009) Unlocking the secrets of the genome. *Nature*. [Online] 459 (7249), 927–930. Available from: doi:10.1038/459927a.

Chalfie, M., Horvitz, H.R. & Sulston, J.E. (1981) Mutations that lead to reiterations in the cell lineages of *C. elegans*. *Cell*. [Online] 24 (1), 59–69. Available from: doi:10.1016/0092-8674(81)90501-8.

Cheetham, S.W., Gruhn, W.H., Ameele, J. Van Den, Krautz, R., et al. (2018) Targeted DamID reveals differential binding of mammalian pluripotency factors. *Development*. [Online] Available from: doi:10.1242/dev.170209.

Chen, R.A.J., Down, T.A., Stempor, P., Chen, Q.B., et al. (2013) The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures. *Genome Research*. [Online] 23 (8), 1339–1347. Available from: doi:10.1101/gr.153668.112.

Chen, R.A.J., Stempor, P., Down, T.A., Zeiser, E., et al. (2014) Extreme HOT regions are CpG-dense promoters in *C. elegans* and humans. *Genome Research*. [Online] 24 (7), 1138–1146. Available from: doi:10.1101/gr.161992.113.

Chen, Z., Eastburn, D.J. & Han, M. (2004) The *Caenorhabditis elegans* Nuclear Receptor Gene *nhr-25* Regulates Epidermal Cell Development. *Molecular and Cellular Biology*. [Online] 24 (17), 7345–7358. Available from: doi:10.1128/mcb.24.17.7345-7358.2004.

Chikina, M.D. & Troyanskaya, O.G. (2012) An effective statistical evaluation of chipseq dataset similarity. *Bioinformatics*. [Online] 28 (5), 607–613. Available from: doi:10.1093/bioinformatics/bts009.

Chisholm, A.D. & Hsiao, T.I. (2012) The *Caenorhabditis elegans* epidermis as a model skin. I: Development, patterning, and growth. *Wiley Interdisciplinary Reviews: Developmental Biology*. [Online] Available from: doi:10.1002/wdev.79.

Chisholm, A.D. & Xu, S. (2012) The *Caenorhabditis elegans* epidermis as a model skin. II: Differentiation and physiological roles. *Wiley Interdisciplinary Reviews: Developmental Biology*. [Online] Available from: doi:10.1002/wdev.77.

Choy, S.W., Wong, Y.M., Ho, S.H. & Chow, K.L. (2007) *C. elegans* SIN-3 and its associated HDAC corepressor complex act as mediators of male sensory ray development. *Biochemical and Biophysical Research Communications*. [Online] 358 (3), 802–807. Available from: doi:10.1016/j.bbrc.2007.04.194.

Conte, D., MacNei, L.T., Walhout, A.J.M. & Mello, C.C. (2015) RNA Interference in *Caenorhabditis elegans*. *Current Protocols in Molecular Biology*. [Online] 2015 (January), 26.3.1-26.3.30. Available from: doi:10.1002/0471142727.mb2603s109.

Cook, D.E., Zdraljevic, S., Roberts, J.P. & Andersen, E.C. (2017) CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Research*. [Online] 45 (D1), D650–D657. Available from: doi:10.1093/nar/gkw893.

Corsi, A.K. (2006) A biochemist's guide to *C. elegans*. *Analytical biochemistry*. [Online] 359 (1), 1. Available from: doi:10.1016/j.bbi.2008.05.010.

Crooks, G., Hon, G., Chandonia, J. & Brenner, S. (2004) WebLogo: a sequence logo generator. *Genome Res*. [Online] 14, 1188–1190. Available from: doi:10.1101/gr.849004.1.

Cunha, A., Azevedo, R.B., Emmons, S.W. & Leroi, A.M. (1999) Variable cell number in nematodes. *Nature*. [Online] 402 (6759), 253. Available from: doi:10.1038/46211.

Daugherty, A.C., Yeo, R.W., Buenrostro, J.D., Greenleaf, W.J., et al. (2017) Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Research*. [Online] 2096–2107. Available from: doi:10.1101/gr.226233.117.Freely.

Davidson, E.H. (2010) Emerging properties of animal gene regulatory networks. *Nature*. [Online] 468 (7326), 911–920. Available from: doi:10.1038/nature09645.

Davidson, E.H. & Erwin, D.H. (2006) Gene regulatory networks and the evolution of animal body plans. *Science*. [Online]. 311 (5762) pp.796–797. Available from: doi:10.1126/science.1113832.

Deal, R.B. & Henikoff, S. (2011) The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nature Protocols*. [Online] 6 (1), 56–68. Available from: doi:10.1038/nprot.2010.175.

Diao, Y., Guo, X., Li, Y., Sun, K., et al. (2012) Pax3/7BP is a Pax7- and Pax3-binding protein that regulates the proliferation of muscle precursor cells by an epigenetic mechanism. *Cell Stem Cell*. [Online] 11 (2), 231–241. Available from: doi:10.1016/j.stem.2012.05.022.

Ding, S.S. & Woollard, A. (2017) Non-muscle myosin II is required for correct fate specification in the *Caenorhabditis elegans* seam cell divisions. *Scientific Reports*. [Online] 7 (1), 3524. Available from: doi:10.1038/s41598-017-01675-7.

Dupuy, D., Li, Q., Deplancke, B., Boxem, M., et al. (2004) A First Version of the *Caenorhabditis elegans* Promoterome. *Genome Research*. [Online] (Hope 1991), 2169–2175. Available from: doi:10.1101/gr.2497604.1.

Endo-Munoz, L., Dahler, A., Teakle, N., Rickwood, D., et al. (2009) E2F7 can regulate proliferation, differentiation, and apoptotic responses in human keratinocytes: Implications for cutaneous squamous cell carcinoma formation. *Cancer Research*. [Online] 69 (5), 1800–1808. Available from: doi:10.1158/0008-5472.CAN-08-2725.

Engler, C., Kandzia, R. & Marillonnet, S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE*. [Online] 3 (11). Available from: doi:10.1371/journal.pone.0003647.

Ernst, J. & Kellis, M. (2012) ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods*. [Online] 9 (3), 215–216. Available from: doi:10.1038/nmeth.1906.

Evans, T. (2006) Transformation and microinjection. *WormBook*. [Online] 1–15. Available from: doi:10.1895/wormbook.1.108.1.

Félix, M.-A. & Barkoulas, M. (2015) Pervasive robustness in biological systems. *Nature Reviews. Genetics*. [Online] 16 (8), 483–496. Available from: doi:10.1038/nrg3949.

Ferré, Q., Charbonnier, G., Sadouni, N., Lopez, F., et al. (2019) OLOGRAM: determining significance of total overlap length between genomic regions sets. *Bioinformatics*. [Online] 36 (November 2019), 1920–1922. Available from: doi:10.1093/bioinformatics/btz810.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., et al. (2010) Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in *Drosophila* Cells. *Cell*. [Online] 143 (2), 212–224. Available from: doi:10.1016/j.cell.2010.09.009.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., et al. (2004) VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*. [Online] 32 (WEB SERVER ISS.), 273–279. Available from: doi:10.1093/nar/gkh458.

Friedlander-Shani, L. & Podbilewicz, B. (2011) Heterochronic control of AFF-1-mediated cell-to-cell fusion in *C. elegans*. *Advances in Experimental Medicine and Biology*. [Online] 713, 5–11. Available from: doi:10.1007/978-94-007-0763-4_2.

Frokjaer-Jensen, C., Davis, M.W., Sarov, M., Taylor, J., et al. (2014) Random and targeted transgene insertion in C. elegans using a modified Mosl transposon. *Nat. Methods*. [Online] 11 (5), 529–534. Available from: doi:10.1038/nmeth.2889.Random.

Frøkjær-Jensen, C., Davis, M.W., Sarov, M., Taylor, J., et al. (2014) Random and targeted transgene insertion in *Caenorhabditis elegans* using a modified Mos1 transposon. *Nature Methods*. [Online] 11 (5), 529–534. Available from: doi:10.1038/nmeth.2889.

Fuxman Bass, J.I., Tamburino, A.M., Mori, A., Beittel, N., et al. (2014) Transcription factor binding to *Caenorhabditis elegans* first introns reveals lack of redundancy with gene promoters. *Nucleic Acids Research*. [Online] 42 (1), 153–162. Available from: doi:10.1093/nar/gkt858.

Gaudet, J. & McGhee, J.D. (2010) Recent advances in understanding the molecular mechanisms regulating *C. elegans* transcription. *Developmental Dynamics*. [Online] 239 (5), 1388–1404. Available from: doi:10.1002/dvdy.22246.

Ghildiyal, M. & Zamore, P.D. (2009) Small silencing RNAs: An expanding universe. *Nature Reviews Genetics*. [Online] 10 (2), 94–108. Available from: doi:10.1038/nrg2504.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*. [Online] 6 (5), 343–345. Available from: doi:10.1038/nmeth.1318.

Gilbert, S.F. (2000) *Developmental Biology*. 6th edition. Sinauer Associates.

Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., et al. (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*. [Online] 143 (4), 540–551. Available from: doi:10.1016/j.cell.2010.10.004.

Gilleard, J.S. & Mcghee, J.D. (2001) Activation of Hypodermal Differentiation in the *Caenorhabditis elegans* Embryo by GATA Transcription Factors ELT-1 and ELT-3. *Molecular and Cellular Biology*. [Online] Vol. 21 (7), 2533–2544. Available from: doi:10.1128/MCB.21.7.2533.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., et al. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*. [Online] 17 (6), 877–885. Available from: doi:10.1101/gr.5533506.

Gissendanner, C.R. & Sluder, A.E. (2000) *nhr-25*, the *Caenorhabditis elegans* ortholog of ftz-f1, is required for epidermal and somatic gonad development. *Developmental Biology*. [Online] 221 (1), 259–272. Available from: doi:10.1006/dbio.2000.9679.

Gleason, J.E. & Eisenmann, D.M. (2010) Wnt signaling controls the stem cell-like asymmetric division of the epithelial seam cells during *C. elegans* larval development. *Developmental Biology*. [Online] Available from: doi:10.1016/j.ydbio.2010.09.005.

Golden, J.W. & Riddle, D.L. (1984) The *Caenorhabditis elegans* dauer larva: Developmental effects of pheromone, food, and temperature. *Developmental Biology*. [Online] 102 (2), 368–378. Available from: doi:10.1016/0012-1606(84)90201-X.

Gorrepati, L., Krause, M.W., Chen, W., Brodigan, T.M., et al. (2015) Identification of Wnt pathway target genes regulating the division and differentiation of larval seam cells and vulval precursor cells in Caenorhabditis elegans. *G3: Genes, Genomes, Genetics*. [Online] 5 (8), 1551–1566. Available from: doi:10.1534/g3.115.017715.

Gorrepati, L., Thompson, K.W. & Eisenmann, D.M. (2013) *C. elegans* GATA factors EGL-18 and ELT-6 function downstream of Wnt signaling to maintain the progenitor fate during larval asymmetric divisions of the seam cells. *Development*. [Online] 140 (10), 2093–2102. Available from: doi:10.1242/dev.091124.

Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., et al. (2015) DNA methylation on N6-adenine in *C. elegans*. *Cell*. [Online] 161 (4), 868–878. Available from: doi:10.1016/j.cell.2015.04.005.

Gritti, N., Kienle, S., Filina, O. & van Zon, J.S. (2016) Long-term time-lapse microscopy of *C. elegans* post-embryonic development. *Nat Commun*. [Online] 7, 12500. Available from: doi:10.1038/ncomms12500.

Gross, D.S. & Garrard, W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annual Review of*

Biochemistry. [Online] 57 (1), 159–197. Available from: doi:10.1146/annurev.bi.57.070188.001111.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., et al. (2009) A Multiparameter Network Reveals Extensive Divergence between *C. elegans* bHLH Transcription Factors. *Cell*. [Online] 138 (2), 314–327. Available from: doi:10.1016/j.cell.2009.04.058.

Haerty, W., Artieri, C., Khezri, N., Singh, R.S., et al. (2008) Comparative analysis of function and interaction of transcription factors in nematodes: Extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics*. [Online] 9, 1–16. Available from: doi:10.1186/1471-2164-9-399.

Hahn, S. (2004) Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural and Molecular Biology*. [Online] 11 (5), 394–403. Available from: doi:10.1038/nsmb763.

Hajduskova, M., Jindra, M., Herman, M.A. & Asahina, M. (2009) The nuclear receptor NHR-25 cooperates with the Wnt/β-catenin asymmetry pathway to control differentiation of the T seam cell in *C. elegans*. *Journal of Cell Science*. [Online] 122 (17), 3051–3060. Available from: doi:10.1242/jcs.052373.

Halder, G., Callaerts, P. & Gehring, W.J. (1995) Induction of ectopic eyes by targeted expression of the eyeless gene in *Drosophila*. *Science*. [Online] 67 (5205), 1788–1792. Available from: doi:10.1126/science.7892602.

Harandi, O.F. & Ambros, V.R. (2015) Control of stem cell self-renewal and differentiation by the heterochronic genes and the cellular asymmetry machinery in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 112 (3), E287–E296. Available from: doi:10.1073/pnas.1422852112.

Harterink, M., Kim, D.H., Middelkoop, T.C., Doan, T.D., et al. (2011) Neuroblast migration along the anteroposterior axis of *C. elegans* is controlled by opposing gradients of Wnts and a secreted Frizzled-related protein. *Development*. [Online] 138 (14), 2915–2924. Available from: doi:10.1242/dev.064733.

Hayes, G.D., Frand, A.R. & Ruvkun, G. (2006) The *mir-84* and *let-7* paralogous microRNA genes of *Caenorhabditis elegans* direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. *Development*. [Online] 133 (23), 4631–4641. Available from: doi:10.1242/dev.02655.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. [Online] 459 (7243), 108–112. Available from: doi:10.1038/nature07829.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*. [Online] 39 (3), 311–318. Available from: doi:10.1038/ng1966.

Heinz, S., Benner, C., Spann, N., Bertolino, E., et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*. [Online] 38 (4), 576–589. Available from: doi:10.1016/j.molcel.2010.05.004.

Heppert, J.K., Dickinson, D.J., Pani, A.M., Higgins, C.D., et al. (2016) Comparative assessment of fluorescent proteins for in vivo imaging in an animal model system. *bioRxiv*. [Online]. 27. Available from: doi:10.1101/040279.

Hintze, M., Koneru, S.L., Gilbert, S.P.R., Katsanos, D., et al. (2020) A cell fate switch in the *Caenorhabditis elegans* seam cell lineage occurs through modulation of the Wnt asymmetry pathway in response to temperature increase. *Genetics*. [Online] 214 (4), 927–939. Available from: doi:10.1534/GENETICS.119.302896.

Ho, M.C.W., Quintero-Cadena, P. & Sternberg, P.W. (2017) Genome-wide discovery of active regulatory elements and transcription factor footprints in *Caenorhabditis elegans* using DNase-seq. *Genome Research*. [Online] 27 (12), 2108–2119. Available from: doi:10.1101/gr.223735.117.

Höflich, J., Berninsone, P., Göbel, C., Gravato-Nobre, M.J., et al. (2004) Loss of *srf-3*-encoded nucleotide sugar transporter activity in *Caenorhabditis elegans* alters surface antigenicity and prevents bacterial adherence. *Journal of Biological Chemistry*. [Online] 279 (29), 30440–30448. Available from: doi:10.1074/jbc.M402429200.

Hong, Y., Roy, R. & Ambros, V. (1998) Developmental regulation of a cyclin-dependent kinase inhibitor controls postembryonic cell cycle progression in *Caenorhabditis elegans*. *Development*. [Online] 125 (18), 3585–3597. Available from: https://dev.biologists.org/content/develop/125/18/3585.full.pdf.

Hope, I.A. (1991) 'Promoter trapping' in *Caenorhabditis elegans*. *Development (Cambridge, England)*. [Online] 113 (2), 399–408. Available from: http://dev.biologists.org/content/develop/113/2/399.full.pdf.

van der Horst, S.E.M., Cravo, J., Woollard, A., Teapal, J., et al. (2019) *C. elegans* Runx/CBFβ suppresses POP-1 TCF to convert asymmetric to proliferative division of stem cell-like seam cells. *Development (Cambridge, England)*. [Online] 146 (22), 1–14. Available from: doi:10.1242/dev.180034.

Horvitz, H.R. & Herskowitz, I. (1992) Mechanisms of asymmetric cell division: Two Bs or not two Bs, that is the question. *Cell*. [Online] 68 (2), 237–255. Available from: doi:10.1016/0092-

8674(92)90468-R.

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., et al. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*. [Online] 12. Available from: doi:10.1186/1471-2105-12-357.

Huang, X., Tian, E., Xu, Y. & Zhang, H. (2009) The *C. elegans engrailed* homolog ceh-16 regulates the self-renewal expansion division of stem cell-like seam cells. *Developmental Biology*. [Online] Available from: doi:10.1016/j.ydbio.2009.07.005.

Hughes, S., Brabin, C., Appleford, P.J. & Woollard, A. (2013) CEH-20/Pbx and UNC-62/Meis function upstream of *rnt-1*/Runx to regulate asymmetric divisions of the *C. elegans* stem-like seam cells. *Biology open*. [Online] 2 (7), 718–727. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3711040&tool=pmcentrez&rendertype= abstract.

Ishiguro, A., Nogi, Y., Hisatake, K., Muramatsu, M., et al. (2000) The Rpb6 Subunit of Fission Yeast RNA Polymerase II Is a Contact Target of the Transcription Elongation Factor TFIIS. *Molecular and Cellular Biology*. [Online] 20 (4), 1263–1270. Available from: doi:10.1128/mcb.20.4.1263-1270.2000.

Jaenisch, R. & Bird, A. (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*. [Online] 33 (3S), 245–254. Available from: doi:10.1038/ng1089.

Jamaladdin, S., Kelly, R.D.W., O'Regan, L., Dovey, O.M., et al. (2014) Histone deacetylase (HDAC) 1 and 2 are essential for accurate cell division and the pluripotency of embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 111 (27), 9840–9845. Available from: doi:10.1073/pnas.1321330111.

Jan, C.H., Friedman, R.C., Ruby, J.G. & Bartel, D.P. (2011) Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*. [Online] 469 (7328), 97–103. Available from: doi:10.1038/nature09616.

Jänes, J., Dong, Y., Schoof, M., Serizay, J., et al. (2018) Chromatin accessibility dynamics across *C. elegans* development and ageing. *eLife*. [Online] 7 (1), 267–272. Available from: doi:10.7554/eLife.37344.

Johnstone, I.L., Shafi, Y. & Barry, J.D. (1992) Molecular analysis of mutations in the *Caenorhabditis elegans* collagen gene *dpy-7*. *The EMBO journal*. [Online] 11 (11), 3857–3863. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=556895&tool=pmcentrez&rendertype=a bstract.

Jones, E., Kimura, H., Vigneron, M., Wang, Z., et al. (2000) Isolation and characterization of monoclonal antibodies directed against subunits of human RNA polymerases I, II, and III. *Experimental Cell Research*. [Online] 254 (1), 163–172. Available from: doi:10.1006/excr.1999.4739.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics*. [Online] 30 (9), 1236–1240. Available from: doi:10.1093/bioinformatics/btu031.

Jorgensen, E.M. & Mango, S.E. (2002) The art and design of genetic screens: *Caenorhabditis elegans*. *Nature Reviews Genetics*. [Online] 3 (5), 356–369. Available from: doi:10.1038/nrg794.

Jose, A.M. & Hunter, C.P. (2007) Transport of sequence-specific RNA interference information between cells. *Annual Review of Genetics*. [Online] 41, 305–330. Available from: doi:10.1146/annurev.genet.41.110306.130216.

Jose, A.M., Smith, J.J. & Hunter, C.P. (2009) Export of RNA silencing from C. elegans tissues does not require the RNA channel SID-1. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 106 (7), 2283–2288. Available from: doi:10.1073/pnas.0809760106.

Joshi, P.M., Riddle, M.R., Djabrayan, N.J. V & Rothman, J.H. (2010) *Caenorhabditis elegans* as a model for stem cell biology. *Developmental Dynamics*. [Online] 239 (5), 1539–1554. Available from: doi:10.1002/dvdy.22296.

Kageyama, R. & Ohtsuka, T. (1999) The Notch-Hes pathway in mammalian neural development. *Cell Research*. [Online] 9 (3), 179–188. Available from: doi:10.1038/sj.cr.7290016.

Kageyama, R., Ohtsuka, T. & Kobayashi, T. (2007) The Hes gene family: Repressors and oscillators that orchestrate embryogenesis. *Development*. [Online] 134 (7), 1243–1251. Available from: doi:10.1242/dev.000786.

Kagoshima, H., Nimmo, R., Saad, N., Tanaka, J., et al. (2007) The *C. elegans* CBF homologue BRO-1 interacts with the Runx factor, RNT-1, to promote stem cell proliferation and self-renewal. *Development*. [Online] 134 (21), 3905–3915. Available from: doi:10.1242/dev.008276.

Kakebeen, A.D., Chitsazan, A.D., Williams, M.C., Saunders, L.M., et al. (2020) Chromatin accessibility dynamics and single cell RNA-seq reveal new regulators of regeneration in neural progenitors. *eLife*. [Online] 9, 1–31. Available from: doi:10.7554/eLife.52648.

Kaletsky, R., Yao, V., Williams, A., Runnels, A.M., et al. (2018) Transcriptome analysis of adult *Caenorhabditis elegans* cells reveals tissue-specific gene and isoform expression. *PLoS Genetics*. [Online] 14 (8). Available from: doi:10.1371/journal.pgen.1007559.

Kamath, R.S. & Ahringer, J. (2003) Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods*.

[Online] 30 (4), 313–321. Available from: doi:10.1016/S1046-2023(03)00050-1.

Katsanos, D., Koneru, S.L., Mestek Boukhibar, L., Gritti, N., et al. (2017) Stochastic loss and gain of symmetric divisions in the *C. elegans* epidermis perturbs robustness of stem cell number. *PLoS Biology*. [Online] 15 (11), 1–31. Available from: doi:10.1371/journal.pgen.1006278.

Kay, S.K., Harrington, H.A., Shepherd, S., Brennan, K., et al. (2017) The role of the Hes1 crosstalk hub in Notch-Wnt interactions of the intestinal crypt. *PLoS Computational Biology*. [Online] 13 (2), 1–28. Available from: doi:10.1371/journal.pcbi.1005400.

Khan, A. & Mathelier, A. (2017) Intervene: A tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*. [Online] 18 (1), 1–8. Available from: doi:10.1186/s12859-017-1708-7.

Kim, W., Underwood, R.S., Greenwald, I. & Shaye, D.D. (2018) Ortholist 2: A new comparative genomic analysis of human and *Caenorhabditis elegans* genes. *Genetics*. [Online] 210 (2), 445–461. Available from: doi:10.1534/genetics.118.301307.

Kimble, J. & Hirsh, D. (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Developmental Biology*. [Online] 70 (2), 396–417. Available from: doi:10.1016/0012-1606(79)90035-6 [Accessed: 27 June 2017].

Klein, A.M. & Simons, B.D. (2011) Universal patterns of stem cell fate in cycling adult tissues. *Development (Cambridge, England)*. [Online] 138 (15), 3103–3111. Available from: doi:10.1242/dev.060103.

Klemm, S.L., Shipony, Z. & Greenleaf, W.J. (2019) Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*. [Online] 20 (4), 207–220. Available from: doi:10.1038/s41576-018-0089-8.

Knoblich, J.A. (2008) Mechanisms of asymmetric stem cell division. *Cell*. [Online] 132 (4), 583–597. Available from: doi:10.1016/j.cell.2008.02.007.

Kobayashi, T., Mizuno, H., Imayoshi, I., Furusawa, C., et al. (2009) The cyclic gene Hes1 contributes to diverse differentiation responses of embryonic stem cells. *Genes and Development*. [Online] Available from: doi:10.1101/gad.1823109.

Koh, K. & Rothman, J.H. (2001) ELT-5 and ELT-6 are required continuously to regulate epidermal seam cell differentiation and cell fusion in *C. elegans*. *Development (Cambridge, England)*. [Online] 128, 2867–2880. Available from: http://dev.biologists.org/content/develop/128/15/2867.full.pdf.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., et al. (2015) The technology and biology of single-cell RNA sequencing. *Molecular Cell*. [Online] 58 (4), 610–620. Available from:

doi:10.1016/j.molcel.2015.04.005.

Kolundzic, E., Ofenbauer, A., Bulut, S.I., Uyar, B., et al. (2018) FACT sets a barrier for cell fate reprogramming in *Caenorhabditis elegans* and human cells. *Developmental Cell*. [Online] 46 (5), 611-626.e12. Available from: doi:10.1016/j.devcel.2018.07.006.

Kondili, M., Fust, A., Preussner, J., Kuenne, C., et al. (2017) UROPA: A tool for Universal RObust Peak Annotation. *Scientific Reports*. [Online] 7 (1), 1–12. Available from: doi:10.1038/s41598-017-02464-y.

Kozak, M. (2001) Constraints on reinitiation of translation in mammals. *Nucleic Acids Research*. [Online] 29 (24), 5226–5232. Available from: doi:10.1093/nar/29.24.5226.

Kozak, M. (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and cellular biology*. [Online] 7 (10), 3438–3445. Available from: doi:10.1128/MCB.7.10.3438.Updated.

Krüger, A. V., Jelier, R., Dzyubachyk, O., Zimmerman, T., et al. (2015) Comprehensive single cell-resolution analysis of the role of chromatin regulators in early *C. elegans* embryogenesis. *Developmental Biology*. [Online] 398 (2), 153–162. Available from: doi:10.1016/j.ydbio.2014.10.014.

Kudron, M.M., Victorsen, A., Gevirtzman, L., Hillier, L.W., et al. (2018) The modern resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics*. [Online] 208 (3), 937–949. Available from: doi:10.1534/genetics.117.300657.

Kuntz, S.G., Schwarz, E.M., DeModena, J.A., De Buysscher, T., et al. (2008) Multigenome DNA sequence conservation identifies Hox *cis*-regulatory elements. *Genome Research*. [Online] 18 (12), 1955–1968. Available from: doi:10.1101/gr.085472.108.

Lam, A.K. & Phillips, B.T. (2017) Wnt signaling polarizes *C. elegans* asymmetric cell divisions during development. In: *Results and Problems in Cell Differentiation*. [Online]. pp. 83–114. Available from: doi:10.1007/978-3-319-53150-2_4.

Lammens, T., Li, J., Leone, G. & De Veylder, L. (2009) Atypical E2Fs: new players in the E2F transcription factor family. *Trends in Cell Biology*. [Online] 19 (3), 111–118. Available from: doi:10.1016/j.tcb.2009.01.002.

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*. [Online] 9 (4), 357–359. Available from: doi:10.1038/nmeth.1923.

Lau, N.C., Lim, L.P., Weinstein, E.G. & Bartel, D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*. [Online] 294 (5543), 858–862.

Available from: doi:10.1126/science.1065062.

Law, J.A. & Jacobsen, S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*. [Online] 11 (3), 204–220. Available from: doi:10.1038/nrg2719.

Lee, R.C., Feinbaum, R.L. & Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*. [Online] 75 (5), 843–854. Available from: doi:10.1016/0092-8674(93)90529-Y.

Levine, M. & Davidson, E.H. (2005) Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 102 (14), 4936–4942. Available from: doi:10.1073/pnas.0408031102.

Lewis, B.P., Burge, C.B. & Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. [Online] 120 (1), 15–20. Available from: doi:10.1016/j.cell.2004.12.035.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. [Online] 25 (16), 2078–2079. Available from: doi:10.1093/bioinformatics/btp352.

Lichtenberg, K.H. de, Funa, N., Nakic, N., Ferrer, J., et al. (2018) Genome-Wide Identification of HES1 Target Genes Uncover Novel Roles for HES1 in Pancreatic Development. *bioRxiv*. [Online] 335869. Available from: doi:10.1101/335869.

Lin, G.G. & Scott, J.G. (2012) FLP/FRT and Cre/lox recombination technology in *C. elegans*. *Methods*. [Online] 100 (2), 130–134. Available from: doi:10.1016/j.pestbp.2011.02.012.Investigations.

Lin, R., Hill, R.J. & Priess, J.R. (1998) POP-1 and anterior-posterior fate decisions in *C. elegans* embryos. *Cell*. [Online] 92 (2), 229–239. Available from: doi:10.1016/S0092-8674(00)80917-4.

Lints, R. & Hall, D.H. (2004) WormAtlas Hermaphrodite Handbook - Reproductive System - Egg-laying Apparatus. *WormAtlas*. [Online] Available from: doi:10.3908/wormatlas.1.24.

Liu, Z., Kirch, S. & Ambros, V. (1995) The *Caenorhabditis elegans* heterochronic gene pathway controls stage-specific transcription of collagen genes. *Development*. 121 (8), 2471–2478.

Loots, G.G. & Ovcharenko, I. (2004) rVISTA 2.0: Evolutionary analysis of transcription factor binding sites. *Nucleic Acids Research*. [Online] 32 (WEB SERVER ISS.), 217–221. Available from: doi:10.1093/nar/gkh383.

Lozano, E., Sáez, A.G., Flemming, A.J., Cunha, A., et al. (2006) Regulation of growth by ploidy in *Caenorhabditis elegans*. *Current Biology*. [Online] 16 (5), 493–498. Available from:

doi:10.1016/j.cub.2006.01.048.

Lu, C., Meyers, B.C. & Green, P.J. (2007) Construction of small RNA cDNA libraries for deep sequencing. *Methods*. [Online] 43 (2), 110–117. Available from: doi:10.1016/j.ymeth.2007.05.002.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., et al. (2005) Elucidation of the small RNA component of the transcriptome. *Science*. [Online] 309 (5740), 1567–1569. Available from: doi:10.1126/science.1114112.

Mallick, A., Ranawade, A. & Gupta, B.P. (2019) Role of PRY-1/Axin in heterochronic miRNA-mediated seam cell development. *BMC Developmental Biology*. [Online] 19 (1), 1–12. Available from: doi:10.1186/s12861-019-0197-5.

Marioni, J.C. & Arendt, D. (2017) How single-cell genomics is changing evolutionary and developmental biology. *Annual Review of Cell and Developmental Biology*. [Online] 33, 537–553. Available from: doi:10.1146/annurev-cellbio-100616-060818.

Marshall, O.J. & Brand, A.H. (2015) Damidseq-pipeline: An automated pipeline for processing DamID sequencing datasets. *Bioinformatics*. [Online] 31 (20), 3371–3373. Available from: doi:10.1093/bioinformatics/btv386.

Marshall, O.J., Southall, T.D., Cheetham, S.W. & Brand, A.H. (2016) Cell-type-specific profiling of protein-DNA interactions without cell isolation using targeted DamID with next-generation sequencing. *Nature protocols*. [Online] 11 (9), 1586–1598. Available from: doi:10.1038/nprot.2016.084.

Martinez, N.J., Ow, M.C., Reece-Hoyes, J.S., Barrasa, M.I., et al. (2008) Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Research*. [Online] 18 (12), 2005–2015. Available from: doi:10.1101/gr.083055.108.

Mathies, L.D., Ray, S., Lopez-Alvillar, K., Arbeitman, M.N., et al. (2019) mRNA profiling reveals significant transcriptional differences between a multipotent progenitor and its differentiated sister. *BMC Genomics*. [Online] 20 (1), 1–15. Available from: doi:10.1186/s12864-019-5821-z.

Matus, D.Q., Lohmer, L.L., Kelley, L.C., Schindler, A.J., et al. (2015) Invasive Cell Fate Requires G1 Cell-Cycle Arrest and Histone Deacetylase-Mediated Changes in Gene Expression. *Developmental Cell*. [Online] 35 (2), 162–174. Available from: doi:10.1016/j.devcel.2015.10.002.

McClure, C.D. & Southall, T.D. (2015) Getting down to specifics: Profiling gene expression and protein-DNA interactions in a cell type-specific manner. *Advances in Genetics*. [Online] 91, 103–151. Available from: doi:10.1016/bs.adgen.2015.06.003.

McGinnis, W., Garber, R.L., Wirz, J., Kuroiwa, A., et al. (1984) A homologous protein-coding sequence

in drosophila homeotic genes and its conservation in other metazoans. *Cell*. [Online] 37 (2), 403–408. Available from: doi:10.1016/0092-8674(84)90370-2.

Mello, C.C., Kramer, J.M., Stinchcomb, D. & Ambros, V. (1992) Efficient gene transfer in *C. elegans*: extrachromosomal maintenance and integration of transforming sequences. *Trends in Genetics*. [Online] 8 (2), 50. Available from: doi:10.1016/0168-9525(92)90342-2.

Mercer, T.R., Edwards, S.L., Clark, M.B., Neph, S.J., et al. (2013) DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genetics*. [Online] 45 (8), 852–859. Available from: doi:10.1038/ng.2677.

Meyer, C.A. & Liu, X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*. [Online] 15 (11), 709–721. Available from: doi:10.1038/nrg3788.

Michaux, G., Legouis, R. & Labouesse, M. (2001) Epithelial biology: Lessons from Caenorhabditis elegans. *Gene*. [Online] 277 (1–2), 83–100. Available from: doi:10.1016/S0378-1119(01)00700-4.

Minakhin, L., Bhagat, S., Brunning, A., Campbell, E.A., et al. (2001) Bacterial RNA polymerase subunit ω and eukaryotic polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 98 (3), 892–897. Available from: doi:10.1073/pnas.98.3.892.

Miska, E.A., Alvarez-Saavedra, E., Abbott, A.L., Lau, N.C., et al. (2007) Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. *PLoS Genetics*. [Online] 3 (12), 2395–2403. Available from: doi:10.1371/journal.pgen.0030215.

Mitxelena, J., Apraiz, A., Vallejo-Rodriguez, J., Malumbres, M., et al. (2016) E2F7 regulates transcription and maturation of multiple microRNAs to restrain cell proliferation. *Nucleic Acids Research*. [Online] 44 (12), 5557–5570. Available from: doi:10.1093/nar/gkw146.

Miyabayashi, T., Palfreyman, M.T., Sluder, a E., Slack, F., et al. (1999) Expression and function of members of a divergent nuclear receptor family in *Caenorhabditis elegans*. *Developmental biology*. [Online] 215 (2), 314–331. Available from: doi:10.1006/dbio.1999.9470.

Mizumoto, K. & Sawa, H. (2007a) Cortical β-catenin and APC regulate asymmetric nuclear β-catenin localization during asymmetric cell division in *C. elegans*. *Developmental Cell*. [Online] 12 (2), 287–299. Available from: doi:10.1016/j.devcel.2007.01.004.

Mizumoto, K. & Sawa, H. (2007b) Two βs or not two βs: regulation of asymmetric division by β-catenin. *Trends in Cell Biology*. [Online] 17 (10), 465–473. Available from: doi:10.1016/j.tcb.2007.08.004.

Morrison, S. & Kimble, J. (2006) Asymmetric and symmteric division in development and cancer. *Nature*

Review. [Online] 441 (29 June 2006), 1068–1074. Available from: https://www.nature.com/articles/nature04956.pdf.

Morrison, S.J. & Spradling, A.C. (2008) Stem cells and niches: Mechanisms that promote stem cell maintenance throughout life. *Cell*. [Online] 132 (4), 598–611. Available from: doi:10.1016/j.cell.2008.01.038.

Murata, K., Hattori, M., Hirai, N., Shinozuka, Y., et al. (2005) Hes1 directly controls cell proliferation through the transcriptional repression of p27Kip1. *Molecular and Cellular Biology*. [Online] 25 (10), 4262–4271. Available from: doi:10.1128/mcb.25.10.4262-4271.2005.

Nance, J. & Frøkjær-Jensen, C. (2019) The Caenorhabditis elegans transgenic toolbox. *Genetics*. [Online] 212 (4), 959–990. Available from: doi:10.1534/genetics.119.301506.

Neumüller, R.A. & Knoblich, J.A. (2009) Dividing cellular asymmetry: Asymmetric cell division and its implications for stem cells and cancer. *Genes and Development*. [Online] 23 (23), 2675–2699. Available from: doi:10.1101/gad.1850809.

Newman, S.A. (2020) Cell differentiation: What have we learned in 50 years? *Journal of Theoretical Biology*. [Online] 485. Available from: doi:10.1016/j.jtbi.2019.110031.

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., et al. (2018) RSAT 2018: Regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*. [Online] 46 (W1), W209–W214. Available from: doi:10.1093/nar/gky317.

Nimmo, R., Antebi, A. & Woollard, A. (2005) *mab-2* encodes RNT-1, a *C. elegans* Runx homologue essential for controlling cell proliferation in a stem cell-like developmental lineage. *Development*. [Online] 132 (22), 5043–5054. Available from: doi:10.1242/dev.02102.

Nimmo, R. & Woollard, A. (2008) Worming out the biology of Runx. *Developmental Biology*. [Online] 313 (2), 492–500. Available from: doi:10.1016/j.ydbio.2007.11.002.

Nimmo, R.A. & Slack, F.J. (2009) An elegant miRror: microRNAs in stem cells, developmental timing and cancer. *Chromosoma*. [Online] 118 (4), 405–418. Available from: doi:10.1007/s00412-009-0210-z.

Noonan, J.P. & McCallion, A.S. (2010) Genomics of long-range regulatory elements. *Annual Review of Genomics and Human Genetics*. [Online] 11 (1), 1–23. Available from: doi:10.1146/annurev-genom-082509-141651.

O'Connell, K.F. (2010) Reverse chunking, a simple and effective method for identifying *unc-119(+)* transgenics. *The Worm Breeder's Gazette*. Volume 18 (Number 2).

Ohsako, S., Hyer, J., Panganiban, G., Oliver, I., et al. (1994) Hairy function as a DNA-binding helix-

loop-helix repressor of *Drosophila* sensory organ formation. *Genes and Development*. [Online] 8 (22), 2743–2755. Available from: doi:10.1101/gad.8.22.2743.

Oliveri, P. & Davidson, E.H. (2007) Built to run, not fail. *Science*. [Online] 315 (5818), 1510–1511. Available from: doi:10.1126/science.1140979.

Ozsolak, F. & Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*. [Online] 12 (2), 87–98. Available from: doi:10.1038/nrg2934.

Page, A. (2007) The cuticle. *WormBook*. [Online] 1–15. Available from: doi:10.1895/wormbook.1.138.1.

Page, B.D., Zhang, W., Steward, K., Blumenthal, T., et al. (1997) ELT-1, a GATA-like transcription factor, is required for epidermal cell fates in *Caenorhabditis elegans* embryos. *Genes and Development*. [Online] 11 (13), 1651–1661. Available from: doi:10.1101/gad.11.13.1651.

Pani, A.M. & Goldstein, B. (2018) Direct visualization of a native Wnt in vivo reveals that a long-range Wnt gradient forms by extracellular dispersal. *eLife*. [Online] 7, 1–22. Available from: doi:10.7554/eLife.38325.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*. [Online] 444 (7118), 499–502. Available from: doi:10.1038/nature05295.

Pfeiffer, B.D., Truman, J.W. & Rubin, G.M. (2012) Using translational enhancers to increase transgene expression in Drosophila. *Proceedings of the National Academy of Sciences*. [Online] 109 (17), 6626–6631. Available from: doi:10.1073/pnas.1204520109.

Piano, F., Gunsalus, K.C., Hill, D.E. & Vidal, M. (2006) *C. elegans* network biology: a beginning. *WormBook : the online review of C. elegans biology*. [Online] 1–20. Available from: doi:10.1895/wormbook.1.118.1.

Pindyurin, A. V., Pagie, L., Kozhevnikova, E.N., Van Arensbergen, J., et al. (2016) Inducible DamID systems for genomic mapping of chromatin proteins in *Drosophila*. *Nucleic Acids Research*. [Online] 44 (12), 5646–5657. Available from: doi:10.1093/nar/gkw176.

Podbilewicz, B. (2006) Cell fusion. *WormBook : the online review of C. elegans biology*. [Online] 1–32. Available from: doi:10.1895/wormbook.1.52.1.

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. [Online] 26 (6), 841–842. Available from: doi:10.1093/bioinformatics/btq033.

Ramialison, M., Waardenberg, A.J., Schonrock, N., Doan, T., et al. (2017) Analysis of steric effects in DamID profiling of transcription factor target genes. *Genomics*. [Online] 109 (2), 75–82. Available from: doi:10.1016/j.ygeno.2017.01.006.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., et al. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*. [Online] 44 (W1), W160–W165. Available from: doi:10.1093/nar/gkw257.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquienelll, A.E., et al. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*. [Online] 403 (6772), 901–906. Available from: doi:10.1038/35002607.

Reyes, A. & Huber, W. (2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*. [Online] 46 (2), 582–592. Available from: doi:10.1093/nar/gkx1165.

van Rijnberk, L.M., van der Horst, S.E.M., van den Heuvel, S. & Ruijtenberg, S. (2017) A dual transcriptional reporter and CDK-activity sensor marks cell cycle entry and progression in C. elegans. *Plos One*. [Online] 12 (2), e0171600. Available from: doi:10.1371/journal.pone.0171600.

RM, T., S, C., PL, K.-C. & JH, R. (1995) A Deficiency Screen To Identify Loci Required for Establishment and Patterning of the Epidermis. *International C. elegans Meeting*. [Online] Available from: https://www.genetics.org/content/genetics/146/1/185.full.pdf.

Rougvie, A.E. & Ambros, V. (1995) The heterochronic gene *lin-29* encodes a zinc finger protein that controls a terminal differentiation event in *Caenorhabditis elegans*. *Development*. 121 (8), 2491–2500.

Sánchez-Romero, M.A., Cota, I. & Casadesús, J. (2015) DNA methylation in bacteria: From the methyl group to the methylome. *Current Opinion in Microbiology*. [Online] 25, 9–16. Available from: doi:10.1016/j.mib.2015.03.004.

Sapir, A., Choi, J., Leikina, E., Avinoam, O., et al. (2007) AFF-1, a FOS-1-regulated fusogen, mediates fusion of the anchor cell in *C. elegans*. *Developmental Cell*. [Online] 12 (5), 683–698. Available from: doi:10.1016/j.devcel.2007.03.003.

Sasai, Y., Kageyama, R., Tagawa, Y., Shigemoto, R., et al. (1992) Two mammalian helix-loop-helix factors structurally related to *Drosophila hairy* and *Enhancer of split*. *Genes and Development*. [Online] 6 (12 PART B), 2620–2634. Available from: doi:10.1101/gad.6.12b.2620.

Sawa, H. & Korswagen, H.C. (2013) Wnt signaling in *C. elegans*. *WormBook*. [Online] 1–18. Available from: doi:10.1895/wormbook.1.

Schlager, B., Röseler, W., Zheng, M., Gutierrez, A., et al. (2006) HAIRY-like transcription factors and the evolution of the nematode vulva equivalence group. *Current Biology*. [Online] 16 (14), 1386–1394. Available from: doi:10.1016/j.cub.2006.06.058.

Schulenburg, H. & Félix, M.A. (2017) The natural biotic environment of *Caenorhabditis elegans*. *Genetics*. [Online] 206 (1), 55–86. Available from: doi:10.1534/genetics.116.195511.

Schuster, E., McElwee, J.J., Tullet, J.M.A., Doonan, R., et al. (2010) DamID in *C. elegans* reveals longevity-associated targets of DAF-16/FoxO. *Molecular Systems Biology*. [Online] 6 (399), 1–6. Available from: doi:10.1038/msb.2010.54.

Schwarz, E.M., Kato, M. & Sternberg, P.W. (2012) Functional transcriptomics of a migrating cell in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 109 (40), 16246–16251. Available from: doi:10.1073/pnas.1203045109.

Sen, S.Q., Chanchani, S., Southall, T.D. & Doe, C.Q. (2019) Neuroblast-specific chromatin landscapes allows the integration of spatial and temporal cues during *Drosophila* neurogenesis. *eLife*. [Online] 8, 399–404. Available from: doi:10.7554/eLife.44036.

Sha, K., Gu, S.G., Pantalena-Filho, L.C., Goh, A., et al. (2010) Distributed probing of chromatin structure in vivo reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in *C. elegans*. *BMC Genomics*. [Online] 11 (1), 1–15. Available from: doi:10.1186/1471-2164-11-465.

Shao, J., He, K., Wang, H., Ho, W.S., et al. (2013) Collaborative regulation of development but independent control of metabolism by two epidermis-specific transcription factors in *Caenorhabditis elegans*. *Journal of Biological Chemistry*. [Online] 288 (46), 33411–33426. Available from: doi:10.1074/jbc.M113.487975.

Shpakovski, G. V, Acker, J., Wintzerith, M., Lacroix, J.F., et al. (1995) Four subunits that are shared by the three classes of RNA polymerase are functionally interchangeable between *Homo sapiens* and *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*. [Online] 15 (9), 4702–4710. Available from: doi:10.1128/mcb.15.9.4702.

Šilhánková, M., Jindra, M. & Asahina, M. (2005) Nuclear receptor NHR-25 is required for cell-shape dynamics during epidermal differentiation in *Caenorhabditis elegans*. *Journal of Cell Science*. [Online] 118 (1), 223–232. Available from: doi:10.1242/jcs.01609.

Slack, F. & Ruvkun, G. (1997) Temporal pattern formation by heterochronic genes. *Annual Review of Genetics*. [Online] 31 (1), 611–634. Available from: doi:10.1146/annurev.genet.31.1.611.

Slack, J.M.W. (2006) *Essential Developmental Biology*. 2nd edition. Blackwell.

Smith, J.A., McGarr, P. & Gilleard, J.S. (2005) The *Caenorhabditis elegans* GATA factor *elt-1* is essential for differentiation and maintenance of hypodermal seam cells and for normal locomotion. *Journal of Cell Science*. [Online] 118 (24), 5709–5719. Available from: doi:10.1242/jcs.02678.

Solari, F. & Ahringer, J. (2000) NURD-complex genes antagonise Ras-induced vulval development in *Caenorhabditis elegans*. *Current Biology*. [Online] 10 (4), 223–226. Available from: doi:10.1016/S0960-9822(00)00343-2.

Solomon, M.J. & Varshavsky, A. (1985) Formaldehyde-mediated DNA-protein crosslinking: A probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 82 (19), 6470–6474. Available from: doi:10.1073/pnas.82.19.6470.

Song, L. & Crawford, G.E. (2010) DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*. [Online] 5 (2), 1–12. Available from: doi:10.1101/pdb.prot5384.

Southall, T., Gold, K., Egger, B., Davidson, C., et al. (2013) Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: Assaying RNA pol II occupancy in neural stem cells. *Developmental Cell*. [Online] Available from: doi:10.1016/j.devcel.2013.05.020.

Southall, T.D., Davidson, C.M., Miller, C., Carr, A., et al. (2014) Dedifferentiation of neurons precedes tumor formation in lola mutants. *Developmental Cell*. [Online] 28 (6), 685–696. Available from: doi:10.1016/j.devcel.2014.01.030.

Spencer, W.C., McWhirter, R., Miller, T., Strasbourger, P., et al. (2014) Isolation of specific neurons from *C. elegans* larvae for gene expression profiling. *PLoS ONE*. [Online] 9 (11), 1–11. Available from: doi:10.1371/journal.pone.0112102.

Spencer, W.C., Zeller, G., Watson, J.D., Henz, S.R., et al. (2011) A spatial and temporal map of *C. elegans* gene expression. *Genome Research*. [Online] 21 (2), 325–341. Available from: doi:10.1101/gr.114595.110.

Spieth, J., Hillier, L.W. & Wilson, R.K. (2005) Evolutionarily conserved elements in vertebrate , insect , worm , and yeast genomes. *Genome Research*. [Online] 1034–1050. Available from: doi:10.1101/gr.3715005.

Spieth, J., Lawson, D., Davis, P., Williams, G., et al. (2014) Overview of gene structure in *C. elegans*. *WormBook : the online review of C. elegans biology*. [Online] 1–18. Available from: doi:10.1895/wormbook.1.65.2.

Spitz, F. & Furlong, E.E.M. (2012) Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*. [Online] 13 (9), 613–626. Available from: doi:10.1038/nrg3207.

St Johnston, D. (2015) The Renaissance of Developmental Biology. *PLoS Biology*. [Online] 13 (5), 1–9. Available from: doi:10.1371/journal.pbio.1002149.

Stathopoulos, A. & Levine, M. (2005) Genomic regulatory networks and animal development.

Developmental Cell. [Online] 9 (4), 449–462. Available from: doi:10.1016/j.devcel.2005.09.005.

van Steensel, B., Delrow, J. & Henikoff, S. (2001) Chromatin profiling using targeted DNA adenine methyltransferase. *Nature genetics*. [Online] 27 (3), 304–308. Available from: doi:10.1038/85871.

van Steensel, B. & Henikoff, S. (2000) Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature biotechnology*. [Online] 18 (4), 424–428. Available from: doi:10.1038/74487.

Di Stefano, L., Jensen, M.R. & Helin, K. (2003) E2F7, a novel E2F featuring DP-independent repression of a subset of E2F-regulated genes. *EMBO Journal*. [Online] 22 (23), 6289–6298. Available from: doi:10.1093/emboj/cdg613.

Steiner, F.A., Talbert, P.B., Kasinathan, S., Deal, R.B., et al. (2012) Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Research*. [Online] 22 (4), 766–777. Available from: doi:10.1101/gr.131748.111.

Stempor, P. & Ahringer, J. (2016) SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome Open Research*. [Online] 1 (0), 14. Available from: doi:10.12688/wellcomeopenres.10004.1.

Von Stetina, S.E., Watson, J.D., Fox, R.M., Olszewski, K.L., et al. (2007) Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the *C. elegans* nervous system. *Genome Biology*. [Online] 8 (7), R135. Available from: doi:10.1186/gb-2007-8-7-r135.

Stiernagle, T. (2006) Maintenance of C. elegans. *WormBook*. [Online] (1999), 1–11. Available from: doi:10.1895/wormbook.1.101.1.

Suggs, B.Z., Latham, A.L., Dawes, A.T. & Chamberlin, H.M. (2018) FACT complex gene duplicates exhibit redundant and non-redundant functions in *C. elegans*. *Developmental Biology*. [Online] 444 (2), 71–82. Available from: doi:10.1016/j.ydbio.2018.10.002.

Sugioka, K., Mizumoto, K. & Sawa, H. (2011) Wnt regulates spindle asymmetry to generate asymmetric nuclear β-catenin in *C. elegans*. *Cell*. [Online] 146 (6), 942–954. Available from: doi:10.1016/j.cell.2011.07.043.

Sulston, J.E., Albertson, D.G. & Thomson, J.N. (1980) The *Caenorhabditis elegans* male: Postembryonic development of nongonadal structures. *Developmental Biology*. [Online] 78 (2), 542–576. Available from: doi:10.1016/0012-1606(80)90352-8.

Sulston, J.E. & Horvitz, H.R. (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental Biology*. [Online] 56 (1), 110–156. Available from: doi:10.1016/0012-1606(77)90158-0.

Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*. [Online] 100 (1), 64–119. Available from: doi:10.1016/0012-1606(83)90201-4 [Accessed: 27 June 2017].

Takebayashi, K., Sasai, Y., Sakai, Y., Watanabe, T., et al. (1994) Structure, chromosomal locus, and promoter analysis of the gene encoding the mouse helix-loop-helix factor HES-1. Negative autoregulation through the multiple N box elements. *Journal of Biological Chemistry*. [Online] 269 (7), 5150–5156. Available from: http://magazine.good.is/articles/chat-history.

Takeshita, H. & Sawa, H. (2005) Asymmetric cortical and nuclear localizations of WRM-1/β-catenin during asymmetric cell division in *C. elegans*. *Genes and Development*. [Online] 19 (15), 1743–1748. Available from: doi:10.1101/gad.1322805.

Tavernarakis, N., Wang, S.L., Dorovkov, M., Ryazanov, A., et al. (2000) Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nature genetics*. [Online] 24 (2), 180–183. Available from: doi:10.1038/72850.

Teuscher, A.C., Jongsma, E., Davis, M.N., Statzer, C., et al. (2019) The *in-silico* characterization of the *Caenorhabditis elegans* matrisome and proposal of a novel collagen classification. *Matrix Biology Plus*. [Online] 1, 100001. Available from: doi:10.1016/j.mbplus.2018.11.001.

Teytelman, L., Thurtle, D.M., Rine, J. & Van Oudenaarden, A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 110 (46), 18602–18607. Available from: doi:10.1073/pnas.1316064110.

Thompson, K.W., Joshi, P., Dymond, J.S., Gorrepati, L., et al. (2016) The Paired-box protein PAX-3 regulates the choice between lateral and ventral epidermal cell fates in *C. elegans*. *Developmental Biology*. [Online] 412 (2), 191–207. Available from: doi:10.1016/j.ydbio.2016.03.002.

Timmons, L. & Fire, A. (1998) Specific interference by ingested dsRNA. *Nature*. [Online] 395 (October), 854. Available from: https://www.nature.com/articles/27579.pdf.

Timmons, L., Tabara, H., Mello, C.C. & Fire, A.Z. (2003) Inducible systemic RNA silencing in *Caenorhabditis elegans*. *Molecular Biology of the Cell*. [Online] 14 (7), 2972–2983. Available from: doi:10.1091/mbc.E03-01-0858.

Tourasse, N.J., Millet, J.R.M. & Dupuy, D. (2017) Quantitative RNA-seq meta-analysis of alternative exon usage in C. elegans. *Genome Research*. [Online] 27 (12), 2120–2128. Available from: doi:10.1101/gr.224626.117.

Tsompana, M. & Buck, M.J. (2014) Chromatin accessibility: A window into the genome. *Epigenetics and*

Chromatin. [Online] 7 (1), 1–16. Available from: doi:10.1186/1756-8935-7-33.

Vissers, J.H.A., Froldi, F., Schröder, J., Papenfuss, A.T., et al. (2018) The Scalloped and Nerfin-1 transcription factors cooperate to maintain neuronal cell fate. *Cell Reports*. [Online] 25 (6), 1561-1576.e7. Available from: doi:10.1016/j.celrep.2018.10.038.

Wacker, I., Schwarz, V., Hedgecock, E.M. & Hutter, H. (2003) *zag-1*, a Zn-finger homeodomain transcription factor controlling neuronal differentiation and axon outgrowth in *C. elegans*. *Development*. [Online] 130 (16), 3795–3805. Available from: doi:10.1242/dev.00570.

Waddington, C.H. (1942) Canalization of development and the inheritance of acquired characters. *Nature*. [Online] 150 (3811), 563–565. Available from: doi:10.1038/150563a0.

Wang, P., Wang, Z. & Liu, J. (2020) Role of HDACs in normal and malignant hematopoiesis. *Molecular Cancer*. [Online] 19 (1), 1–21. Available from: doi:10.1186/s12943-019-1127-7.

Wang, X., Liu, M., Li, W., Suh, C.D., et al. (2009) The function of a spindle checkpoint gene *bub-1* in *C. elegans* development. *PLoS ONE*. [Online] 4 (6). Available from: doi:10.1371/journal.pone.0005912.

Waring, D. a & Kenyon, C. (1990) Selective silencing of cell communication influences anteroposterior pattern formation in *C. elegans*. *Cell*. [Online] 60, 123–131. Available from: doi:0092-8674(90)90722-Q [pii].

Waring, D.A., Wrischnik, L. & Kenyon, C. (1992) Cell signals allow the expression of a pre-existent neural pattern in *C. elegans*. *Development (Cambridge, England)*. [Online] 116 (2), 457–466. Available from: http://dev.biologists.org/content/develop/116/2/457.full.pdf.

Wheeler, J.M. & Thomas, J.H. (2006) Identification of a novel gene family involved in osmotic stress response in *Caenorhabditis elegans*. *Genetics*. [Online] 174 (3), 1327–1336. Available from: doi:10.1534/genetics.106.059089.

Wilson, S. & Filipp, F.V. (2018) A network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer. *npj Systems Biology and Applications*. [Online] 4 (1), 1–10. Available from: doi:10.1038/s41540-018-0061-4.

Wines, D.R., Talbert, P.B., Clark, D. V. & Henikoff, S. (1996) Introduction of a DNA methyltransferase into *Drosophila* to probe chromatin structure in vivo. *Chromosoma*. [Online] 104 (5), 332–340. Available from: doi:10.1007/s004120050123.

Winn, J., Carter, M., Avery, L. & Cameron, S. (2011) Hox and a newly identified E2F co-repress cell death in *Caenorhabditis elegans*. *Genetics*. [Online] 188 (4), 897–905. Available from: doi:10.1534/genetics.111.128421.

Winston, W.M., Molodowitch, C. & Hunter, C.P. (2002) Systemic RNAi in C. elegans Requires the Putative Transmembrane Protein SID-1. *Science*. [Online] 295 (March), 2456–2459. Available from: doi:10.1126/science.1068836.

Wrischnik, L. a & Kenyon, C.J. (1997) The role of *lin-22, a hairy/enhancer of split* homolog, in patterning the peripheral nervous system of *C. elegans*. *Development*. [Online] 124 (15), 2875–2888. Available from: http://dev.biologists.org/content/develop/124/15/2875.full.pdf.

Wu, F., Olson, B.G. & Yao, J. (2016) DamID-seq: Genome-wide mapping of protein-DNA interactions by high throughput sequencing of adenine-methylated DNA fragments. *Journal of Visualized Experiments*. [Online] (107), 1–11. Available from: doi:10.3791/53620.

Wu, J., Duggan, A. & Chalfie, M. (2001) Inhibition of touch cell fate by *egl-44* and *egl-46* in *C. elegans*. *Genes and Development*. [Online] 15 (6), 789–802. Available from: doi:10.1101/gad.857401.

Yamamoto, Y., Takeshita, H. & Sawa, H. (2011) Multiple Wnts redundantly control polarity orientation in *Caenorhabditis elegans* epithelial stem cells. *PLoS Genetics*. [Online] 7 (10). Available from: doi:10.1371/journal.pgen.1002308.

Yang, Z., Edenberg, H.J. & Davis, R.L. (2005) Isolation of mRNA from specific tissues of *Drosophila* by mRNA tagging. *Nucleic Acids Research*. [Online] 33 (17), 1–9. Available from: doi:10.1093/nar/gni149.

Yip, Z.C. & Heiman, M.G. (2016) Duplication of a single neuron in *C. elegans* reveals a pathway for dendrite tiling by mutual repulsion. *Cell Reports*. [Online] 15 (10), 2109–2117. Available from: doi:10.1016/j.celrep.2016.05.003.

Zeitlinger, J. & Stark, A. (2010) Developmental gene regulation in the era of genomics. *Developmental Biology*. [Online] 339 (2), 230–239. Available from: doi:10.1016/j.ydbio.2009.12.039.

Zhang, G., Huang, H., Liu, D., Cheng, Y., et al. (2015) N6-methyladenine DNA modification in *Drosophila*. *Cell*. [Online] 161 (4), 893–906. Available from: doi:10.1016/j.cell.2015.04.018.

Zhang, S. & Kuhn, J.R. (2013) Cell isolation and culture. *WormBook*. [Online] 1 (February). Available from: doi:10.1895/wormbook.1.15.

Zhao, H., Sun, Z., Wang, J., Huang, H., et al. (2014) CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. [Online] 30 (7), 1006–1007. Available from: doi:10.1093/bioinformatics/btt730.

Zhong, W. & Sternberg, P.W. (2006) Genome-wide prediction of *C. elegans* genetic interactions. *Science*. [Online] 311 (5766), 1481–1484. Available from: doi:10.1126/science.1123287.

Appendix

Appendix A: Resources

A.1 List of strains used

Strain	Background	Genotype
N2	N2	wild isolate *C. elegans from Bristol, UK*
JR667	N2	*unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V*
EG6699	N2	*ttTi5605 II; unc-119(ed3) III; oxEx1578.*
MBA10	N2	*wIs51[SCMp::GFP + unc-119(+)] V; nhr-25(ku217) X*
MBA81	N2	*lin-22(icb38) IV; wIs51[SCMp::GFP + unc-119(+)] V*
MBA145	N2	*lin-22(icb38) IV; egIs1[dat-1p::GFP] IV; wIs51[SCMp::GFP + unc-119(+)] V*
MBA216	N2	*lin-22(icb38)IV; egIs1[dat-1p::GFP]IV; wIs51[SCMp::GFP + unc-119(+)] V; icbEx54[pDK4 (wrt-2p::wormCherry::lin-22:Dam::unc-54 3'UTR + cb-unc119), pMA122, pCFJ601, pGH8, pCFJ104, myo-2;;dsRed]*
MBA250	N2	*icbIs2[arf-3::pes-10::GFP:CAAX::unc-54] I; icbSi2[dpy-7p::mCherry:H2B::unc-54 3'UTR+cb-unc-119] IV; wIs51[SCMp::GFP+unc-119(+)] V*
MBA268	N2	*lin-22(icb49) egIs1[dat-1p::GFP] IV; wIs51[SCMp::GFP + unc-119(+)] V*
MBA292	N2	*icbSi5[pDK4(wrt-2p::wormCherry::lin-22:dam::unc-54 3'UTR+ cb-unc-119)] II; unc-119(ed3) III*
MBA444	N2	*icbSi10[pDK8(wrt-2p::wormCherry::NLS-GFP:dam::unc-54 3'UTR+ cb-unc-119)] II; unc-119(ed3) III*
MBA445	N2	*icbSi11[pPB10(wrt-2p::wormCherry::nhr-25:dam::unc-54 3'UTR+ cb-unc-119)] II; unc-119(ed3) III*
MBA467	N2	*icbSi25[pDK16(srf-3ap::GFPo-H2B::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III*
MBA468	N2	*icbSi26[pDK1(psrf-3ap::GFPo-H2B::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III*
MBA488	N2	*icbEx121[pDK18(dpy-7syn1::mCherry-H2B::unc-54 3'UTR) pBJ36, pRF4]; icbIs2[arf-3::GFP:CAAX::unc-54 3'UTR]I; wIs51[SCMp::GFP + unc-119(+)] V*
MBA489	N2	*icbEx122[pDK18(dpy-7syn1::mCherry-H2B::unc-54 3'UTR) pBJ36, pRF4]; icbIs2[arf-3::GFP:CAAX::unc-54 3'UTR]I; wIs51[SCMp::GFP + unc-119(+)] V*
MBA490	N2	*icbEx123[pDK18(dpy-7syn1::mCherry-H2B::unc-54 3'UTR) pBJ36, pRF4]; icbIs2[arf-3::GFP:CAAX::unc-54 3'UTR]I; wIs51[SCMp::GFP + unc-119(+)] V*
MBA496	N2	*icbSi32[pDK26(srf-3bp::GFPo-H2B::unc-54 3'UTR +cb-unc-119)] II; unc-119(ed3) III*
MBA497	N2	*icbSi33[pDK26(srf-3bp::GFPo-H2B::unc-54 3'UTR +cb-unc-119)] II; unc-119(ed3) III*
MBA540	N2	*icbSi42[pDK32(srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III*
MBA541	N2	*icbSi43[pDK32(srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III*

Strain	Background	Genotype
MBA542	N2	icbSi44[pDK32(srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III
MBA650	N2	elt-1(ku491) IV; wIs51[SCMp::GFP + unc-119(+)] V
MBA687	N2	icbSi71[pDK62(cb-unc-119 + dpy-7syn1::wormCherry::Dam-myc:ama-1::unc-54 3'UTR)] II; unc-119(ed3) III
MBA688	N2	icbSi72[pDK65(dpy-7syn1::wormCherry::Dam-myc:rpb-6::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III
MBA692	N2	icbSi76[pDK54(srf-3i1::pes-10::wormCherry::Dam-myc:NLS-GFP::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III
MBA693	N2	icbSi77[pDK64(dpy-7syn1::wormCherry::Dam-myc:NLS-GFP::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III
MBA694	N2	icbSi78[pDK46(cb-unc-119 + srf-3i1::pes-10::wormCherry::Dam-myc:ama-1::unc-54 3'UTR)] II; unc-119(ed3) III
MBA698	N2	icbSi82[pDK55(srf-3i1::pes-10::wormCherry::Dam-myc:rpb-6::unc-54 3'UTR + cb-unc-119)]; unc-119(ed3) III
MBA705	N2	icbSi84[pDK49(wrt-2p::lin-22:dam::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III
MBA707	N2	icbSi86[pDK50(wrt-2p::NLS-GFP:dam::unc-54 3'UTR + cb-unc-119)] II; unc-119(ed3) III
MBA718	N2	icbEx177[pDK59(lin-17CRE1::Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119), myo-2::dsRed]
MBA721	N2	icbEx180[pDK60(lin-17CRE2::Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119), myo-2::dsRed]
MBA744	N2	lin-22(icb49) IV; egIs1[dat-1p::GFP]IV; icbEx177[pDK59(lin-17CRE1::Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119), myo-2::dsRed]
MBA745	N2	lin-22(icb49) IV; egIs1[dat-1p::gfp]IV; icbEx180[pDK60(lin-17CRE2::Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119), myo-2::dsRed]
MBA803	N2	wIs51[SCMp::GFP + unc-119(+)] V; nhr-25(ku217) X; icbEx184[pPB10(wrt-2p::wormCherry::nhr-25:dam::unc-54 3'UTR + cb-unc-119), pMA122, pCFJ601, pGH8, pCFJ104, myo-2;:dsRed]
MBA1110	N2	unc-119(ed3) III; icbSi3[dpy-7::GFP:H2B::unc-54 3'UTR + cb-unc-119]; icbEx255[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]
MBA1133	N2	icbEx256[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]; wIs51[scm::GFP + unc-119(+)] V
MBA1134	N2	icbEx257[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]; wIs51[scm::GFP + unc-119(+)]V
MBA1135	N2	icbEx258[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]; wIs51[scm::GFP + unc-119(+)]V
MBA1136	N2	icbEx259[pDK134(srf-3i1::Δpes-10::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]; wIs51[scm::GFP + unc-119(+)]V
MBA1137	N2	icbEx260[pDK134(srf-3i1::Δpes-10::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]; wIs51[scm::GFP + unc-119(+)]V
MBA1138	N2	icbEx261[pDK133(srf-3i1::Δpes-10::mir-42-44::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1139	N2	icbEx262[pDK133(srf-3i1::Δpes-10::mir-42-44::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1140	N2	icbEx263[pDK133(srf-3i1::Δpes-10:::mir-42-44::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1141	N2	icbEx267[pDK147(dpy-7syn1::mir-42-44::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1142	N2	icbEx265[pDK139(srf-3i1::Δpes-10:::mir-47::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V

Strain	Background	Genotype
MBA1143	N2	icbEx266[pDK139(srf-3i1::Δpes-10::::mir-47::p10 3'UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1144	N2	icbEx264[pDK139(srf-3i1::Δpes-10::::mir-47::p10 3'UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1145	N2	icbEx268[pDK148(dpy-7syn1::mir-47::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1146	N2	icbEx269[pDK148(dpy-7syn1::mir-47::p10 3UTR), myo-2::dsRed, pBJ36]; unc-119(e2498::Tc1) III; wIs51[SCMp::GFP + unc-119(+)] V
MBA1147	N2	icbEx270[pDK145(CATaDa rps-25 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1148	N2	icbEx271[pDK145(CATaDa rps-25 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1149	N2	icbEx272[pDK146(CATaDa F22B7.3 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1150	N2	icbEx273[pDK146(CATaDa F22B7.3 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1151	N2	icbEx274[pDK140(CATaDa K02A2.5-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1152	N2	icbEx275[pDK140(CATaDa K02A2.5-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1153	N2	icbEx276[pDK141(CATaDa nhr-25 -upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1154	N2	icbEx277[pDK141(CATaDa nhr-25 -upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1155	N2	icbEx278[pDK150(CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1156	N2	icbEx279[pDK150(CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1157	N2	icbEx280[pDK150(CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1158	N2	icbEx281[pDK153(CATaDa K02B2.6 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1159	N2	icbEx282[pDK153(CATaDa K02B2.6 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1161	N2	icbEx284[pDK149(CATaDa nhr-4 proximal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1162	N2	icbEx285[pDK149(CATaDa nhr-4 proximal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1175	N2	icbEx288[pDK152(CATaDa Y38F1A.8 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1176	N2	icbEx289[pDK152(CATaDa Y38F1A.8 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1177	N2	icbEx290[pDK154(CATaDa nhr-4 distal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1178	N2	icbEx291[pDK154(CATaDa nhr-4 distal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]
MBA1192	N2	icbEx292[pDK158(srf-3i1-mut::Δpes-10::outron::>hda-1 fragment>::srf-3a intron5::<hda-1 fragment<::p10 3UTR), myo-2::dsRed, pBJ36]]; wIs51[SCMp::GFP + unc-119(+)] V

A.2 List of RNAi clones used

WormBase Gene ID	Public Name	Sequence Name	Ahringer RNAi library Geneservice_location
WBGene00003623	nhr-25	F11C1.6	X-6I19
WBGene00009899	efl-3	F49E12.6	II-6K03
WBGene00000275	bub-1	R06C7.8	I-3H11
WBGene00001834	hda-1	C53A5.3	V-9F11
WBGene00009672	F43G9.12	F43G9.12	I-4C12
WBGene00001974	hmg-4	T20B12.8	III-3P10
WBGene00010369	chd-1	H06O01.2	I-3M20
WBGene00003664	nhr-74	C27C7.3	I-5P17
WBGene00001976	hmg-11	T05A7.4	II-3N12
WBGene00006554	tbx-35	ZK177.10	II-4O24
WBGene00001835	hda-2	C08B11.2	II-5N08
WBGene00001971	hmg-1.1	Y48B6A.14	II-9G13
WBGene00017757	bra-2	F23H11.1	III-1E22
WBGene00001470	baz-2	ZK783.4	III-4E11
WBGene00006970	zag-1	F28F9.1	IV-1P04
WBGene00003606	nhr-7	F54D1.4	IV-6C03
WBGene00020062	nhr-270	R13D11.8	V-1B09
WBGene00003649	nhr-59	T27B7.1	V-2M07
WBGene00001210	egl-46	K11G9.4	V-4L02
WBGene00003717	nhr-127	T13F3.3	V-10J08
WBGene00000482	chd-3	T14G8.1	X-6G01
WBGene00007433	swsn-7	C08B11.3	II-5N10

Appendix

A.3 List of oligos used

Oligo name	Sequence 5' -> 3'	Working oligo name
DK11	GTGGCATGGATGAATTGTATAAGTAGTAACATGACGTCATTCCTGTGCTC	lin-22pPB7gibF
DK12	CAAAAAGCGCGATTTTTCTTGACTAGTCCCACTCTACTTCGATCCTCAA	lin-22pPB7gibR
DK15	GGTGGCATGGATGAATTGTATAAGTAGTAAATGACTGCTCCAAAGAAGAA	NLS:GFPu-DamF
DK16	AAAAGCGCGGATTTTTCTTGACTAGTCCTAGTTTGTATAGTTCATCCATGC	NLS:GFPu-DamR
DK17	CTTACTTGCACTTATAATACGACTCACTAGCAACTTTGTATAGAAAAGTTGAACG	(r-rec)_attR4F
DK18	AAGCATTGCTCATCAATTTGTTGCAACGAACAGGTCACTATCAGTCAAAATAAAATCATTATTTGCGGCCGCATTAGGCACCCCCAGGCT	pDest+(attL1)R
DK19	TTCGTTGCAACAAATTGATGAGCAATGCTTTTTTATAATGCCAACTTTGTACAAAAAAGCAGGCTATGGTCTCAAAGGGTGAAGAAGATA	(attL1)-wCHERF
DK20	CGATTTTCTTCATCCTAGGGTTACTACTTATACAATTCATCC	wC-(Avr2-dam)R
DK21	GTATAAGTAGTAACCCTAGGATGAAGAAAAATCGCGCTTT	(wC-Avr2)-DAMF
DK22	GATGGCTAGCACTTAATTAATGCATGCCGCAGATCCTCTTCAGAGATGAG	MYC-(mcs-u54)R
DK23	TGCGGCATGCATTAATTAAGTGCTAGCCATCTCGCGCCCGTGCCTCTGAC	(myc-mcs)-U54F
DK24	GGCCCGGGCTACGTAATACGACTCACTTAAAACAGTTATGTTTGGTATATTG	U54-(backb)R
DK27	CTCATCTCTGAAGAGGATCTGCGGCATGGAATGGCCGATGAGGACGATTATC	rpb-6U.TaDaF
DK28	CGAGATGGCTAGCACTTAATTAACTGCATGCTACCAATCAGCGAGTTGAA	rpb-6U.TaDaR
DK33	CTTACTTGCACTTATAATACGACTCACTAGCCTAGCTTTTCACTTTGGTTTTCACG	(BB)Psrf-3F
DK34	TTCTCCTTTACTCATCCTAGGATTCCAGATGGTACCTGAAA	Psrf-3(GFPo)R
DK35	GGTACCATCTGGAATCCTAGGATGAGTAAAGGAGAAGAATT	(Psrf-3)GFPoF
DK36	CCGATCCCCCGGGCACTTAAGCTTGTAGAGCTCGTCCATTC	GFPo(H2B)R
DK37	GACGAGCTCTACAAGCTTAAGTGCCCGGGGGATCGGTGGAG	(GFPo)H2BF
DK38	CACGGCGCGAAGTAGTGTTAATTAATTACTTGCTGGAAGTGTACT	H2B(unc54)R
DK39	ACTTCCCAGCAAGTAATTAATTAACATCTCGCGCCCGTGCCTCT	(H2B)unc54F
DK40	GGCCCGGGCTACGTAATACGACTCACTTAAAAACAGTTATGTTTGGTATA	unc54(BB)R
DK41	AGGCCTTGACTAGAGGGTACCAGAGCTCACCAACTTTGTATAGAAAAGTT	R4-L1dockp304F
DK42	CCACTTCAAAAAGCGCGATTTTTCTTCATCCTAGGGTTACTACTTATAC	R4-L1dockp304R
DK43	CTCATCTCTGAAGAGGATCTGCGGCATGCTATGACTGCTCCAAAGAAGAA	NLS-GFPpDK10F
DK44	GCGAGATGGCTAGCACTTAATTAATGCATGCTATTTGTATAGTTCATCCA	NLS-GFPpDK10R
DK46	ACATTTTGTTCTTGATAAGTTTAAACATTTATGGTCTCAAAGGGTGAAGA	dpy7syn1/2mCHF
DK47	GTAATTGGACTTAGAAGTCAGAGGCAATTTTTACTTGCTGGAAGTGTACT	dpy7syn1/2mCHR
DK48	GGGGACAAGTTTGTACAAAAAAGCAGGCTATGGTCAGCAAGGGAGAGGCA	B1wScarletF
DK49	CTTGTAGAGCTCGTCCATTCCT	wScarletnSTPR
DK50	AGGAATGGACGAGCTCTACAAGTGCCCGGGGATCGGTGGAGC	(wScrlt)-H2BF
DK59	AATTCTTCCTTTACTCATCCTAGGATTTCCTTTGATTGTGAAGATTC	Psrf-3hodgkR
DK64	TTGCACTTATAATACGACTCACTAGGCTAGGTGAGTTCTTAACTTAACCT	srf-3i1pDK16F
DK65	CGGGGATCCTCTAGAGTCGACCAAGGCCTGATTTCCTTTTGATTGTGAAG	srf-3i1pDK16R
DK66	GGCCGTGAATCTTCACAATCAAAAGGAAATCAGGCCTTGGTCGACTCTAG	(srf3i1)pes10F
DK67	TGAACAATTCTTCTCCTTTACTCATCCTAGCTCCAAGCAAGGGTCCTCCT	pes10(pDK16)R
DK89	GGGGACAACTTTGTATAGAAAAGTTGGTGAGTTCTTAACTTAACCT	srf-3i1attB4
DK90	GGGGACTGCTTTTTTGTACAAACTTGCTCGAGCTCCAAGCAAGGGTCCTCCT	pes-10XholattB1r
DK102	CTCATTGACAATCAACCTGCCAATTGTTTCTCGGATGACCTAGGATGACGTCATTCCTGTGCTC	(pwrt-2-XmaJ)lin-22F
DK103	GTAAGTCTTAACAAAAGTTCAATTGGCTCGATCAAGTAAAATTT	pCFJ151Munlrepair
DK107	GACCCCAGTTAACCCATCGATTTTTTGCAAATTAC	d-pes-10(pDK45)R
DK108	TTGCACTTATAATACGACTCACTAGGCTAGTCCATCTCTCTTCTTCATTC	(pDK49)NLS-GFPF
DK113	ACCTGCCAATTGTTTCTCGGATGACCTAGGATGACTGCTCCAAAGAAGAA	dpy-7syn1attB4
DK114	GGGGACAACTTTGTATAGAAAAGTTGCGGGAATTCCTTTGATAACG	dpy-7syn1attB1r
DK115	TTGCACTTATAATACGACTCACTAGGCTAGGTGGTGCAAGCCTTGTCTC	(pDK16)lin-17CRE1F
DK116	TTGCAAAAAATCGATGGGTTAACTGGGGTCCGACC	lin-17CRE1(d-pes10)R
DK117	GACCCCAGTTAACCCATCGATTTTTTGCAAATTAC	(lin-17CRE1)dpes-10F
DK118	TTGCACTTATAATACGACTCACTAGGCTAGTCCATCTCTCTTCTTCATTC	(pDK16)lin-17CRE2F
DK119	TTGCAAAAAATCGATAAAGACGCACACCACCCGAT	lin-17CRE2(dpes-10)R
DK120	GTGGTGTGCGTCTTTATCGATTTTTTGCAAATTAC	(lin-17CRE2)dpes-10F
DK179	AAGCTGGGAGCTCCACCGCGGTGGCGGCCGCAGATCTGTTAACTCGAATCG	(pBS)MCSp103UTRR
DK186	TGAGCGCGCGTAATACGACTCACTATAGGGCATATGTCTAGACGGGGAATTCCTTTGATAACG	(pBS)XbaI/NdeIpDK109F

Oligo name	Sequence 5' -> 3'	Working oligo name
DK203	TCGAAGACTAAGGTCTGTCAGTGGAGAGGGTGAAGG	hpGFP np F
DK204	TCGAAGACTAAAGCCCATGCCATGTGTAATCCCAGC	hpGFP np R
DK212	GTCTTCGAGAGAGGAAGACCGGCCTCGTTGGTTTTGACGATCGATTAGT	Bpil np over F
DK213	GATTATTTAAATTTCAGATGCAGGTGGGTCTTCGAGAGAGGAAGACCGG	Bpil np over ext F
DK214	GCTCTCTCCGTCTCGGCTTCCTGGAATATAAGAAATTTTGGGGAATTCC	Esp3I np over R
DK215	CCGGGGAATTCGATATCTTACAGGTGGAGACGCTCTCTCCGTCTCGGCTT	Esp3I np over ext R
DK217	TTATACTGTAATTTTTAACTTTCAGCCTAGTTCGCGGACCTTTGTGGGTG	(dpes10) mir-42-44 F
DK218	ATTTGTTATTTTAAAAACGATTCATTTAATTAATAAACCTTACAGGACGGAAATAC	(p10 3UTR) mir-42-44 R
DK219	TATTCTTACATTTTGTTCTTGATAACCTAGTTCGCGGACCTTTGTGGGTG	(dyp7-syn1) mir-42-44 F
DK220	TTATACTGTAATTTTTAACTTTCAGCCTAGGAGAGCCGACTGAAACTGAA	(dpes10) mir-47 F
DK221	ATTTGTTATTTTAAAAACGATTCATTTAATGATTACCGCACACCTTTGCC	(p10 3UTR) mir-47 R
DK222	TATTCTTACATTTTGTTCTTGATAACCTAGGAGAGCCGACTGAAACTGAA	(dyp7-syn1) mir-47 F
DK223	TGCATGCCTGCAGGCCTTGGTCGACTCTAGTCTAGTTAAAAATCTCGGATGA	CATaDa K02A2.5 F
DK224	TGCCAGCAAAACCTTATGCATTTTTGCTAGACCGGAATCAAGAAGATTAGGTTGA	CATaDa K02A2.5 R
DK225	TGCATGCCTGCAGGCCTTGGTCGACTCTAGTTTCTGTGTGGGTCTTGCCA	CATaDa K07F5.15 F
DK226	TGCCAGCAAAACCTTATGCATTTTGCTAGTCGCAAATTGTTTTGTGGT	CATaDa K07F5.15 R
DK227	TGCATGCCTGCAGGCCTTGGTCGACTCTAGGAACAAAGACTAAATGAAAAAT	CATaDa nhr-4 F
DK228	TGCCAGCAAAACCTTATGCATTTTTGCTAGACTGCTTCTCAGTAGTGTTTG	CATaDa nhr-4 R
DK229	TGCATGCCTGCAGGCCTTGGTCGACTCTAGATCTCACACAGGTGATATGTTAGAC	CATaDa sma-5 F
DK230	TGCCAGCAAAACCTTATGCATTTTGCTAGTAGCAACACCCTCATCAAAGCA	CATaDa sma-5 R
DK231	TGCATGCCTGCAGGCCTTGGTCGACTCTAGTTCCAAAATTTAGCAACGTGCC	CATaDa Y38F1A.8 F
DK232	TGCCAGCAAAACCTTATGCATTTTGCTAGGGGCAAACGGCAAATTTTA	CATaDa Y38F1A.8 R
DK233	TGCATGCCTGCAGGCCTTGGTCGACTCTAGAGATTTTGAAGAGCAAATTTA	CATaDa nhr-25 F
DK234	TGCCAGCAAAACCTTATGCATTTTGCTAGTGCAAAGTTTTGTGAGAGTTG	CATaDa nhr-25 R
DK235	TGCATGCCTGCAGGCCTTGGTCGACTCTAGACAAGACTAATCAAGGGAATACGGT	CATaDa nhr-25 down F
DK236	TGCCAGCAAAACCTTATGCATTTTGCTAGGCGATTGTAGGATGAACTTTCAGG	CATaDa nhr-25 down R
DK237	TGCATGCCTGCAGGCCTTGGTCGACTCTAGGATTTCACTGGATTTCCACCTCCG	CATaDa F22B7.3 F
DK238	TGCCAGCAAAACCTTATGCATTTTGCTAGAGAACATTTACTGTCCACCTCCG	CATaDa F22B7.3 R
DK239	TGCATGCCTGCAGGCCTTGGTCGACTCTAGTCCGTCGACTTTCAAAAGAAGTTT	CATaDa rps-25 F
DK240	TGCCAGCAAAACCTTATGCATTTTTGCTAGTAGGCTGTAGGGTAGGCCG	CATaDa rps-25 R
DK241	TGCATGCCTGCAGGCCTTGGTCGACTCTAGTGGAATACATTTTCTAACAACCTCT	CATaDa K02B2.6 F
DK242	TGCCAGCAAAACCTTATGCATTTTGCTAGTGATGTGGTAGTTGCCGGAG	CATaDa K02B2.6 R
DK243	GAG GTC GAC GGT ATC GAT AAG CTT GAT ATC GCT AGC GTG AGT TCT TAA CTT AAC C	(pDK130) srf-311 F
DK244	GTT AGT ACG CAT GTT TAA ACT TAC GTC GAC CCT AGG CTG AAA GTT AAA AAT TAC A	(outron) d-pes-10 R
DK247	TCGAAGACTAAGGTATGACTCGTTTTCACAGCGAC	hda-1 hairpin F
DK248	TCGAAGACTAAAGCTCGGTGATTCCATCGCGAAGT	hda-1 hairpin R
DK249	AATTGCAAAAATTGAATTTTCAGAAGCTACTAAAATATAAAATATGGTT	srf-311 mut Bpil F
DK250	AACCATAATTATATTATTAGTCTTTCGAAAAATTCAATTTTTGCAATT	srf-311 mut Bpil R
MB100	ATCGGGAGGCGAACCTAACTG	NM3884
MB106	AGGCAGAATGTGAACAAGACTCG	NM3880
MB270	CGGGAATTCCTTTGATAACGATAATTGGGTTATTGTGTC	dpy-7synF1
MB271	CGGGAATTCTCATTTATCACACCCTGTTCTT	dpy-synF2
PB6	GGGGACCCACTTTGTACAAGAAAGCTGGGTTTACTTGCTGGAAGTGTACTTG	GFP:H2BR1
PB7	ACCCTTTGAGACCATTCATCCGAGAAAACAATTGGCAGGT	wrt2-cherry.REV
PB8	TTGTTTCTCGGATGAATGGTCTCAAAGGGTGAAGAAGATAACA	wrt2-cherry.FOR
PB13	CACGGCGCGCGAGATGTTAACCGGCCTTTTTTCGCGG	dam-stop-unc54.REV
PB14	GCGAAAAAGCCGGTTAACATCTCGCGCCCGTG	dam-stop-unc54.FOR
PB15	ACTAGGCCCCGGGCTACGTAATACGACTCACAAACAGTTATGTTTGGTATATTGGGAATGTATTCTG	unc54-bboneRev
PB16	TGATCTTACTTGCACTTATAATACGACTCAGGTTACTACTTATACAATTCATTCCATGCCACCTGTCG	bbone-wrt2.For
PB17	AGCGCGATTTTTTCTTGACTAGTCCTAGGTTACTACTTATACAATTCATTCCATGCCACCTGTCG	cherry-stop-avrII-spel-noStart-dam.REV
PB18	GAATTGTATAAGTAGTAACCTAGGACTAGTCAGTAAGAAAAATCGCGCTTTTTTGAAGTGG	cherry-stop-avrII-spel-noStart-dam.FOR
PB19	GTGGCATGGATGAATTGTATAAGTAGTAACATGACTGACGTCGAGAGGATG	Cherry-nhr25for
PB20	TCAAAAAGCGCGATTTTTTCTTGACTAGTCTTCCTGATCCTCCTGATGCCATGTACGGCACAGC	nhr25-GGSG-DamRev

A.4 Transgenes created in this study

Extrachromosomal array number	Genetic contents	Composition of injection mix
icbEx54	[pDK4 (wrt-2p::wormCherry::lin-22:Dam::unc-54 3'UTR + cb-unc-119), pMA122, pCFJ601, pGH8, pCFJ104, myo-2;;dsRed]	pDK4 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbEx118	[pDK23(col-10p::wrmScarlet-H2B::unc-54 3'UTR +cb-unc-119) pBJ36, pRF4]	pDK23 10 ng/µl, pRF4 35 ng/µl, pBJ36 80 ng/µl
icbEx119	[pDK23(col-10p::wrmScarlet-H2B::unc-54 3'UTR +cb-unc-119) pBJ36, pRF4]	pDK23 10 ng/µl, pRF4 35 ng/µl, pBJ36 80 ng/µl
icbEx120	[pDK23(col-10p::wrmScarlet-H2B::unc-54 3'UTR +cb-unc-119) pBJ36, pRF4]	pDK23 10 ng/µl, pRF4 35 ng/µl, pBJ36 80 ng/µl
icbEx121	[pDK18(dpy-7syn1::mCherry-H2B::unc-54 3'UTR) pBJ36, pRF4]	pDK18 10 ng/µl, pRF4 35 ng/µl, pBJ36 80 ng/µl
icbEx122	[pDK18(dpy-7syn1::mCherry-H2B::unc-54 3'UTR) pBJ36, pRF4]	pDK18 10 ng/µl, pRF4 35 ng/µl, pBJ36 80 ng/µl
icbEx123	[pDK18(dpy-7syn1::mCherry-H2B::unc-54 3'UTR) pBJ36, pRF4]	pDK18 10 ng/µl, pRF4 35 ng/µl, pBJ36 80 ng/µl
icbEx177	[pDK59(lin-17CRE1::Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119), myo-2::dsRed]	pDK59 c1 25 ng/µl, myo-2;;dsRed 5 ng/µl, pBJ36 100 ng/µl
icbEx180	[pDK60(lin-17CRE2::Δpes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119), myo-2::dsRed]	pDK60 c7 25 ng/µl, myo-2;;dsRed 5 ng/µl, pBJ36 100 ng/µl
icbEx184	[pPB10(wrt-2p::wormCherry::nhr-25:dam::unc-54 3'UTR + cb-unc-119), pMA122, pCFJ601, pGH8, pCFJ104, myo-2;;dsRed]	pPB10 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbEx255	[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]	pDK130 50 ng/µl, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx256	[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]	pDK130 50 ng/µl, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx257	[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]	pDK130 50 ng/µl, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx258	[pDK130(dpy-7syn1::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]	pDK130 50 ng/µl, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx259	[pDK134(srf-3i1::Δpes-10::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]	pDK134 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx260	[pDK134(srf-3i1::Δpes-10::outron::>GFP-frag>::srf-3a intron5::<GFP-frag<::p10 3UTR), myo-2::dsRed, pBJ36]	pDK134 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx261	[pDK133(srf-3i1::Δpes-10::mir-42-44::p10 3'UTR), myo-2::dsRed, pBJ36]	pDK133 50 ng/µl, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx262	[pDK133(srf-3i1::Δpes-10::mir-42-44::p10 3'UTR), myo-2::dsRed, pBJ36]	pDK133 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx263	[pDK133(srf-3i1::Δpes-10::mir-42-44::p10 3'UTR), myo-2::dsRed, pBJ36]	pDK133 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx264	[pDK139(srf-3i1::Δpes-10::mir-47::p10 3'UTR), myo-2::dsRed, pBJ36]	pDK139 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul
icbEx265	[pDK139(srf-3i1::Δpes-10::mir-47::p10 3'UTR), myo-2::dsRed, pBJ36]	pDK139 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul

Extrachromosomal array number	Genetic contents	Composition of injection mix
icbEx266	[pDK139(srf-3i1::Δpes-10::mir-47::p10 3'UTR), myo-2::dsRed, pBJ36]	pDK139 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx267	[pDK147(dpy-7syn1::mir-42-44::p10 3UTR), myo-2::dsRed, pBJ36]	pDK147 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx268	[pDK148(dpy-7syn1::mir-47::p10 3UTR), myo-2::dsRed, pBJ36]	pDK148 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx269	[pDK148(dpy-7syn1::mir-47::p10 3UTR), myo-2::dsRed, pBJ36]	pDK148 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx270	[pDK145(CATaDa rps-25 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK145 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx271	[pDK145(CATaDa rps-25 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK145 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx272	[pDK146(CATaDa F22B7.3 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK146 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx273	[pDK146(CATaDa F22B7.3 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK146 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx274	[pDK140(CATaDa K02A2.5-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK140 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx275	[pDK140(CATaDa K02A2.5-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK140 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx276	[pDK141(CATaDa nhr-25 -upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK141 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx277	[pDK141(CATaDa nhr-25 -upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK141 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx278	[pDK150(CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK150 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx279	[pDK150(CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK150 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx280	[pDK150(CATaDa nhr-25 downstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK150 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx281	[pDK153(CATaDa K02B2.6 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK153 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx282	[pDK153(CATaDa K02B2.6 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK153 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx284	[pDK149(CATaDa nhr-4 proximal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK149 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx285	[pDK149(CATaDa nhr-4 proximal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK149 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx288	[pDK152(CATaDa Y38F1A.8 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK152 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx289	[pDK152(CATaDa Y38F1A.8 upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK152 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx290	[pDK154(CATaDa nhr-4 distal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK154 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx291	[pDK154(CATaDa nhr-4 distal-upstream element::pes-10::GFP:LacZ::unc-54 3'UTR), myo-2::dsRed, pBJ36]	pDK154 50 ng/µl, myo-2::dsRed 5 ng/µl, pBJ36 70 ng/µl
icbEx292	[pDK158(srf-3i1-mut::delta-pes-10::outron5::>hda-1 fragment>::srf-3a intron5::<hda-1 fragment<::p10 3UTR),myo-2::dsRed, pBJ36]	pDK158 50 ng/ul, myo-2::dsRed 5 ng/ul, pBJ36 70 ng/ul

Single-copy insertion number	Genetic contents	Composition of injection mix
icbSi5	[pDK4(wrt-2p::wormCherry::lin-22:dam::unc-54 3'UTR+ cb-unc-119)]	pDK4 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi10	[pDK8(wrt-2p::wormCherry::NLS-GFP:dam::unc-54 3'UTR+ cb-unc-119)]	pDK8 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi11	[pPB10(wrt-2p::wormCherry::nhr-25:dam::unc-54 3'UTR+ cb-unc-119)]	pPB10 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi25	[pDK16(srf-3ap::GFPo-H2B::unc-54 3'UTR + cb-unc-119)]	pDK16 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi26	[pDK16(srf-3ap::GFPo-H2B::unc-54 3'UTR + cb-unc-119I	pDK16 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi32	[pDK26(srf-3bp::GFPo-H2B::unc-54 3'UTR +cb-unc-119)]	pDK26 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi33	[pDK26(srf-3bp::GFPo-H2B::unc-54 3'UTR +cb-unc-119)]	pDK26 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi42	[pDK32(srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119)]	pDK32 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi43	[pDK32(srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119)]	pDK32 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi44	[pDK32(srf-3i1::pes-10::GFPo-H2B::unc-54 3'UTR + cb-unc-119)]	pDK32 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi71	[pDK62(cb-unc-119 + dpy-7syn1::wormCherry::Dam-myc:ama-1::unc-54 3'UTR)]	pDK62 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi72	[pDK65(dpy-7syn1::wormCherry::Dam-myc::rpb-6::unc-54 3'UTR + cb-unc-119)]	pDK65 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi76	[pDK54(srf-3i1::pes-10::wormCherry::Dam-myc:NLS-GFP::unc-54 3'UTR + cb-unc-119)]	pDK54 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi77	[pDK64(dpy-7syn1::wormCherry::Dam-myc::NLS-GFP::unc-54 3'UTR + cb-unc-119)]	pDK64 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi78	[pDK46(cb-unc-119 + srf-3i1::pes-10::wormCherry::Dam-myc:ama-1::unc-54 3'UTR)]	pDK46 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi82	[pDK55(srf-3i1::pes-10::wormCherry::Dam-myc::rpb-6::unc-54 3'UTR + cb-unc-119)]	pDK55 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi84	[pDK49(wrt-2p::lin-22:dam::unc-54 3'UTR + cb-unc-119)]	pDK49 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl
icbSi86	[pDK50(wrt-2p::NLS-GFP:dam::unc-54 3'UTR + cb-unc-119)]	pDK50 50 ng/µl, pCFJ601 50 ng/µl, pMA122 10 ng/µl, pGH8 10 ng/µl, myo-2;;dsRed 2,5 ng/µl, pCFJ104 5 ng/µl

A.5 List of smFISH probes used

elt-1 (1:50)	*lin-39* (1:50)	*egl-18* (1:50)
gttagcatcacgataatgca	gatgtggtcatcttaattct	cgtcattatgctgatcgaca
agattcactttatttcggga	atctgtggatgacggtgatg	agcacttcgtggtgttgttg
ccgacaactccatctaacat	attcaggagctgtagctctc	ctacacggctcatctgacgg
cattcgtgttctgcatatca	gaggatgaagacgaagagct	cttctgtaactgtttgcaac
gacggagcaaagagtccaac	ccacagaagatgtggatgat	tgctgattgtctttgcaaca
gatggagtttgtgtcaggat	gatgatggaattccagatgc	ttgtccattcgctccataac
atgttttcctgctcaattgg	ctgtcattggatcatatcca	ctcgtcgagccgatactgaa
aattcaacgggtttccttc	tgagcagaaagtgcagcaga	gctcattgttctctttgagc
tccagaagaagtgccgagag	gtcggatcataataacttcc	gatgagaccgatgagctttt
aatggtgcaatggatgcagc	gtccttgacttgaggcaaaa	cggtgatggtgaggcttttc
gttggttgctgatgtattgt	attggatattgaggacctcc	tctcgacaagcttcggagag
tggataactactcggtgtgg	tagcatagtgactgatctcc	cctgatactggagcgactac
cctccagcatattgatagtt	gggtacttgttactgatgga	aagtctggaagtggactcgc
atgcatccatatcagttgtc	tgcaccgcctgaattcttat	ggatcaaacatgaatccgtt
attgacatatccattccact	aacacgtgtcatccatggat	gcatcattccatttggattt
cggattgttttgtgttccaa	gtgaacctcctgtagttgaa	ctcacggattgttgattctc
gagttgtgtcgttggataga	ttgtgtatgctgttcgttgt	cacggatttcgattgtggtg
gcgagtgtatcatatccata	tccagctctaatacttgatt	gatccattggatcttcaatt
cagtaattccagatgctgtc	tttgtgtgtatgaaattcct	gactcttcctgtttcacatc
tacgatagttccatttcctt	aattctacgcttcctcgtca	agtgcatccaaaaggttctg
tgttttgagtgattggctgt	gcatcaatgaatgagctact	gttgctgctaaactgtgctg
ggtgagaggttgactgttat	attttgacttgtctttcggt	tggtggaggtctggaagaac
gatgatcctgaagacgtcga	gcttcattcgtcgattttga	gatgatgatgatggtggaga
gagtttgctgatgagcttga	ggtttatccttattttcttt	ttcgatgaccgcttgtactt
ggtgttcttcggtgtagaag	aaatggcatcatcggaggtg	ttgaactgtcttggctttgg
atgatcggtttgctttggag	gaccgaatggtagatttgca	ctcgtctcggttttctcaaa
atcttccgtgctgaattgag	gattgaaaagtgggaaccgg	acggattgcagacaagcttc
ctccacaattgacacactca		gtgcaatcgatagtagagcc
cgacgccataatggagtatt		atttctattggtcggcgaac
atgcgttgcagaggtagttt		ttgttggatgtggttttttgc
ttcatcttgaagtagaggcc		actctttttccttgttcttt
ttcaccaatggacgagcatg		tggttgaagatctgtgttgc
tttttgagcgttctgctgtc		atcgtcggcatttgtgtgag
acgacagttgacacactcga		ttgaatgtgttgatggctcc
atttcttctccagagtgttg		ctgcgtgaattgcgagattt
acaagcattgcacactggat		tgctattcatgagctcttga
cgttccaccttatgaagttt		
atccttcttcatggtgattg		
ttgcgattacgggtctgaat		
tttcttcattcttcgcgatc		
gaagatgttgtaggcattcc		
ccactttgatcaagttcgat		
agtgttttcattccccaaa		
tgtcggtgtcatgagcattg		
tggatgctgggaatgcatat		
tcctcgatcgagttgaagta		
ggcacgttttatactccaat		
tccaaagtccaccatcattg		

efl-3 (1:50)	cki-1 (1:5)	rnt-1 (1:20)
agagccgaagatagttgtca	caacgacgagcagaagacat	gagcacaggaatgacgtcat
ccgggtggaatattttcttt	ccaaattcgagtcctggagc	ccatcggattcaatttcagt
tctggatctgttgttcgatt	atgcgcttaacagcatcttc	atcttcttgcatctggagat
ctggagatggtcttggaaca	atttctggctttcttcctgg	tggacagctgtgaaattaca
gaagcatcggagagttgtga	gagtctccagttcaaagtcg	tctccattagaggtttgttt
gtcacttctagatcgtcatc	ataaacgaatccagcagagc	tttattcgagctctccgttt
cccaaactcttttcttttct	ggaacacaattctctggaat	gttttagttgtgacagtgac
ctgtctcttcattgatagca	aactttggttctgtagaact	cgataccttatgctcatctt
attcattttccttgcaacgg	agcatgtggttctgacagtg	acttgtttttcggaggtat
cgtagattcgtcgtttttca	gttgagctgatgtccagcga	cctgatctgaagcaatgcta
ttgttagttttctgcatggc	agagctcaatggagtcaaag	taattccggcaacttgccga
aagagactctaatccttgcc	tcctccttatcagatgtgct	cacagccctagttgtaattg
ctctggaagaccttcttcaa	gagctgttgggatccatcag	tttttctaaaaccggccgt
cacgtggtgatgacaactca	cggttcctcttcatcttcga	acaacatacccctcggaaaa
gagccgacaatatccttgaa	gctctctgaattgccacttc	aactcggaaattgtccgccg
ggtcggtgttgaagtacatg	catcttctgctgacgctttg	gatgaagtatagcgagtggc
acgagtgtcaactctcgatc	ttcttcttacgggaaactgc	gagtttggccagtcattata
tcaaaaaccgtctgcacagc	acagcttgtttggagacaac	ttcacggcggcggacaattt
cttttcggatattcttcggg	gtgaagatcacattcaccgg	aattttccgaggggtatgtt
aactgtgctagctacatcga	cgttggacgacgagattttg	ggtcgtccggtttagaaaa
aaaaccttctgtttccggat	agcatgaagatcgagttctg	tataagagctgtgcaaccgg
tataagcgtctgcatcgact		tcaaacctgcctacggaata
aaccaggacattggcaatgt		atctctcgcgctacttaaaa
accttttaatgagcccaag		tgcacgttattcgcgaattc
cgggattttctttgttccaa		gatttttttacaacttgcca
ggttcaggaccacagtaaac		aatagaggcggaggctatta
aaacatcgaagcttccggtt		tggaatacccaaaagaagcc
gtgaactactcagaaggcgt		
caattttgtcagtctgtgct		
ttttgttctgagagcgttct		
tgcgattttttggagtattcc		
tggtgatgatctgacagctg		
ccacctctgcgaacataaac		
tatctcattttctcagctgc		
attgatgctactgatcccaa		
gatgtcattggatgctgtga		
ggaagttgaagaggagccat		
ggctttggaatctctggaag		
tgtaaactgcttttgagcct		
attgaatcgaggcctagacg		
ttgcacaggtgtatagtcag		
agatacgagtgggcgtattg		
gttgcgaagattcttgtgga		
aatgttttggcttgacgtct		
cttcgattcgccaagaatgt		
tgtgctcgaacgtattctga		
cacctggaatgcggaagatg		
ctccgaagactttttttggt		

Appendix B: Additional tables

B.1 Sequencing results summary statistics for all samples of this study

Sequencing results maping and quality statistics

| | | | | | Mapping to WBcel235 assembly | | | | | | | Quality control statistics fastq-stats 2000000 reads sample | | | | |
| | | | | | Aligned 0 times | | Aligned 1 | | Aligned >1 | | | | | | | |
Strain	Stage	Rep	Total reads	Overall alignment	Reads #	%	Reads #	%	Reads #	%	xtimes genomic coverage	Uncut DamID adaptors	Primer-dimer (fwd)	Primer-dimer (rev)	Initial GATC sequence	Internal GATC sites
LIN-22:DAM TaDa																
MBA292	L2	1	30465427	93.81%	1885058	6.19%	28023014	91.98%	557355	1.83%	41	19589	6430	3508	2501	188021
MBA444	L2	1	30634768	94.38%	1722191	5.62%	27311867	89.15%	1600710	5.23%	40	17112	8987	7434	2424	174085
MBA292	L2	2	27375573	94.24%	1577434	5.76%	25272570	92.32%	525569	1.92%	37	19268	8285	4895	2585	226250
MBA444	L2	2	27459606	93.81%	1700080	6.19%	24384507	88.80%	1375019	5.01%	36	19448	9129	6994	3011	230877
MBA292	L4	1	28551918	95.28%	1348327	4.72%	26472929	92.72%	730662	2.56%	39	20887	6213	3156	2677	236497
MBA444	L4	1	33434533	96.20%	1271620	3.80%	30235385	90.43%	1927528	5.77%	44	13626	6455	3757	2892	175152
MBA292	L4	2	27397958	95.61%	1202172	4.39%	25633718	93.56%	562068	2.05%	37	35259	9351	2336	3770	255849
MBA444	L4	2	28653463	92.23%	2225903	7.77%	24742317	86.35%	1685243	5.88%	36	22032	20300	20271	3137	196925
NHR-25:DAM TaDa																
MBA444	L2	1	29348674	88.98%	3234737	11.02%	24851160	84.68%	1262777	4.30%	36	28325	12761	10756	1666	204960
MBA445	L2	1	24611396	25.93%	18229171	74.07%	6059055	24.62%	323170	1.31%	9	41169	144652	160379	5171	647643
MBA444	L2	2	274417459	64.07%	9851842	35.93%	16625880	60.64%	939737	3.43%	24	21842	39038	39667	1440	258316
MBA445	L2	2	251116949	54.76%	11363180	45.24%	12424130	49.47%	1329639	5.29%	18	23150	87701	87533	12971	353255
MBA444	L4	1	27922799	93.69%	1761862	6.31%	24625525	88.19%	1535412	5.50%	36	31360	11767	8413	3270	244516
MBA445	L4	1	24931664	59.12%	10192558	40.88%	12714590	51.00%	2024516	8.12%	19	41505	41424	39175	30890	335971
MBA444	L4	2	27933355	96.66%	932681	3.34%	25530161	91.40%	1470513	5.26%	37	17452	5841	2098	1830	174626
MBA445	L4	2	26540704	36.94%	16736144	63.06%	9554417	36.00%	250143	0.94%	14	23043	49976	50233	2432	298974

Sequencing results maping and quality statistics

Seam and hypodermis RNApol TaDa

Strain	Stage	Rep	Mapping to WBcel235 assembly									Quality control statistics fastq-stats 2000000 reads sample				
			Total reads	Overall alignment	Aligned 0 times		Aligned 1		Aligned >1		xtimes genomic coverage	Uncut DamID adaptors	Primer-dimer (fwd)	Primer-dimer (rev)	Initial GATC sequence	Internal GATC sites
					Reads #	%	Reads #	%	Reads #	%						
MBA692	L2	1	38677810	44.32%	21537024	55.68%	15714778	40.63%	1426008	3.69%	23	12452	272432	281319	2051	697878
MBA698	L2	1	37007535	61.99%	14064945	38.01%	13654568	36.90%	9288022	25.10%	20	17091	120132	115589	2257	339580
MBA693	L2	1	42990226	76.05%	10298253	23.95%	30571241	71.11%	2120732	4.93%	45	14742	122684	107311	939	304518
MBA688	L2	1	30871231	95.49%	1392496	4.51%	18540163	60.06%	10938572	35.43%	27	21161	9534	6690	3352	200423
MBA692	L2	2	40664354	50.77%	20019148	49.23%	19118670	47.02%	1526536	3.75%	28	16553	204125	207777	1833	597930
MBA698	L2	2	36754420	86.07%	5120432	13.93%	18078872	49.19%	13555116	36.88%	26	26440	49388	49482	4884	247705
MBA693	L2	2	35124166	65.72%	12040804	34.28%	21507061	61.23%	1576301	4.49%	31	15088	102232	95705	2287	315030
MBA688	L2	2	41329805	92.70%	3016342	7.30%	21670746	52.43%	16642717	40.27%	32	26344	20992	18632	4750	217227
MBA692	L4	1	36975870	36.84%	23355783	63.16%	12530630	33.89%	1089457	2.95%	18.2					
MBA698	L4	1	23933259	85.57%	3454722	14.43%	11722758	48.98%	8755779	36.58%	17	20554	24447	23171	3076	213389
MBA693	L4	1	29880094	45.23%	16365722	54.77%	12520569	41.90%	993803	3.33%	18	15223	104533	96574	1810	289622
MBA688	L4	1	19406621	96.75%	630816	3.25%	12260719	63.18%	6515086	33.57%	18	30068	8138	3254	2911	217608
MBA692	L4	2	28411452	75.64%	6922386	24.36%	19963489	70.27%	1525577	5.37%	29	12676	112216	103786	1696	298101
MBA698	L4	2	34240698	81.95%	6179301	18.05%	13716138	40.06%	14345259	41.90%	20.0					
MBA693	L4	2	27561932	92.95%	1942722	7.05%	24263097	88.03%	1356113	4.92%	35	21645	20156	18503	1505	168847
MBA688	L4	2	23634780	94.14%	1385258	5.86%	13547979	57.32%	8701543	36.82%	20	25124	22422	20448	6780	234445

Sequencing results maping and quality statistics

Seam and hypodermis DAM:AMA-1 TaDa

Strain	Stage	Rep	Mapping to WBcel235 assembly								
			Total reads	Overall alignment	Aligned 0 times		Aligned 1		Aligned >1		xtimes genomic coverage
					Reads #	%	Reads #	%	Reads #	%	
MBA694	L2		35015699	5.68%	33027136	94.32%	1882502	5.38%	106061	0.30%	2.7
MBA692	L2		34529333	40.00%	20717715	60.00%	12843928	37.20%	967690	2.80%	18.7
MBA687	L2		28289804	1.17%	27959663	98.83%	269254	0.95%	60887	0.22%	0.4
MBA693	L2		38331244	40.34%	22866560	59.66%	14512181	37.86%	952503	2.48%	21.1
MBA694	L4		30927732	17.56%	25497489	82.44%	5112292	16.53%	317951	1.03%	7.4
MBA692	L4		36975870	36.84%	23355783	63.16%	12530630	33.89%	1089457	2.95%	18.2
MBA687	L4		37874323	1.08%	37466885	98.92%	333532	0.88%	73906	0.20%	0.5
MBA693	L4		38950805	6.57%	36392824	93.43%	2359851	6.06%	198130	0.51%	3.4

B.2 NHR-25 TOMTOM motif similarity complete results

Name	Species	Class	Optimal_offset	p-value	E-value	q-value	Overlap	Query consensus	Target consensus	Orientation
Esrra	Mus musculus	Nuclear receptors with C4 zinc fingers	0	1.40E-10	1.96E-07	3.93E-07	10	ATGACCTTGA	ATGACCTTGAA	-
Esrrg	Mus musculus	Nuclear receptors with C4 zinc fingers	0	9.81E-08	0.000138	9.99E-05	10	ATGACCTTGA	ATGACCTTGA	-
ESRRB	Homo sapiens	Nuclear receptors with C4 zinc fingers	1	1.07E-07	0.00015	9.99E-05	10	ATGACCTTGA	TATGACCTTGA	-
ZAP1	Saccharomyces cerevisiae	C2H2 zinc finger factors	1	1.77E-05	0.024848	0.012424	10	ATGACCTTGA	CATGACCTTTAAGGT	-
RORA	Homo sapiens	Nuclear receptors with C4 zinc fingers	-1	2.56E-05	0.035991	0.014396	9	ATGACCTTGA	TGACCTTGAT	-
Nr5a2	Mus musculus	Nuclear receptors with C4 zinc fingers	1	0.000108	0.152052	0.050684	10	ATGACCTTGA	GCTGACCTTGAACTT	-
NR4A2	Rattus norvegicus	Nuclear receptors with C4 zinc fingers	0	0.001043	1.46463	0.40746	8	ATGACCTTGA	GTGACCTT	-
NR4A1	Homo sapiens	Nuclear receptors with C4 zinc fingers	0	0.001287	1.80717	0.40746	10	ATGACCTTGA	GTGACCTTTT	-
NR2F2	Homo sapiens	Nuclear receptors with C4 zinc fingers	1	0.001344	1.88657	0.40746	10	ATGACCTTGA	TTTGACCTTTG	-
PPARA::RXRA	Homo sapiens	Nuclear receptors with C4 zinc fingers::Nuclear receptors with C4 zinc fingers	0	0.001451	2.0373	0.40746	10	ATGACCTTGA	ATGACCTTTGACCTACAT	-
NR1H4	Homo sapiens	Nuclear receptors with C4 zinc fingers	3	0.001716	2.40918	0.438033	8	ATGACCTTGA	TCAATGACCTA	+
Rarb(var.2)	Mus musculus	Nuclear receptors with C4 zinc fingers	-1	0.004616	6.48022	1	9	ATGACCTTGA	TGACCTTTAGTTGACCT	-
RARA(var.2)	Homo sapiens	Nuclear receptors with C4 zinc fingers	-1	0.0064	8.98527	1	9	ATGACCTTGA	TGACCTTTGCATGACCT	-
MYB3R1	Arabidopsis thaliana	Helix-Turn-Helix	4	0.006644	9.32823	1	10	ATGACCTTGA	AAATATGACCGTTAG	+

B.3 LIN-22 TOMTOM motif similarity complete results

Target_ID	Name	Species	Class	Optimal_offset	p-value	E-value	q-value	Overlap	Query_consensus	Target_consensus	Orientation
MA0930.1	ABF3	Arabidopsis thaliana	Basic leucine zipper factors (bZIP)	-1	0.000228	0.319712	0.091279	8	CACAGGTGTT	ACACGTGT	+
MA0941.1	ABF2	Arabidopsis thaliana	Basic leucine zipper factors (bZIP)	3	0.000812	1.1394	0.129482	10	CACAGGTGTT	CATGACACGTGTT	+
MA0931.1	ABI5	Arabidopsis thaliana	Basic leucine zipper factors (bZIP)	1	0.001607	2.2564	0.170698	9	CACAGGTGTT	TGACACGTGG	+
MA0549.1	BZR2	Arabidopsis thaliana	0	1	0.001879	2.63846	0.183747	10	CACAGGTGTT	TCACACGTGCG	-
MA1207.1	GT3a	Arabidopsis thaliana	Helix-Turn-Helix	0	0.002337	3.28175	0.206429	10	CACAGGTGTT	AACACGTGTTAAATTTG	+
MA0583.1	RAV1(var.2)	Arabidopsis thaliana	AP2/ERF domain	2	0.00352	4.94263	0.247872	10	CACAGGTGTT	GCCTCAGGTGAT	-
MA1333.1	AT4G18890	Arabidopsis thaliana	Other	0	0.003917	5.4994	0.261552	10	CACAGGTGTT	CGCACGTGTGA	+
MA1332.1	AT4G36780	Arabidopsis thaliana	Other	0	0.004505	6.32474	0.262772	10	CACAGGTGTT	CGCACGTGTGA	-
MA0569.1	MYC4	Arabidopsis thaliana	Basic helix-loop-helix factors (bHLH)	-1	0.004993	7.00962	0.262772	8	CACAGGTGTT	GCACGTGT	-
MA1331.1	AT1G78700	Arabidopsis thaliana	Other	1	0.006543	9.18596	0.278433	10	CACAGGTGTT	CCACACGTGCG	-
MA0545.1	hlh-1	Caenorhabditis elegans	Basic helix-loop-helix factors (bHLH)	0	0.000245	0.34375	0.091279	10	CACAGGTGTT	GACAGCTGTTC	-
MA0249.1	twi	Drosophila melanogaster	Basic helix-loop-helix factors (bHLH)	1	0.001997	2.80322	0.188489	10	CACAGGTGTT	TCGCATATGTTG	+
MA0086.2	sna	Drosophila melanogaster	C2H2 zinc finger factors	2	0.003531	4.95751	0.247872	10	CACAGGTGTT	GCAACAGGTGCAA	+
MA0783.1	PKNOX2	Homo sapiens	Homeo domain factors	1	6.59E-05	0.092498	0.091279	10	CACAGGTGTT	TGACAGGTGTCA	+
MA0522.2	TCF3	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.000144	0.202304	0.091279	10	CACAGGTGTT	AGCAGGTGTT	-
MA0782.1	PKNOX1	Homo sapiens	Homeo domain factors	1	0.000151	0.212132	0.091279	10	CACAGGTGTT	TGACAGGTGTCA	+
MA0820.1	FIGLA	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.000184	0.257859	0.091279	10	CACAGGTGTT	AACAGGTGGT	-
MA0665.1	MSC	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.000267	0.374483	0.091279	10	CACAGGTGTT	AACAGCTGTT	-
MA0745.1	SNAI2	Homo sapiens	C2H2 zinc finger factors	0	0.000301	0.422553	0.091552	9	CACAGGTGTT	AACAGGTGT	+
MA0824.1	ID4	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.000363	0.510296	0.093615	10	CACAGGTGTT	GACAGGTGTA	-
MA0797.1	TGIF2	Homo sapiens	Homeo domain factors	1	0.000407	0.571683	0.093615	10	CACAGGTGTT	TGACAGCTGTCA	+
MA0819.1	CLOCK	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.00041	0.576101	0.093615	10	CACAGGTGTT	AACACGTGTT	-

Target_ID	Name	Species	Class	Optimal_offset	p-value	E-value	q-value	Overlap	Query_consensus	Target_consensus	Orientation
MA0743.1	SCRT1	Homo sapiens	C2H2 zinc finger factors	4	0.000497	0.698232	0.101682	10	CACAGGTGTT	GAGCAACAGGTGGTT	+
MA0667.1	MYF6	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.000617	0.866266	0.105574	10	CACAGGTGTT	AACAGGTGTT	-
MA0796.1	TGIF1	Homo sapiens	Homeo domain factors	1	0.000892	1.25266	0.129482	10	CACAGGTGTT	TGACAGCTGTCA	+
MA0649.1	HEY2	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.001397	1.96148	0.170698	10	CACAGGTGTT	GGCACGTGTC	-
MA0669.1	NEUROG2	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.001464	2.05574	0.170698	10	CACAGGTGTT	GACATATGTT	-
MA0691.1	TFAP4	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.002834	3.97875	0.232943	10	CACAGGTGTT	AACAGCTGAT	+
MA0823.1	HEY1	Homo sapiens	Basic helix-loop-helix factors (bHLH)	0	0.00329	4.61885	0.247872	10	CACAGGTGTT	GACACGTGCC	+
MA0807.1	TBX5	Homo sapiens	T-Box factors	-3	0.003653	5.12821	0.249996	7	CACAGGTGTT	AGGTGTGA	+
MA0806.1	TBX4	Homo sapiens	T-Box factors	-3	0.004557	6.39831	0.262772	7	CACAGGTGTT	AGGTGTGA	+
MA0744.1	SCRT2	Homo sapiens	C2H2 zinc finger factors	4	0.005072	7.12094	0.262772	9	CACAGGTGTT	ATGCAACAGGTGG	+
MA0103.3	ZEB1	Homo sapiens	Homeo domain factors	1	0.006246	8.76942	0.278433	10	CACAGGTGTT	GCGCAGGTGGG	-
MA0091.1	TAL1::TCF3	Homo sapiens	Basic helix-loop-helix factors (bHLH)::Basic helix-loop-helix factors (bHLH)	0	0.006308	8.85664	0.278433	10	CACAGGTGTT	AACAGATGGTCG	-
MA1108.1	MXI1	Homo sapiens	Basic helix-loop-helix factors (bHLH)	1	0.006611	9.28126	0.278433	10	CACAGGTGTT	GGGCACGTGGTCG	-
MA0801.1	MGA	Homo sapiens	T-Box factors	-3	0.007109	9.98047	0.280998	7	CACAGGTGTT	AGGTGTGA	+
MA0832.1	Tcf21	Mus musculus	Basic helix-loop-helix factors (bHLH)	2	0.001621	2.27601	0.170698	10	CACAGGTGTT	GCAACAGCTGTTGT	+
MA0004.1	Arnt	Mus musculus	Basic helix-loop-helix factors (bHLH)	-2	0.004413	6.19551	0.262772	6	CACAGGTGTT	CACGTG	-
MA0816.1	Ascl2	Mus musculus	Basic helix-loop-helix factors (bHLH)	0	0.004654	6.53411	0.262772	10	CACAGGTGTT	AGCAGCTGCT	-
MA0626.1	Npas2	Mus musculus	Basic helix-loop-helix factors (bHLH)	0	0.005384	7.55897	0.26321	10	CACAGGTGTT	GACACGTGCC	-
MA0608.1	Creb3l2	Mus musculus	Basic leucine zipper factors (bZIP)	0	0.006097	8.55959	0.278433	9	CACAGGTGTT	GCCACGTGT	+
MA1033.1	OJ1058_F05.8	Oryza sativa Japonica Group	Basic leucine zipper factors (bZIP)	-1	0.001034	1.45216	0.141583	8	CACAGGTGTT	ACACGTGT	+
MA0310.1	HAC1	Saccharomyces cerevisiae	Basic leucine zipper factors (bZIP)	-2	0.002932	4.11716	0.232943	8	CACAGGTGTT	CACGTGTC	-
MA0335.1	MET4	Saccharomyces cerevisiae	Basic leucine zipper factors (bZIP)	1	0.005487	7.70336	0.263532	7	CACAGGTGTT	CCACAGTT	-

B.4 LIN-22:DAM L2 gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichmen	P value	Q value
supramolecular polymer GO:0099081	60	130	2.2	7.90E-21	9.50E-19
cellular developmental process GO:0048869	1.80E+02	261	1.5	5.70E-12	3.40E-10
neurogenesis GO:0022008	59	108	1.8	2.00E-11	8.20E-10
cell part morphogenesis GO:0032990	37	75	2	8.90E-11	2.70E-09
cell projection organization GO:0030030	63	110	1.8	2.00E-10	4.70E-09
neuron development GO:0048666	36	72	2	3.30E-10	6.60E-09
regulation of neuron differentiation GO:0045664	24	53	2.2	6.20E-10	1.10E-08
neuron projection guidance GO:0097485	26	53	2.1	1.30E-08	1.90E-07
actin filament-based process GO:0030029	35	66	1.9	2.70E-08	3.60E-07
post-embryonic development GO:0009791	97	139	1.4	2.20E-06	2.70E-05
localization of cell GO:0051674	28	50	1.8	6.70E-06	7.30E-05
cell junction GO:0030054	39	63	1.6	2.30E-05	0.00023
cell projection GO:0042995	1.20E+02	162	1.3	3.50E-05	0.00033
biological adhesion GO:0022610	19	35	1.8	7.00E-05	0.0006
small GTPase mediated signal transduction GO:0007264	26	41	1.6	0.00043	0.0034
phosphorus metabolic process GO:0006793	2.50E+02	296	1.2	0.00053	0.004
cell body GO:0044297	44	63	1.4	0.00068	0.0048
kinase binding GO:0019900	24	38	1.6	0.00073	0.0049
oviposition GO:0018991	31	47	1.5	0.00078	0.0049
response to endogenous stimulus GO:0009719	35	51	1.5	0.0014	0.0086
calcium ion binding GO:0005509	34	50	1.5	0.0015	0.0086
male anatomical structure morphogenesis GO:0090598	23	36	1.5	0.002	0.011
post-embryonic animal organ development GO:0048569	35	50	1.4	0.0025	0.013
potassium ion transmembrane transport GO:0071805	20	30	1.5	0.0058	0.029
dephosphorylation GO:0016311	52	68	1.3	0.006	0.029
ribonucleotide binding GO:0032553	2.20E+02	256	1.1	0.0064	0.03
purine nucleotide binding GO:0017076	2.20E+02	253	1.1	0.0071	0.032
nucleoside phosphate binding GO:1901265	2.50E+02	282	1.1	0.0083	0.035
transcription regulatory region sequence-specific DNA binding GO:0000976	51	66	1.3	0.0085	0.035
RNA polymerase II regulatory region DNA binding GO:0001012	48	62	1.3	0.01	0.041
aging GO:0007568	49	62	1.3	0.018	0.071
dendritic tree GO:0097447	26	35	1.4	0.021	0.077
response to nitrogen compound GO:1901698	24	33	1.4	0.022	0.082
regulatory region nucleic acid binding GO:0001067	59	72	1.2	0.023	0.082
reproductive system development GO:0061458	25	33	1.3	0.028	0.096

B.5 LIN-22:DAM L4 gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichmer	P value	Q value
supramolecular polymer GO:0099081	59	115	2	1.90E-14	2.30E-12
actin filament-based process GO:0030029	35	70	2	3.20E-10	1.90E-08
cellular developmental process GO:0048869	1.70E+02	242	1.4	1.70E-08	6.70E-07
cell junction GO:0030054	39	68	1.8	3.50E-07	1.00E-05
post-embryonic development GO:0009791	96	141	1.5	3.70E-07	1.00E-05
response to endogenous stimulus GO:0009719	35	59	1.7	4.80E-06	9.60E-05
cell part morphogenesis GO:0032990	37	62	1.7	5.00E-06	9.60E-05
neurogenesis GO:0022008	59	89	1.5	9.30E-06	0.00014
cell projection organization GO:0030030	62	93	1.5	1.10E-05	0.00015
neuron development GO:0048666	36	59	1.7	1.50E-05	0.00018
regulation of neuron differentiation GO:0045664	24	41	1.7	6.80E-05	0.00074
localization of cell GO:0051674	28	46	1.7	0.00011	0.0011
neuron projection guidance GO:0097485	25	41	1.6	0.00038	0.0035
kinase binding GO:0019900	23	38	1.6	0.00056	0.0048
small GTPase binding GO:0031267	19	32	1.7	0.0008	0.0064
post-embryonic animal organ development GO:0048569	35	51	1.5	0.0011	0.0079
biological adhesion GO:0022610	19	31	1.6	0.0014	0.0099
aging GO:0007568	48	67	1.4	0.0016	0.011
regulatory region nucleic acid binding GO:0001067	58	77	1.3	0.0026	0.016
RNA polymerase II regulatory region DNA binding GO:0001012	47	64	1.4	0.0032	0.019
transcription regulatory region sequence-specific DNA binding GO:0000976	50	67	1.3	0.0041	0.024
calcium ion binding GO:0005509	34	47	1.4	0.0056	0.03
male anatomical structure morphogenesis GO:0090598	23	34	1.5	0.0056	0.03
passive transmembrane transporter activity GO:0022803	75	95	1.3	0.0062	0.031
embryo development ending in birth or egg hatching GO:0009792	67	85	1.3	0.0063	0.031
phosphorus metabolic process GO:0006793	2.50E+02	281	1.1	0.0065	0.031
regulation of nucleobase-containing compound metabolic process GO:0019219	2.10E+02	238	1.2	0.0075	0.033
negative regulation of metabolic process GO:0009892	1.20E+02	140	1.2	0.0083	0.036
multicellular organism growth GO:0035264	21	31	1.5	0.0083	0.036
small GTPase mediated signal transduction GO:0007264	25	35	1.4	0.013	0.054
cation binding GO:0043169	3.20E+02	356	1.1	0.018	0.07
nucleoside phosphate metabolic process GO:0006753	41	53	1.3	0.02	0.076
oviposition GO:0018991	31	40	1.3	0.025	0.091
potassium ion transmembrane transport GO:0071805	20	27	1.4	0.028	0.098

B.6 LIN-22:DAM L2-only gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichmer	P value	Q value
cell projection organization GO:0030030	20	36	1.8	0.00027	0.032
neurogenesis GO:0022008	19	34	1.8	0.00038	0.032
neuron projection guidance GO:0097485	8.2	18	2.2	0.00055	0.032
neuron development GO:0048666	12	23	2	0.00056	0.032
regulation of neuron differentiation GO:0045664	7.7	17	2.2	0.00062	0.032
cellular developmental process GO:0048869	57	80	1.4	0.00083	0.032
cell part morphogenesis GO:0032990	12	22	1.8	0.002	0.034
cell projection GO:0042995	39	56	1.4	0.0028	0.043
nucleoside binding GO:0001882	14	24	1.7	0.003	0.043
potassium ion transmembrane transport GO:0071805	6.4	13	2	0.0044	0.053
reproductive system development GO:0061458	7.9	15	1.9	0.0056	0.061
development of primary sexual characteristics GO:0045137	6.6	13	2	0.0062	0.062
cell body GO:0044297	14	23	1.6	0.007	0.065

B.7 NHR-25:DAM L2 gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichment	P value	Q value
supramolecular polymer GO:0099081	68	117	1.7	5.30E-11	6.40E-09
actin filament-based process GO:0030029	40	63	1.6	3.00E-05	0.0018
post-embryonic development GO:0009791	1.10E+02	148	1.3	3.30E-05	0.0018
organic acid metabolic process GO:0006082	1.00E+02	135	1.4	3.90E-05	0.0018
molting cycle GO:0042303	24	41	1.7	5.00E-05	0.0018
cellular developmental process GO:0048869	2.00E+02	243	1.2	0.0004	0.008
cation binding GO:0043169	3.70E+02	428	1.2	0.0004	0.008
aging GO:0007568	56	77	1.4	0.00071	0.011
male anatomical structure morphogenesis GO:0090598	27	40	1.5	0.0017	0.022
response to endogenous stimulus GO:0009719	40	55	1.4	0.003	0.037
structural constituent of cuticle GO:0042302	37	51	1.4	0.0036	0.039
biological adhesion GO:0022610	22	33	1.5	0.0037	0.039
localization of cell GO:0051674	32	45	1.4	0.0046	0.043
neurogenesis GO:0022008	67	86	1.3	0.0049	0.043
nuclear outer membrane-endoplasmic reticulum membrane network GO:0042175	56	72	1.3	0.0065	0.052
negative regulation of metabolic process GO:0009892	1.30E+02	159	1.2	0.0079	0.059
regulation of neuron differentiation GO:0045664	27	38	1.4	0.0089	0.063
process utilizing autophagic mechanism GO:0061919	23	32	1.4	0.011	0.076
calcium ion binding GO:0005509	39	51	1.3	0.012	0.079
cell part morphogenesis GO:0032990	42	55	1.3	0.013	0.079
kinase binding GO:0019900	27	37	1.4	0.013	0.079
reproductive system development GO:0061458	28	38	1.4	0.015	0.083
nucleoside phosphate metabolic process GO:0006753	48	60	1.3	0.019	0.1

B.8 NHR-25:DAM L4 gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichment	P value	Q value
structural constituent of cuticle GO:0042302	38	71	1.9	3.00E-09	3.60E-07
molting cycle GO:0042303	24	51	2.1	4.80E-09	3.60E-07
organic acid metabolic process GO:0006082	1.00E+02	155	1.5	6.40E-09	3.60E-07
cation binding GO:0043169	3.80E+02	440	1.2	0.00037	0.011
post-embryonic development GO:0009791	1.10E+02	144	1.3	0.00063	0.015
response to endogenous stimulus GO:0009719	41	59	1.4	0.00073	0.015
supramolecular polymer GO:0099081	70	93	1.3	0.00077	0.015
cellular developmental process GO:0048869	2.10E+02	246	1.2	0.00098	0.015
aging GO:0007568	57	78	1.4	0.0011	0.015
nuclear outer membrane-endoplasmic reticulum membrane network GO:0042175	57	78	1.4	0.0011	0.015
nucleoside phosphate metabolic process GO:0006753	49	68	1.4	0.0011	0.015
cell junction GO:0030054	46	64	1.4	0.0012	0.015
kinase binding GO:0019900	28	41	1.5	0.0023	0.021
IRE1-mediated unfolded protein response GO:0036498	27	40	1.5	0.0025	0.021
localization of cell GO:0051674	33	47	1.4	0.0028	0.022
ribose phosphate metabolic process GO:0019693	37	51	1.4	0.0051	0.038
purine nucleotide metabolic process GO:0006163	35	48	1.4	0.0052	0.038
male anatomical structure morphogenesis GO:0090598	27	39	1.4	0.0054	0.038
response to topologically incorrect protein GO:0035966	45	60	1.3	0.0059	0.038
identical protein binding GO:0042802	28	39	1.4	0.0073	0.044
iron ion binding GO:0005506	27	38	1.4	0.0081	0.046
neurogenesis GO:0022008	69	86	1.2	0.011	0.059
multicellular organism growth GO:0035264	25	35	1.4	0.011	0.059
biological adhesion GO:0022610	23	32	1.4	0.011	0.059
actin filament-based process GO:0030029	41	53	1.3	0.016	0.076
RNA polymerase II regulatory region DNA binding GO:0001012	56	69	1.2	0.02	0.093

B.9 NHR-25:DAM L2-only gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichment	P value	Q value
supramolecular polymer GO:0099081	19	37	1.9	5.30E-05	0.0063

B.10 NHR-25:DAM L4-only gene-set complete GO-term analysis results

Term	Expected	Observed	Enrichment	P value	Q value
organic acid metabolic process GO:0006082	31	55	1.8	1.90E-05	0.0023
iron ion binding GO:0005506	8.3	19	2.3	0.00021	0.012
molting cycle GO:0042303	7.4	16	2.1	0.0012	0.046
structural constituent of cuticle GO:0042302	12	22	1.9	0.0012	0.046
IRE1-mediated unfolded protein response GO:0036498	8.3	16	1.9	0.0036	0.088
nucleoside phosphate metabolic process GO:0006753	15	25	1.7	0.0044	0.088

B.11 NHR-25:DAM L2 and NHR-25 ChIP-seq L2 common gene-set complete tissue-enrichment term analysis results

Term	Expected	Observed	Enrichment	P value	Q value
sex organ WBbt:0008422	1.90E+02	289	1.5	2.30E-12	6.70E-10
midbody WBbt:0005740	92	152	1.7	1.60E-10	2.40E-08
hermaphrodite WBbt:0007849	2.60E+02	348	1.3	5.70E-09	5.50E-07
somatic gonad WBbt:0005785	63	107	1.7	2.40E-08	1.70E-06
outer labial sensillum WBbt:0005501	3.30E+02	428	1.3	4.70E-08	2.70E-06
PVD WBbt:0006831	3.30E+02	419	1.3	9.40E-08	4.50E-06
epithelial system WBbt:0005730	5.80E+02	697	1.2	2.80E-07	1.20E-05
excretory duct cell WBbt:0004540	5.3	16	3	8.00E-06	0.00029
spermatheca WBbt:0005319	67	101	1.5	1.10E-05	0.00037
reproductive tract WBbt:0005744	2.10E+02	272	1.3	1.20E-05	0.00037
P5 WBbt:0006774	3.7	12	3.2	3.20E-05	0.00085
P6 WBbt:0006775	3.9	12	3.1	4.70E-05	0.0011
P7 WBbt:0006776	4	12	3	6.70E-05	0.0015
P10 WBbt:0006779	4.5	13	2.9	7.60E-05	0.0016
touch receptor neuron WBbt:0005237	2.20E+02	277	1.2	9.40E-05	0.0018
vulA WBbt:0006762	6.5	16	2.5	0.00013	0.0024
P8 WBbt:0006777	3.7	11	2.9	0.00016	0.0027
P1 WBbt:0006770	3.7	11	2.9	0.00016	0.0027
P11 WBbt:0004410	6.6	16	2.4	0.00017	0.0027
P4.p WBbt:0006892	7.3	17	2.3	0.0002	0.0028
P3.p WBbt:0006891	7.3	17	2.3	0.0002	0.0028
P2 WBbt:0006771	3.9	11	2.9	0.00022	0.0029
Q cell WBbt:0008598	3.9	11	2.9	0.00022	0.0029
P8.p WBbt:0006896	7.4	17	2.3	0.00024	0.0029
P4 WBbt:0006773	4	11	2.8	0.0003	0.0035
P9 WBbt:0006778	4	11	2.8	0.0003	0.0035
P5.p WBbt:0006893	8.2	18	2.2	0.00033	0.0035
P6.p WBbt:0006894	8.2	18	2.2	0.00033	0.0035
P7.p WBbt:0006895	8.3	18	2.2	0.00039	0.0039
vulC WBbt:0006765	7.3	16	2.2	0.00061	0.0059
P12 WBbt:0004409	5.4	13	2.4	0.00065	0.0061
gonadal primordium WBbt:0008366	3.30E+02	387	1.2	0.00096	0.0087
somatic cell WBbt:0008378	6.9	15	2.2	0.001	0.0088
spermathecal-uterine junction WBbt:0006756	7	15	2.1	0.0012	0.01
ABprappaa WBbt:0006350	4	10	2.5	0.0012	0.01
ABprappap WBbt:0006220	4.1	10	2.4	0.0016	0.013
vulB2 WBbt:0006764	6.6	14	2.1	0.0017	0.013
Ealp WBbt:0006546	4.8	11	2.3	0.0019	0.014
vulB1 WBbt:0006763	6.7	14	2.1	0.002	0.015
ABarpppaa WBbt:0006041	3.7	9	2.4	0.0028	0.02
P7.pp WBbt:0006984	3.7	9	2.4	0.0028	0.02
Epla WBbt:0006661	5	11	2.2	0.0029	0.02
Eprp WBbt:0006507	5.7	12	2.1	0.0031	0.021
HSN WBbt:0006830	15	25	1.7	0.0032	0.021
P7.pa WBbt:0006983	3.9	9	2.3	0.0035	0.023

Term	Expected	Observed	Enrichment	P value	Q value
Cpapa WBbt:0005962	5.1	11	2.2	0.0036	0.023
Eala WBbt:0006104	4.5	10	2.2	0.004	0.025
Cpaap WBbt:0006594	4.5	10	2.2	0.004	0.025
Cpaaa WBbt:0006212	4.5	10	2.2	0.004	0.025
excretory system WBbt:0005736	1.10E+02	135	1.2	0.0043	0.025
excretory secretory system WBbt:0006850	1.10E+02	135	1.2	0.0045	0.025
Cpapp WBbt:0005897	5.3	11	2.1	0.0052	0.029
Eplp WBbt:0006496	5.3	11	2.1	0.0052	0.029
ABprapapa WBbt:0006510	4.1	9	2.2	0.0055	0.03
ABpraapaa WBbt:0006108	4.1	9	2.2	0.0055	0.03
Epra WBbt:0006321	5.4	11	2	0.0063	0.032
vulD WBbt:0006766	8.3	15	1.8	0.007	0.036
Eara WBbt:0006161	4.9	10	2	0.0073	0.036
Earp WBbt:0006646	4.9	10	2	0.0073	0.036
hyp4 WBbt:0004687	6.2	12	1.9	0.0075	0.036
Psub1 WBbt:0006874	27	39	1.4	0.008	0.038
vulF WBbt:0006768	7.7	14	1.8	0.0084	0.039
ABplapppp WBbt:0006656	3.7	8	2.1	0.0094	0.043
ABplappaa WBbt:0006371	3.7	8	2.1	0.0094	0.043
ABprapaap WBbt:0006624	3.7	8	2.1	0.0094	0.043
ABpraappp WBbt:0006270	4.4	9	2	0.01	0.044
vulE WBbt:0006767	7.3	13	1.8	0.011	0.05
ABarpppap WBbt:0006251	3.9	8	2.1	0.012	0.05
Ear WBbt:0006370	3.9	8	2.1	0.012	0.05
ABarpaapp WBbt:0006620	3.9	8	2.1	0.012	0.05
ABpraapap WBbt:0006062	3.9	8	2.1	0.012	0.05
ABprapapp WBbt:0006290	4	8	2	0.014	0.057
ABarppapp WBbt:0006240	4	8	2	0.014	0.057
intestinal muscle WBbt:0005796	23	33	1.4	0.015	0.06
ABplappap WBbt:0006067	4.1	8	2	0.017	0.065
ABarpaapa WBbt:0005844	4.1	8	2	0.017	0.065
Epl WBbt:0006000	4.1	8	2	0.017	0.065
ABplaappp WBbt:0005948	4.1	8	2	0.017	0.065
amphid socket cell WBbt:0008379	4.8	9	1.9	0.017	0.065
ABpraappa WBbt:0006035	4.8	9	1.9	0.017	0.065
ABplaaapa WBbt:0006680	4.8	9	1.9	0.017	0.065
uterine muscle WBbt:0005342	15	23	1.5	0.017	0.065
hermaphrodite distal tip cell WBbt:0006863	22	31	1.4	0.02	0.069
Epr WBbt:0006547	4.2	8	1.9	0.02	0.069
anal depressor muscle WBbt:0004292	54	68	1.3	0.021	0.07
AB WBbt:0004015	14	20	1.5	0.027	0.09
ABplaappa WBbt:0006519	4.4	8	1.8	0.028	0.093
ABplpaapa WBbt:0006115	3.7	7	1.9	0.028	0.093
ABarppaa WBbt:0006465	3.7	7	1.9	0.028	0.093

B.12 NHR-25 ChIP-seq L2 genes not shared with TaDa tissue-enrichment analysis complete results

Term	Expected	Observed	Enrichment	P value	Q value
germ line WBbt:0005784	1.90E+03	2404	1.3	1.20E-40	3.40E-38
reproductive system WBbt:0005747	2.30E+03	2730	1.2	4.30E-26	6.20E-24
gonadal primordium WBbt:0008366	8.10E+02	987	1.2	8.70E-14	8.40E-12
nerve ring WBbt:0006749	2.50E+02	314	1.3	4.10E-07	3.00E-05
thermosensory neuron WBbt:0005838	6.20E+02	697	1.1	0.00022	0.013
sex organ WBbt:0008422	4.70E+02	532	1.1	0.00033	0.016
midbody WBbt:0005740	2.20E+02	264	1.2	0.00042	0.018
anal depressor muscle WBbt:0004292	1.30E+02	163	1.2	0.00048	0.018
hermaphrodite distal tip cell WBbt:0006863	53	73	1.4	0.00079	0.025
touch receptor neuron WBbt:0005237	5.40E+02	603	1.1	0.00091	0.026
lateral nerve cord WBbt:0006769	1.10E+02	142	1.2	0.00093	0.026
somatic gonad WBbt:0005785	1.50E+02	185	1.2	0.0013	0.031
dorsal nerve cord WBbt:0006750	1.70E+02	198	1.2	0.002	0.044
gonad arm WBbt:0008629	65	84	1.3	0.0025	0.052
Q cell WBbt:0008598	9.3	16	1.7	0.0043	0.084
PVD WBbt:0006831	8.00E+02	857	1.1	0.0045	0.084
spermatheca WBbt:0005319	1.60E+02	190	1.2	0.0047	0.084

B.13 *srf-3i1* L2 gene-set GO-term and tissue-enrichment analysis complete results

GO-term analysis

Term	Expected	Observed	Enrichment	P value	Q value
structural constituent of ribosome GO:0003735	25	68	2.7	8.40E-17	1.00E-14
peptide biosynthetic process GO:0043043	62	115	1.9	2.60E-12	1.50E-10
molting cycle GO:0042303	19	46	2.5	1.80E-10	7.30E-09
post-embryonic development GO:0009791	89	143	1.6	6.80E-10	2.00E-08
embryo development ending in birth or egg hatching GO:0009792	62	107	1.7	1.30E-09	3.00E-08
aging GO:0007568	44	78	1.8	7.70E-08	1.50E-06
structural constituent of cuticle GO:0042302	30	56	1.9	2.40E-07	4.20E-06
male anatomical structure morphogenesis GO:0090598	21	43	2	7.80E-07	1.20E-05
negative regulation of metabolic process GO:0009892	1.10E+02	148	1.4	1.40E-05	0.00019
cellular developmental process GO:0048869	1.60E+02	206	1.3	2.20E-05	0.00026
ribonucleoprotein complex subunit organization GO:0071826	18	34	1.8	6.50E-05	0.00071
ribonucleoprotein granule GO:0035770	18	32	1.8	0.00024	0.0024
multicellular organism growth GO:0035264	20	34	1.7	0.00028	0.0026
reproduction GO:0000003	1.40E+02	175	1.3	0.00047	0.004
negative regulation of transcription by RNA polymerase II GO:0000122	27	43	1.6	0.0005	0.004
biological adhesion GO:0022610	19	32	1.7	0.00053	0.004
regulation of cellular amide metabolic process GO:0034248	28	44	1.6	0.00055	0.004
RNA splicing via transesterification reactions GO:0000375	23	37	1.6	0.0006	0.004
reproductive system development GO:0061458	22	36	1.6	0.001	0.0064
regulation of protein metabolic process GO:0051246	92	119	1.3	0.0011	0.0066
actin binding GO:0003779	21	32	1.6	0.0028	0.016
post-embryonic animal organ development GO:0048569	32	46	1.4	0.003	0.016
development of primary sexual characteristics GO:0045137	19	29	1.5	0.0046	0.024
biosynthetic process GO:0009058	4.00E+02	441	1.1	0.0054	0.027
macromolecule biosynthetic process GO:0009059	3.10E+02	346	1.1	0.0068	0.033
modification-dependent macromolecule catabolic process GO:0043632	43	56	1.3	0.015	0.067
neuron development GO:0048666	34	44	1.3	0.022	0.096
response to endogenous stimulus GO:0009719	32	42	1.3	0.023	0.099

Tissue enrichment analysis

Term	Expected	Observed	Enrichment	P value	Q value
sex organ WBbt:0008422	2.90E+02	424	1.5	6.00E-18	1.70E-15
midbody WBbt:0005740	1.40E+02	208	1.5	3.00E-11	4.30E-09
hermaphrodite WBbt:0007849	3.80E+02	500	1.3	5.60E-11	5.40E-09
dorsal nerve cord WBbt:0006750	1.00E+02	161	1.6	3.90E-10	2.90E-08
anal depressor muscle WBbt:0004292	80	131	1.6	1.30E-09	7.60E-08
lateral nerve cord WBbt:0006769	69	117	1.7	1.70E-09	8.10E-08
PVD WBbt:0006831	4.80E+02	603	1.2	3.20E-09	1.30E-07
outer labial sensillum WBbt:0005501	4.90E+02	608	1.2	1.20E-08	4.20E-07
touch receptor neuron WBbt:0005237	3.30E+02	413	1.3	4.40E-07	1.40E-05
nerve ring WBbt:0006749	1.50E+02	207	1.4	5.40E-07	1.60E-05
somatic gonad WBbt:0005785	93	137	1.5	9.80E-07	2.60E-05
ABarpppap WBbt:0006251	5.7	17	3	1.50E-06	3.60E-05
excretory duct cell WBbt:0004540	7.9	21	2.7	1.60E-06	3.60E-05
Ealp WBbt:0006546	7	19	2.7	3.40E-06	7.10E-05
ABpraappp WBbt:0006270	6.5	18	2.8	3.90E-06	7.60E-05
ABprapapa WBbt:0006510	6	17	2.8	4.40E-06	8.00E-05
ABarpppaa WBbt:0006041	5.5	16	2.9	4.80E-06	8.20E-05
P3.p WBbt:0006891	11	25	2.3	5.20E-06	8.40E-05
P4.p WBbt:0006892	11	25	2.3	5.20E-06	8.40E-05
P6.p WBbt:0006894	12	27	2.2	6.30E-06	9.10E-05
P5.p WBbt:0006893	12	27	2.2	6.30E-06	9.10E-05
ABplpappa WBbt:0006232	6.2	17	2.7	7.20E-06	9.50E-05
P8.p WBbt:0006896	11	25	2.3	7.30E-06	9.50E-05
Epla WBbt:0006661	7.4	19	2.6	8.40E-06	0.0001
P7.p WBbt:0006895	12	27	2.2	8.60E-06	0.0001
tail WBbt:0005741	3.30E+02	403	1.2	1.10E-05	0.00012
ABarppapp WBbt:0006240	5.9	16	2.7	1.30E-05	0.00014
ABprapapp WBbt:0006290	5.9	16	2.7	1.30E-05	0.00014
spermatheca WBbt:0005319	99	138	1.4	1.30E-05	0.00014
ABpraappa WBbt:0006035	7	18	2.6	1.50E-05	0.00015
ABprappap WBbt:0006220	6	16	2.7	2.10E-05	0.0002
Earp WBbt:0006646	7.2	18	2.5	2.30E-05	0.00021
ABprapaap WBbt:0006624	5.5	15	2.7	2.50E-05	0.00022
MSpppp WBbt:0006409	5.5	15	2.7	2.50E-05	0.00022
ABplappaa WBbt:0006371	5.5	15	2.7	2.50E-05	0.00022
Eplp WBbt:0006496	7.9	19	2.4	2.80E-05	0.00022
P10 WBbt:0006779	6.7	17	2.5	2.80E-05	0.00022
P11 WBbt:0004410	9.7	22	2.3	2.80E-05	0.00022
Epra WBbt:0006321	8	19	2.4	4.00E-05	0.0003
epithelial system WBbt:0005730	8.60E+02	963	1.1	5.60E-05	0.00041
ABprappaa WBbt:0006350	5.9	15	2.6	6.10E-05	0.00043
ABplapapp WBbt:0006413	5.9	15	2.6	6.10E-05	0.00043
ABplaaapa WBbt:0006680	7	17	2.4	6.20E-05	0.00043
ABplpappp WBbt:0006390	6.5	16	2.4	7.60E-05	0.0005
ABprappp WBbt:0006702	7.2	17	2.4	9.00E-05	0.00058
ABarpaapa WBbt:0005844	6	15	2.5	9.30E-05	0.00058
ABpraapaa WBbt:0006108	6	15	2.5	9.30E-05	0.00058
ABplappap WBbt:0006067	6	15	2.5	9.30E-05	0.00058
Eala WBbt:0006104	6.7	16	2.4	0.00011	0.00066
ABplpaapa WBbt:0006115	5.5	14	2.5	0.00011	0.00066
ABarpaapp WBbt:0006620	5.7	14	2.5	0.00017	0.00096
ABplaaapp WBbt:0006136	5.7	14	2.5	0.00017	0.00096
Cpapa WBbt:0005962	7.5	17	2.3	0.00018	0.00099
gonadal primordium WBbt:0008366	4.90E+02	563	1.1	0.00024	0.0013
P7 WBbt:0006776	5.9	14	2.4	0.00025	0.0013
Eprp WBbt:0006507	8.4	18	2.1	0.00027	0.0014
Eara WBbt:0006161	7.2	16	2.2	0.00032	0.0016
Cpapp WBbt:0005897	7.9	17	2.2	0.00034	0.0017
ABplpppap WBbt:0006665	6	14	2.3	0.00036	0.0018

Appendix

Term	Expected	Observed	Enrichment	P value	Q value
MSaapp WBbt:0006425	6	14	2.3	0.00036	0.0018
ABplppppp WBbt:0006574	6	14	2.3	0.00036	0.0018
ABprpapaa WBbt:0006446	6	14	2.3	0.00036	0.0018
Cpaaa WBbt:0006212	6.7	15	2.2	0.00041	0.0019
ABprpaapa WBbt:0006047	5.5	13	2.4	0.00046	0.0021
ABpraapp WBbt:0006335	5.5	13	2.4	0.00046	0.0021
ABplapppp WBbt:0006656	5.5	13	2.4	0.00046	0.0021
P5 WBbt:0006774	5.5	13	2.4	0.00046	0.0021
P8 WBbt:0006777	5.5	13	2.4	0.00046	0.0021
ABalaapp WBbt:0006553	5.5	13	2.4	0.00046	0.0021
P12 WBbt:0004409	8	17	2.1	0.00046	0.0021
ABalaaap WBbt:0005982	6.2	14	2.3	0.00052	0.0021
ABplaaaap WBbt:0006625	5.7	13	2.3	0.00066	0.0026
P6 WBbt:0006775	5.7	13	2.3	0.00066	0.0026
MSppaa WBbt:0006531	5.7	13	2.3	0.00066	0.0026
ABpraapap WBbt:0006062	5.7	13	2.3	0.00066	0.0026
ABalaapa WBbt:0006130	5.7	13	2.3	0.00066	0.0026
ABplappp WBbt:0006470	6.4	14	2.2	0.00072	0.0027
ABprappa WBbt:0006269	6.4	14	2.2	0.00072	0.0027
ABalppaa WBbt:0005944	7	15	2.1	0.00077	0.0028
anal sphincter muscle WBbt:0005798	22	36	1.6	0.00081	0.0029
P9 WBbt:0006778	5.9	13	2.2	0.00092	0.0033
hermaphrodite distal tip cell WBbt:0006863	32	49	1.5	0.00095	0.0033
ABalaaaa WBbt:0006427	6.5	14	2.1	0.00099	0.0035
Eal WBbt:0006441	6	13	2.2	0.0013	0.0044
Epl WBbt:0006000	6	13	2.2	0.0013	0.0044
ABprpppap WBbt:0006237	6	13	2.2	0.0013	0.0044
ABplaaaaa WBbt:0006348	6	13	2.2	0.0013	0.0044
Cpaap WBbt:0006594	6.7	14	2.1	0.0013	0.0044
Caaap WBbt:0006267	7.4	15	2	0.0014	0.0045
anchor cell WBbt:0004522	13	23	1.8	0.0016	0.0052
somatic cell WBbt:0008378	10	19	1.9	0.0016	0.0052
P1 WBbt:0006770	5.5	12	2.2	0.0016	0.0052
MSpaap WBbt:0005878	5.5	12	2.2	0.0016	0.0052
ABprpppa WBbt:0005943	5.5	12	2.2	0.0016	0.0052
ABplpppaa WBbt:0006222	5.5	12	2.2	0.0016	0.0052
ABarppaa WBbt:0006465	5.5	12	2.2	0.0016	0.0052
Caaaa WBbt:0005899	6.2	13	2.1	0.0017	0.0052
ABprpappp WBbt:0005847	6.2	13	2.1	0.0017	0.0052
Epr WBbt:0006547	6.2	13	2.1	0.0017	0.0052
Ear WBbt:0006370	5.7	12	2.1	0.0023	0.0066
ABplppppa WBbt:0006352	5.7	12	2.1	0.0023	0.0066
P2 WBbt:0006771	5.7	12	2.1	0.0023	0.0066
ABprpapa WBbt:0006259	5.7	12	2.1	0.0023	0.0066
ABalappa WBbt:0006157	5.7	12	2.1	0.0023	0.0066
Caapa WBbt:0006123	6.4	13	2	0.0023	0.0066
ABprpappa WBbt:0006088	6.4	13	2	0.0023	0.0066
ABprpppaa WBbt:0006552	6.4	13	2	0.0023	0.0066
hyp4 WBbt:0004687	9.2	17	1.8	0.0028	0.0075
excretory system WBbt:0005736	1.60E+02	193	1.2	0.0029	0.0076
body wall WBbt:0005742	8.5	16	1.9	0.0029	0.0077
excretory secretory system WBbt:0006850	1.60E+02	193	1.2	0.003	0.0078
ABalppap WBbt:0006112	5.9	12	2	0.0031	0.0079
P4 WBbt:0006773	5.9	12	2	0.0031	0.0079
MSaaap WBbt:0006160	5.9	12	2	0.0031	0.0079
ABprppppp WBbt:0005983	5.9	12	2	0.0031	0.0079
ABpraaaa WBbt:0006442	6.5	13	2	0.0031	0.0079
ABalpppa WBbt:0006649	6.5	13	2	0.0031	0.0079
ABplaappa WBbt:0006519	6.5	13	2	0.0031	0.0079
intestinal muscle WBbt:0005796	34	49	1.4	0.0033	0.008

Appendix

Term	Expected	Observed	Enrichment	P value	Q value
gonad arm WBbt:0008629	40	55	1.4	0.0037	0.0089
pm7 WBbt:0003721	6.7	13	1.9	0.004	0.0097
ABalpapa WBbt:0006573	6.7	13	1.9	0.004	0.0097
ABplaappp WBbt:0005948	6	12	2	0.0041	0.0097
vulA WBbt:0006762	9.5	17	1.8	0.0043	0.01
uterine muscle WBbt:0005342	23	34	1.5	0.0045	0.011
ABarpaaa WBbt:0006398	7.5	14	1.9	0.005	0.012
ABalpaaa WBbt:0006557	6.9	13	1.9	0.0052	0.012
anal region WBbt:0006919	6.9	13	1.9	0.0052	0.012
ABaraaaa WBbt:0006360	6.2	12	1.9	0.0053	0.012
ABarapaa WBbt:0006515	6.2	12	1.9	0.0053	0.012
ABalpaap WBbt:0005934	6.2	12	1.9	0.0053	0.012
ABprppaa WBbt:0005984	6.2	12	1.9	0.0053	0.012
MSapa WBbt:0005898	5.5	11	2	0.0053	0.012
ABpraaap WBbt:0006534	5.5	11	2	0.0053	0.012
hyp6 WBbt:0004679	9.9	17	1.7	0.0065	0.014
MSapp WBbt:0006036	5.7	11	1.9	0.007	0.015
ABarappp WBbt:0006524	5.7	11	1.9	0.007	0.015
MSappa WBbt:0006717	5.7	11	1.9	0.007	0.015
ABarpapa WBbt:0006603	5.7	11	1.9	0.007	0.015
ABprapap WBbt:0006678	5.7	11	1.9	0.007	0.015
ABprpppp WBbt:0006179	5.7	11	1.9	0.007	0.015
MSappp WBbt:0006125	5.7	11	1.9	0.007	0.015
vulC WBbt:0006765	11	18	1.7	0.0072	0.015
Cppa WBbt:0006168	7.9	14	1.8	0.0079	0.016
ABpraapa WBbt:0006302	6.5	12	1.8	0.0087	0.017
HSN WBbt:0006830	22	32	1.5	0.009	0.018
hyp5 WBbt:0004685	8.7	15	1.7	0.0091	0.018
pm3 WBbt:0003740	6.7	12	1.8	0.011	0.022
ABplaapaa WBbt:0005866	6	11	1.8	0.012	0.022
ABplaapap WBbt:0005887	6	11	1.8	0.012	0.022
vulB2 WBbt:0006764	9.7	16	1.6	0.012	0.023
Capa WBbt:0006444	6.9	12	1.7	0.014	0.026
vulB1 WBbt:0006763	9.9	16	1.6	0.015	0.028
vulD WBbt:0006766	12	19	1.6	0.015	0.028
ABarappa WBbt:0006005	5.5	10	1.8	0.015	0.029
ABprppap WBbt:0006346	6.4	11	1.7	0.018	0.034
ABplppap WBbt:0006028	6.4	11	1.7	0.018	0.034
ABplppaa WBbt:0006170	6.4	11	1.7	0.018	0.034
ABaraaap WBbt:0005861	5.7	10	1.8	0.019	0.035
Capp WBbt:0006098	6.5	11	1.7	0.022	0.04
Cppp WBbt:0006268	6.5	11	1.7	0.022	0.04
excretory cell WBbt:0005812	1.50E+02	167	1.1	0.025	0.045
MSpapp WBbt:0006201	6	10	1.7	0.029	0.052
ABaraapa WBbt:0005853	6	10	1.7	0.029	0.052
pm5 WBbt:0003737	6.9	11	1.6	0.033	0.057
pm6 WBbt:0003724	7.7	12	1.6	0.035	0.062
ABaraapp WBbt:0006153	6.2	10	1.6	0.036	0.062
P7.pp WBbt:0006984	5.5	9	1.6	0.039	0.067
ABplpppp WBbt:0006647	6.4	10	1.6	0.043	0.074
reproductive tract WBbt:0005744	3.10E+02	341	1.1	0.046	0.078
spermathecal-uterine junction WBbt:0006756	10	15	1.4	0.046	0.079
P7.pa WBbt:0006983	5.7	9	1.6	0.047	0.08
Z3 WBbt:0004575	8.9	13	1.5	0.05	0.084
Z2 WBbt:0004576	8.9	13	1.5	0.05	0.084
vulF WBbt:0006768	11	16	1.4	0.054	0.089
uterine seam cell WBbt:0006789	7.4	11	1.5	0.054	0.089
ABplpapp WBbt:0006420	5.9	9	1.5	0.057	0.093
vulE WBbt:0006767	11	15	1.4	0.06	0.098

B.14 *srf-3i1* L4 gene-set GO-term and tissue-enrichment analysis complete results

GO-term analysis

Term	Expected	Observed	Enrichment	P value	Q value
structural constituent of cuticle GO:0042302	33	95	2.9	1.30E-26	1.60E-24
structural constituent of ribosome GO:0003735	29	76	2.6	4.50E-19	2.70E-17
peptide biosynthetic process GO:0043043	70	140	2	2.20E-18	8.90E-17
molting cycle GO:0042303	21	59	2.8	9.40E-17	2.80E-15
post-embryonic development GO:0009791	1.00E+02	161	1.6	3.30E-11	7.90E-10
embryo development ending in birth or egg hatching GO:0009792	70	118	1.7	4.40E-10	8.70E-09
ribonucleoprotein complex subunit organization GO:0071826	21	45	2.2	1.60E-08	2.80E-07
aging GO:0007568	50	84	1.7	1.50E-07	2.20E-06
male anatomical structure morphogenesis GO:0090598	24	47	2	4.70E-07	6.20E-06
post-embryonic animal organ development GO:0048569	36	62	1.7	1.60E-06	1.90E-05
biosynthetic process GO:0009058	4.50E+02	525	1.2	1.20E-05	0.00013
regulation of cellular amide metabolic process GO:0034248	31	53	1.7	1.90E-05	0.00019
multicellular organism growth GO:0035264	22	39	1.8	5.70E-05	0.00052
macromolecule biosynthetic process GO:0009059	3.50E+02	409	1.2	7.60E-05	0.00065
modification-dependent macromolecule catabolic process GO:0043632	49	70	1.4	0.00038	0.003
regulation of protein metabolic process GO:0051246	1.00E+02	134	1.3	0.00051	0.0038
reproduction GO:0000003	1.60E+02	193	1.2	0.00068	0.0047
reproductive system development GO:0061458	25	40	1.6	0.00068	0.0047
RNA splicing via transesterification reactions GO:0000375	25	40	1.6	0.00081	0.0051
response to topologically incorrect protein GO:0035966	40	57	1.4	0.0013	0.0079
IRE1-mediated unfolded protein response GO:0036498	24	37	1.5	0.0016	0.0088
regulation of neuron differentiation GO:0045664	25	38	1.5	0.0018	0.01
organic cyclic compound metabolic process GO:1901360	4.30E+02	486	1.1	0.0025	0.013
cellular aromatic compound metabolic process GO:0006725	4.30E+02	476	1.1	0.0026	0.013
heterocycle metabolic process GO:0046483	4.20E+02	474	1.1	0.0031	0.015
amide transport GO:0042886	93	116	1.3	0.0033	0.015
cellular macromolecule localization GO:0070727	1.10E+02	132	1.2	0.0034	0.015
development of primary sexual characteristics GO:0045137	21	32	1.5	0.0041	0.017
cellular developmental process GO:0048869	1.80E+02	210	1.2	0.0042	0.017
membrane-enclosed lumen GO:0031974	2.00E+02	230	1.2	0.0063	0.025
small GTPase binding GO:0031267	20	30	1.5	0.0066	0.025
organic acid metabolic process GO:0006082	90	111	1.2	0.007	0.026
biological adhesion GO:0022610	21	31	1.5	0.0077	0.028
envelope GO:0031975	69	86	1.3	0.0087	0.031
organelle GO:0043226	1.00E+03	1088	1.1	0.0094	0.032
ribonucleoprotein granule GO:0035770	20	29	1.4	0.014	0.047
vesicle GO:0031982	69	85	1.2	0.016	0.053
regulation of nucleobase-containing compound metabolic process GO:0019219	2.20E+02	243	1.1	0.02	0.061
oviposition GO:0018991	32	42	1.3	0.02	0.062
response to nitrogen compound GO:1901698	25	34	1.4	0.021	0.062
ribose phosphate metabolic process GO:0019693	33	42	1.3	0.033	0.096
negative regulation of metabolic process GO:0009892	1.20E+02	139	1.1	0.033	0.096

Tissue enrichment analysis

Term	Expected	Observed	Enrichment	P value	Q value
midbody WBbt:0005740	1.50E+02	264	1.8	1.10E-22	3.30E-20
sex organ WBbt:0008422	3.20E+02	460	1.5	2.80E-18	4.10E-16
epithelial system WBbt:0005730	9.50E+02	1168	1.2	3.00E-15	2.90E-13
hermaphrodite WBbt:0007849	4.20E+02	558	1.3	3.90E-13	2.80E-11
PVD WBbt:0006831	5.40E+02	677	1.3	1.10E-11	6.40E-10
touch receptor neuron WBbt:0005237	3.60E+02	479	1.3	4.30E-11	2.10E-09
outer labial sensillum WBbt:0005501	5.40E+02	680	1.2	1.20E-10	4.90E-09
somatic gonad WBbt:0005785	1.00E+02	164	1.6	1.40E-10	5.00E-09
lateral nerve cord WBbt:0006769	77	128	1.7	3.60E-10	1.20E-08
gonadal primordium WBbt:0008366	5.40E+02	668	1.2	2.50E-09	7.30E-08
P10 WBbt:0006779	7.4	23	3.1	7.00E-09	1.90E-07
dorsal nerve cord WBbt:0006750	1.10E+02	168	1.5	9.40E-09	2.30E-07
ABarpppap WBbt:0006251	6.3	20	3.2	3.00E-08	6.70E-07
vulA WBbt:0006762	11	28	2.7	3.40E-08	7.00E-07
nerve ring WBbt:0006749	1.70E+02	231	1.4	3.70E-08	7.10E-07
P4.p WBbt:0006892	12	30	2.5	5.00E-08	9.10E-07
P3.p WBbt:0006891	12	30	2.5	5.00E-08	9.10E-07
P11 WBbt:0004410	11	28	2.6	5.60E-08	9.10E-07
P7 WBbt:0006776	6.5	20	3.1	6.20E-08	9.40E-07
P8.p WBbt:0006896	12	30	2.5	7.90E-08	1.10E-06
P6.p WBbt:0006894	13	32	2.4	1.00E-07	1.40E-06
P5.p WBbt:0006893	13	32	2.4	1.00E-07	1.40E-06
P7.p WBbt:0006895	13	32	2.4	1.50E-07	1.90E-06
P6 WBbt:0006775	6.3	19	3	2.00E-07	2.40E-06
ABarppapp WBbt:0006240	6.5	19	2.9	3.90E-07	4.50E-06
P9 WBbt:0006778	6.5	19	2.9	3.90E-07	4.50E-06
tail WBbt:0005741	3.70E+02	452	1.2	4.70E-07	5.10E-06
P5 WBbt:0006774	6.1	18	3	6.50E-07	6.70E-06
P8 WBbt:0006777	6.1	18	3	6.50E-07	6.70E-06
ABarpppaa WBbt:0006041	6.1	18	3	6.50E-07	6.70E-06
gonad arm WBbt:0008629	44	74	1.7	6.50E-07	6.70E-06
P12 WBbt:0004409	8.9	23	2.6	7.70E-07	6.90E-06
ABpraappp WBbt:0006270	7.2	20	2.8	7.70E-07	6.90E-06
vulB2 WBbt:0006764	11	26	2.4	9.80E-07	8.40E-06
anal depressor muscle WBbt:0004292	88	129	1.5	1.50E-06	1.20E-05
vulB1 WBbt:0006763	11	26	2.4	1.50E-06	1.20E-05
vulD WBbt:0006766	13	30	2.2	1.90E-06	1.50E-05
excretory duct cell WBbt:0004540	8.7	22	2.5	2.10E-06	1.60E-05
ABprapapp WBbt:0006290	6.5	18	2.8	2.20E-06	1.60E-05
Epla WBbt:0006661	8.1	21	2.6	2.20E-06	1.60E-05
Epra WBbt:0006321	8.9	22	2.5	3.30E-06	2.30E-05
P1 WBbt:0006770	6.1	17	2.8	3.70E-06	2.60E-05
ABprapaap WBbt:0006624	6.1	17	2.8	3.70E-06	2.60E-05
ABplappaa WBbt:0006371	6.1	17	2.8	3.70E-06	2.60E-05
Ealp WBbt:0006546	7.8	20	2.6	3.70E-06	2.60E-05
ABpraappa WBbt:0006035	7.8	20	2.6	3.70E-06	2.60E-05
ABarpaapa WBbt:0005844	6.7	18	2.7	3.80E-06	2.60E-05
ABprapapa WBbt:0006510	6.7	18	2.7	3.80E-06	2.60E-05
ABprappap WBbt:0006220	6.7	18	2.7	3.80E-06	2.60E-05
ABarpaapp WBbt:0006620	6.3	17	2.7	6.50E-06	3.80E-05
P2 WBbt:0006771	6.3	17	2.7	6.50E-06	3.80E-05
ABplpappa WBbt:0006232	6.8	18	2.6	6.50E-06	3.80E-05
hermaphrodite distal tip cell WBbt:0006863	36	60	1.7	9.80E-06	5.40E-05
ABprappaa WBbt:0006350	6.5	17	2.6	1.10E-05	6.00E-05
P4 WBbt:0006773	6.5	17	2.6	1.10E-05	6.00E-05
ABplapapp WBbt:0006413	6.5	17	2.6	1.10E-05	6.00E-05
ABpraapaa WBbt:0006108	6.7	17	2.6	1.80E-05	9.30E-05
ABplappap WBbt:0006067	6.7	17	2.6	1.80E-05	9.30E-05
ABplpaapa WBbt:0006115	6.1	16	2.6	1.90E-05	9.30E-05
MSpppp WBbt:0006409	6.1	16	2.6	1.90E-05	9.30E-05

Appendix

Term	Expected	Observed	Enrichment	P value	Q value
Eara WBbt:0006161	7.9	19	2.4	2.50E-05	0.00012
Earp WBbt:0006646	7.9	19	2.4	2.50E-05	0.00012
Cpaaa WBbt:0006212	7.4	18	2.4	2.70E-05	0.00013
ABplaaapp WBbt:0006136	6.3	16	2.5	3.10E-05	0.00014
vulC WBbt:0006765	12	25	2.1	3.20E-05	0.00014
Eplp WBbt:0006496	8.7	20	2.3	3.30E-05	0.00015
spermatheca WBbt:0005319	1.10E+02	146	1.3	6.20E-05	0.00027
ABplaaapa WBbt:0006680	7.8	18	2.3	6.40E-05	0.00027
ABplpappp WBbt:0006390	7.2	17	2.4	7.10E-05	0.0003
ABplppppp WBbt:0006574	6.7	16	2.4	7.90E-05	0.00033
ABplapppp WBbt:0006656	6.1	15	2.5	8.60E-05	0.00035
ABprpaapa WBbt:0006047	6.1	15	2.5	8.60E-05	0.00035
Eprp WBbt:0006507	9.2	20	2.2	0.0001	0.0004
Eala WBbt:0006104	7.4	17	2.3	0.00011	0.00042
Cpaap WBbt:0006594	7.4	17	2.3	0.00011	0.00042
ABplaaaap WBbt:0006625	6.3	15	2.4	0.00013	0.00051
ABpraapap WBbt:0006062	6.3	15	2.4	0.00013	0.00051
MSppaa WBbt:0006531	6.3	15	2.4	0.00013	0.00051
Caaap WBbt:0006267	8.1	18	2.2	0.00014	0.00051
anal region WBbt:0006919	7.6	17	2.2	0.00016	0.00058
excretory system WBbt:0005736	1.80E+02	221	1.2	0.00018	0.00066
excretory secretory system WBbt:0006850	1.80E+02	221	1.2	0.0002	0.00069
Cpapa WBbt:0005962	8.3	18	2.2	0.0002	0.00069
ABalppaa WBbt:0005944	7.8	17	2.2	0.00023	0.00079
ABprpapaa WBbt:0006446	6.7	15	2.3	0.00031	0.001
ABplpppap WBbt:0006665	6.7	15	2.3	0.00031	0.001
ABplaaaaa WBbt:0006348	6.7	15	2.3	0.00031	0.001
ABprpppap WBbt:0006237	6.7	15	2.3	0.00031	0.001
ABprappp WBbt:0006702	7.9	17	2.1	0.00033	0.0011
ABpraapp WBbt:0006335	6.1	14	2.3	0.00035	0.0011
Cpapp WBbt:0005897	8.7	18	2.1	0.00039	0.0012
ABalpaap WBbt:0005934	6.8	15	2.2	0.00045	0.0014
ABprppaa WBbt:0005984	6.8	15	2.2	0.00045	0.0014
Caaaa WBbt:0005899	6.8	15	2.2	0.00045	0.0014
ABplpppppa WBbt:0006352	6.3	14	2.2	0.00052	0.0016
ABprpapa WBbt:0006259	6.3	14	2.2	0.00052	0.0016
ABalaapa WBbt:0006130	6.3	14	2.2	0.00052	0.0016
anchor cell WBbt:0004522	14	26	1.8	0.00053	0.0016
ABalpaaa WBbt:0006557	7.6	16	2.1	0.00054	0.0016
Psub1 WBbt:0006874	44	64	1.4	0.0006	0.0017
ABprappa WBbt:0006269	7	15	2.1	0.00064	0.0018
ABprpppaa WBbt:0006552	7	15	2.1	0.00064	0.0018
ABplappp WBbt:0006470	7	15	2.1	0.00064	0.0018
vulE WBbt:0006767	12	22	1.9	0.00074	0.0021
MSaaap WBbt:0006160	6.5	14	2.2	0.00075	0.0021
ABalppap WBbt:0006112	6.5	14	2.2	0.00075	0.0021
vulF WBbt:0006768	13	23	1.8	0.00077	0.0021
somatic cell WBbt:0008378	11	21	1.9	0.0009	0.0024
ABplaappa WBbt:0006519	7.2	15	2.1	0.0009	0.0024
ABalaaaa WBbt:0006427	7.2	15	2.1	0.0009	0.0024
ABplaappp WBbt:0005948	6.7	14	2.1	0.0011	0.0028
MSaapp WBbt:0006425	6.7	14	2.1	0.0011	0.0028
spermathecal-uterine junction WBbt:0006756	11	21	1.8	0.0012	0.003
ABalpapa WBbt:0006573	7.4	15	2	0.0012	0.0032
ABarppaa WBbt:0006465	6.1	13	2.1	0.0013	0.0032
MSpaap WBbt:0005878	6.1	13	2.1	0.0013	0.0032
ABpraaap WBbt:0006534	6.1	13	2.1	0.0013	0.0032
ABplpppaa WBbt:0006222	6.1	13	2.1	0.0013	0.0032
ABprpppa WBbt:0005943	6.1	13	2.1	0.0013	0.0032
ABalaapp WBbt:0006553	6.1	13	2.1	0.0013	0.0032

Appendix

Term	Expected	Observed	Enrichment	P value	Q value
hyp4 WBbt:0004687	10	19	1.9	0.0013	0.0032
ABarapaa WBbt:0006515	6.8	14	2	0.0015	0.0036
ABaraaaa WBbt:0006360	6.8	14	2	0.0015	0.0036
ABalaaap WBbt:0005982	6.8	14	2	0.0015	0.0036
ABalappa WBbt:0006157	6.3	13	2.1	0.0018	0.0041
MSappa WBbt:0006717	6.3	13	2.1	0.0018	0.0041
ABarappp WBbt:0006524	6.3	13	2.1	0.0018	0.0041
ABarpaaa WBbt:0006398	8.3	16	1.9	0.0019	0.0042
ABplppaa WBbt:0006170	7	14	2	0.0021	0.0046
ABplppap WBbt:0006028	7	14	2	0.0021	0.0046
ABprppap WBbt:0006346	7	14	2	0.0021	0.0046
Caapa WBbt:0006123	7	14	2	0.0021	0.0046
ABprppppp WBbt:0005983	6.5	13	2	0.0025	0.0054
ABpraaaa WBbt:0006442	7.2	14	1.9	0.0028	0.006
ABalpppa WBbt:0006649	7.2	14	1.9	0.0028	0.006
body wall WBbt:0005742	9.4	17	1.8	0.0033	0.0071
ABplaapaa WBbt:0005866	6.7	13	2	0.0034	0.0071
MSpapp WBbt:0006201	6.7	13	2	0.0034	0.0071
ABplaapap WBbt:0005887	6.7	13	2	0.0034	0.0071
hyp6 WBbt:0004679	11	19	1.7	0.0035	0.0073
ABarappa WBbt:0006005	6.1	12	2	0.0041	0.0084
ABprpappp WBbt:0005847	6.8	13	1.9	0.0045	0.0092
reproductive tract WBbt:0005744	3.50E+02	390	1.1	0.0049	0.0099
ABaraaap WBbt:0005861	6.3	12	1.9	0.0055	0.011
ABarpapa WBbt:0006603	6.3	12	1.9	0.0055	0.011
MSappp WBbt:0006125	6.3	12	1.9	0.0055	0.011
ABprapap WBbt:0006678	6.3	12	1.9	0.0055	0.011
ABprpppp WBbt:0006179	6.3	12	1.9	0.0055	0.011
ABprpappa WBbt:0006088	7	13	1.9	0.0059	0.012
amphid socket cell WBbt:0008379	7.8	14	1.8	0.0063	0.012
Capp WBbt:0006098	7.2	13	1.8	0.0077	0.015
ABpraapa WBbt:0006302	7.2	13	1.8	0.0077	0.015
Cppa WBbt:0006168	8.7	15	1.7	0.0082	0.015
ABaraapa WBbt:0005853	6.7	12	1.8	0.0096	0.018
Eal WBbt:0006441	6.7	12	1.8	0.0096	0.018
Epl WBbt:0006000	6.7	12	1.8	0.0096	0.018
MSapa WBbt:0005898	6.1	11	1.8	0.012	0.022
P7.pp WBbt:0006984	6.1	11	1.8	0.012	0.022
ABaraapp WBbt:0006153	6.8	12	1.8	0.012	0.022
Epr WBbt:0006547	6.8	12	1.8	0.012	0.022
Capa WBbt:0006444	7.6	13	1.7	0.013	0.023
excretory cell WBbt:0005812	1.60E+02	186	1.2	0.013	0.024
Ear WBbt:0006370	6.3	11	1.8	0.015	0.027
P7.pa WBbt:0006983	6.3	11	1.8	0.015	0.027
MSapp WBbt:0006036	6.3	11	1.8	0.015	0.027
ABplpapa WBbt:0006087	7	12	1.7	0.016	0.027
ABplpapp WBbt:0006420	6.5	11	1.7	0.019	0.034
Cppp WBbt:0006268	7.2	12	1.7	0.019	0.034
hyp5 WBbt:0004685	9.6	15	1.6	0.022	0.038
ABplpaap WBbt:0006077	6.7	11	1.7	0.024	0.041
ABplpaaa WBbt:0006315	6.7	11	1.7	0.024	0.041
ABplpppa WBbt:0006423	6.8	11	1.6	0.03	0.051
ABplpppp WBbt:0006647	7	11	1.6	0.037	0.062
ABprpaaa WBbt:0006167	6.5	10	1.5	0.046	0.077
uterine seam cell WBbt:0006789	8.1	12	1.5	0.051	0.084
C WBbt:0003810	8.1	12	1.5	0.051	0.084
Psub3 WBbt:0006875	12	17	1.4	0.051	0.084
rectal valve cell WBbt:0005797	8.3	12	1.4	0.06	0.097

B.15 *dpy-7syn1* L2 gene-set GO-term and tissue-enrichment analysis complete results

GO-terms

Term	Expected	Observed	Enrichment	P value	Q value
peptide biosynthetic process GO:0043043	74	143	1.9	1.30E-17	1.60E-15
molting cycle GO:0042303	22	59	2.7	1.20E-15	7.20E-14
structural constituent of ribosome GO:0003735	30	71	2.3	1.20E-14	4.80E-13
structural constituent of cuticle GO:0042302	35	76	2.2	3.60E-13	1.10E-11
aging GO:0007568	53	102	1.9	4.80E-13	1.20E-11
post-embryonic development GO:0009791	1.10E+02	171	1.6	2.90E-12	5.70E-11
ribonucleoprotein complex subunit organization GO:0071826	22	46	2.1	2.80E-08	4.70E-07
organic acid metabolic process GO:0006082	95	140	1.5	2.60E-07	3.90E-06
regulation of cellular amide metabolic process GO:0034248	33	57	1.7	4.10E-06	5.40E-05
regulation of protein metabolic process GO:0051246	1.10E+02	151	1.4	5.90E-06	7.10E-05
embryo development ending in birth or egg hatching GO:0009792	73	106	1.4	1.60E-05	0.00017
male anatomical structure morphogenesis GO:0090598	25	44	1.7	3.20E-05	0.00032
ribonucleoprotein granule GO:0035770	21	37	1.7	0.00012	0.0011
post-embryonic animal organ development GO:0048569	38	58	1.5	0.00015	0.0013
biosynthetic process GO:0009058	4.70E+02	538	1.1	0.00017	0.0014
response to endogenous stimulus GO:0009719	38	58	1.5	0.00018	0.0014
multicellular organism growth GO:0035264	23	39	1.7	0.00019	0.0014
response to topologically incorrect protein GO:0035966	42	60	1.4	0.001	0.0066
IRE1-mediated unfolded protein response GO:0036498	25	39	1.5	0.0011	0.0071
nucleoside phosphate binding GO:1901265	2.70E+02	310	1.2	0.0021	0.012
vesicle GO:0031982	73	92	1.3	0.0061	0.035
ion homeostasis GO:0050801	30	42	1.4	0.0067	0.036
cellular aromatic compound metabolic process GO:0006725	4.50E+02	493	1.1	0.0076	0.039
macromolecule biosynthetic process GO:0009059	3.60E+02	405	1.1	0.0082	0.04
ribose phosphate metabolic process GO:0019693	35	47	1.4	0.0086	0.041
heterocycle metabolic process GO:0046483	4.50E+02	491	1.1	0.009	0.041
extracellular space GO:0005615	52	67	1.3	0.0094	0.042
ribonucleotide binding GO:0032553	2.40E+02	273	1.1	0.0095	0.042
nucleoside phosphate metabolic process GO:0006753	45	59	1.3	0.0096	0.042
organic cyclic compound metabolic process GO:1901360	4.60E+02	501	1.1	0.01	0.042
reproduction GO:0000003	1.70E+02	191	1.2	0.012	0.046
cation binding GO:0043169	3.60E+02	393	1.1	0.014	0.051
negative regulation of transcription by RNA polymerase II GO:0000122	32	43	1.3	0.014	0.051
purine nucleotide binding GO:0017076	2.40E+02	268	1.1	0.015	0.051
purine nucleotide metabolic process GO:0006163	32	43	1.3	0.016	0.054
reproductive system development GO:0061458	27	36	1.4	0.019	0.062

Tissue enrichment analysis

Term	Expected	Observed	Enrichment	P value	Q value
epithelial system WBbt:0005730	1.00E+03	1353	1.3	2.60E-32	7.70E-30
midbody WBbt:0005740	1.60E+02	264	1.7	8.20E-19	1.20E-16
sex organ WBbt:0008422	3.40E+02	480	1.4	1.90E-17	1.80E-15
PVD WBbt:0006831	5.70E+02	732	1.3	5.50E-14	4.00E-12
touch receptor neuron WBbt:0005237	3.90E+02	523	1.3	5.80E-14	4.00E-12
outer labial sensillum WBbt:0005501	5.80E+02	736	1.3	6.60E-13	3.20E-11
gonadal primordium WBbt:0008366	5.80E+02	733	1.3	1.10E-12	4.60E-11
excretory system WBbt:0005736	1.90E+02	279	1.5	1.40E-12	5.00E-11
hermaphrodite WBbt:0007849	4.50E+02	585	1.3	1.60E-12	5.00E-11
excretory secretory system WBbt:0006850	1.90E+02	279	1.5	1.60E-12	5.00E-11
lateral nerve cord WBbt:0006769	82	139	1.7	1.00E-11	2.70E-10
dorsal nerve cord WBbt:0006750	1.20E+02	181	1.5	8.10E-10	1.90E-08
excretory duct cell WBbt:0004540	9.3	27	2.9	2.20E-09	5.00E-08
excretory cell WBbt:0005812	1.70E+02	242	1.4	2.90E-09	6.00E-08
somatic gonad WBbt:0005785	1.10E+02	163	1.5	2.70E-08	5.20E-07
nerve ring WBbt:0006749	1.80E+02	240	1.4	1.50E-07	2.80E-06
anal depressor muscle WBbt:0004292	94	136	1.5	1.20E-06	2.10E-05
spermatheca WBbt:0005319	1.20E+02	163	1.4	1.30E-06	2.10E-05
P10 WBbt:0006779	7.9	20	2.5	3.90E-06	6.00E-05
P7 WBbt:0006776	6.9	18	2.6	5.90E-06	8.60E-05
P8 WBbt:0006777	6.5	17	2.6	9.50E-06	0.00013
P5 WBbt:0006774	6.5	17	2.6	9.50E-06	0.00013
Epl WBbt:0006000	7.1	18	2.5	1.00E-05	0.00013
P8.p WBbt:0006896	13	27	2.1	1.50E-05	0.00018
Ear WBbt:0006370	6.7	17	2.5	1.60E-05	0.00019
P6 WBbt:0006775	6.7	17	2.5	1.60E-05	0.00019
Epr WBbt:0006547	7.3	18	2.5	1.70E-05	0.00019
P7.p WBbt:0006895	14	29	2	2.30E-05	0.00024
Eplp WBbt:0006496	9.3	21	2.3	2.50E-05	0.00025
Epla WBbt:0006661	8.7	20	2.3	2.60E-05	0.00025
P9 WBbt:0006778	6.9	17	2.5	2.80E-05	0.00026
P3.p WBbt:0006891	13	26	2.1	3.30E-05	0.0003
P4.p WBbt:0006892	13	26	2.1	3.30E-05	0.0003
Epra WBbt:0006321	9.5	21	2.2	3.80E-05	0.00032
Ealp WBbt:0006546	8.3	19	2.3	4.30E-05	0.00035
tail WBbt:0005741	3.90E+02	460	1.2	4.50E-05	0.00036
P1 WBbt:0006770	6.5	16	2.5	4.50E-05	0.00036
P5.p WBbt:0006893	14	28	2	4.80E-05	0.00037
P6.p WBbt:0006894	14	28	2	4.80E-05	0.00037
Eala WBbt:0006104	7.9	18	2.3	6.90E-05	0.0005
P2 WBbt:0006771	6.7	16	2.4	7.30E-05	0.00052
Eprp WBbt:0006507	9.9	21	2.1	8.10E-05	0.00056
P11 WBbt:0004410	11	23	2	0.00013	0.00086
P12 WBbt:0004409	9.5	20	2.1	0.00013	0.00086
hermaphrodite distal tip cell WBbt:0006863	38	59	1.5	0.00013	0.00086
gonad arm WBbt:0008629	47	69	1.5	0.00016	0.001
hyp6 WBbt:0004679	12	23	2	0.00017	0.0011
Eal WBbt:0006441	7.1	16	2.3	0.00018	0.0011
Earp WBbt:0006646	8.5	18	2.1	0.00023	0.0013
Caaaa WBbt:0005899	7.3	16	2.2	0.00027	0.0016

Term	Expected	Observed	Enrichment	P value	Q value
spermathecal-uterine junction WBbt:0006756	12	23	1.9	0.00042	0.0024
P4 WBbt:0006773	6.9	15	2.2	0.00044	0.0025
vulB1 WBbt:0006763	12	22	1.9	0.00049	0.0027
vulC WBbt:0006765	13	23	1.8	0.00073	0.0039
vulA WBbt:0006762	11	21	1.9	0.00076	0.004
Cpapp WBbt:0005897	9.3	18	1.9	0.00088	0.0046
vulB2 WBbt:0006764	11	21	1.8	0.001	0.0051
thermosensory neuron WBbt:0005838	4.50E+02	505	1.1	0.001	0.0051
Caaap WBbt:0006267	8.7	17	2	0.001	0.0051
ABplappp WBbt:0006470	7.5	15	2	0.0013	0.0064
Cpapa WBbt:0005962	8.9	17	1.9	0.0014	0.0066
anchor cell WBbt:0004522	15	26	1.7	0.0015	0.0068
vulF WBbt:0006768	13	23	1.7	0.0019	0.0089
ABarpaapa WBbt:0005844	7.1	14	2	0.0021	0.0095
Eara WBbt:0006161	8.5	16	1.9	0.0022	0.0096
ABarppaa WBbt:0006465	6.5	13	2	0.0024	0.011
Cpaaa WBbt:0006212	7.9	15	1.9	0.0025	0.011
ABalaapa WBbt:0006130	6.7	13	1.9	0.0033	0.014
anal region WBbt:0006919	8.1	15	1.9	0.0034	0.014
ABarpaaa WBbt:0006398	8.9	16	1.8	0.0038	0.016
ABplppaa WBbt:0006170	7.5	14	1.9	0.0039	0.016
ABplppap WBbt:0006028	7.5	14	1.9	0.0039	0.016
amphid socket cell WBbt:0008379	8.3	15	1.8	0.0044	0.018
ABalppaa WBbt:0005944	8.3	15	1.8	0.0044	0.018
ABplapapp WBbt:0006413	6.9	13	1.9	0.0046	0.018
Psub1 WBbt:0006874	47	63	1.3	0.0049	0.019
Capp WBbt:0006098	7.7	14	1.8	0.0052	0.02
Cppp WBbt:0006268	7.7	14	1.8	0.0052	0.02
ABpraappp WBbt:0006270	7.7	14	1.8	0.0052	0.02
vulD WBbt:0006766	14	23	1.6	0.0055	0.02
ABprappp WBbt:0006702	8.5	15	1.8	0.0058	0.021
ABplappap WBbt:0006067	7.1	13	1.8	0.0061	0.022
ABprappap WBbt:0006220	7.1	13	1.8	0.0061	0.022
Cppa WBbt:0006168	9.3	16	1.7	0.0063	0.022
Cpaap WBbt:0006594	7.9	14	1.8	0.0069	0.023
hyp4 WBbt:0004687	11	18	1.7	0.0071	0.024
ABplappaa WBbt:0006371	6.5	12	1.8	0.0072	0.024
MSpppp WBbt:0006409	6.5	12	1.8	0.0072	0.024
ABprppaa WBbt:0005984	7.3	13	1.8	0.0081	0.026
ABalaaap WBbt:0005982	7.3	13	1.8	0.0081	0.026
ABaraaaa WBbt:0006360	7.3	13	1.8	0.0081	0.026
vulE WBbt:0006767	13	20	1.6	0.0092	0.029
ABarpaapp WBbt:0006620	6.7	12	1.8	0.0096	0.03
ABalappa WBbt:0006157	6.7	12	1.8	0.0096	0.03
ABprapap WBbt:0006678	6.7	12	1.8	0.0096	0.03
ABarpppap WBbt:0006251	6.7	12	1.8	0.0096	0.03
MSppaa WBbt:0006531	6.7	12	1.8	0.0096	0.03
ABarpapa WBbt:0006603	6.7	12	1.8	0.0096	0.03

Term	Expected	Observed	Enrichment	P value	Q value
ABprpppaa WBbt:0006552	7.5	13	1.7	0.011	0.031
ABprappa WBbt:0006269	7.5	13	1.7	0.011	0.031
ABpraappa WBbt:0006035	8.3	14	1.7	0.011	0.033
ABalppap WBbt:0006112	6.9	12	1.7	0.013	0.036
ABprappaa WBbt:0006350	6.9	12	1.7	0.013	0.036
ABalpppa WBbt:0006649	7.7	13	1.7	0.014	0.038
ABalaaaa WBbt:0006427	7.7	13	1.7	0.014	0.038
intestine WBbt:0005772	1.40E+03	1468	1	0.016	0.044
ABpraapaa WBbt:0006108	7.1	12	1.7	0.016	0.044
ABalaapp WBbt:0006553	6.5	11	1.7	0.019	0.052
ABpraapp WBbt:0006335	6.5	11	1.7	0.019	0.052
ABarpppaa WBbt:0006041	6.5	11	1.7	0.019	0.052
P7.pp WBbt:0006984	6.5	11	1.7	0.019	0.052
ABplpappa WBbt:0006232	7.3	12	1.6	0.021	0.053
Capa WBbt:0006444	8.1	13	1.6	0.021	0.055
AB WBbt:0004015	23	32	1.4	0.022	0.055
P7.pa WBbt:0006983	6.7	11	1.6	0.025	0.062
ABplaaapp WBbt:0006136	6.7	11	1.6	0.025	0.062
ABarappp WBbt:0006524	6.7	11	1.6	0.025	0.062
ABaraaap WBbt:0005861	6.7	11	1.6	0.025	0.062
reproductive tract WBbt:0005744	3.70E+02	403	1.1	0.025	0.062
ABprppap WBbt:0006346	7.5	12	1.6	0.026	0.062
Caapa WBbt:0006123	7.5	12	1.6	0.026	0.062
ABplaaapa WBbt:0006680	8.3	13	1.6	0.026	0.063
excretory gland cell WBbt:0005776	9.1	14	1.5	0.027	0.063
uterine muscle WBbt:0005342	27	35	1.3	0.03	0.071
MSaaap WBbt:0006160	6.9	11	1.6	0.031	0.072
ABarppapp WBbt:0006240	6.9	11	1.6	0.031	0.072
ABpraaaa WBbt:0006442	7.7	12	1.6	0.032	0.073
ABplaappp WBbt:0005948	7.1	11	1.6	0.038	0.087
ABprpapaa WBbt:0006446	7.1	11	1.6	0.038	0.087
MSaapp WBbt:0006425	7.1	11	1.6	0.038	0.087
ABprpppap WBbt:0006237	7.1	11	1.6	0.038	0.087
ABplpppap WBbt:0006665	7.1	11	1.6	0.038	0.087
hyp5 WBbt:0004685	10	15	1.5	0.039	0.087
ABalpapa WBbt:0006573	7.9	12	1.5	0.039	0.087
MSapa WBbt:0005898	6.5	10	1.5	0.046	0.099
ABplapppp WBbt:0006656	6.5	10	1.5	0.046	0.099
ABpraaap WBbt:0006534	6.5	10	1.5	0.046	0.099
ABarappa WBbt:0006005	6.5	10	1.5	0.046	0.099
ABprpppa WBbt:0005943	6.5	10	1.5	0.046	0.099
MSpaap WBbt:0005878	6.5	10	1.5	0.046	0.099
ABaraapp WBbt:0006153	7.3	11	1.5	0.047	0.099
Caap WBbt:0005921	7.3	11	1.5	0.047	0.099
ABalpaap WBbt:0005934	7.3	11	1.5	0.047	0.099
ABarapaa WBbt:0006515	7.3	11	1.5	0.047	0.099
ABalpaaa WBbt:0006557	8.1	12	1.5	0.047	0.099

B.16 *dpy-7syn1* L4 gene-set GO-term and tissue-enrichment analysis complete results

GO-terms

Term	Expected	Observed	Enrichment	P value	Q value
structural constituent of cuticle GO:0042302	35	100	2.8	2.30E-28	2.80E-26
molting cycle GO:0042303	22	69	3.1	2.40E-23	1.50E-21
peptide biosynthetic process GO:0043043	74	146	2	9.60E-19	3.80E-17
structural constituent of ribosome GO:0003735	30	71	2.3	1.50E-14	4.40E-13
post-embryonic development GO:0009791	1.10E+02	173	1.6	9.90E-13	2.40E-11
aging GO:0007568	53	98	1.9	2.70E-11	5.30E-10
organic acid metabolic process GO:0006082	95	145	1.5	1.90E-08	3.20E-07
ribonucleoprotein complex subunit organization GO:0071826	22	46	2.1	3.20E-08	4.70E-07
embryo development ending in birth or egg hatching GO:0009792	74	116	1.6	5.20E-08	6.90E-07
multicellular organism growth GO:0035264	23	45	1.9	9.70E-07	1.20E-05
regulation of cellular amide metabolic process GO:0034248	33	57	1.7	4.70E-06	5.00E-05
post-embryonic animal organ development GO:0048569	38	62	1.6	1.10E-05	0.00011
regulation of protein metabolic process GO:0051246	1.10E+02	150	1.4	1.20E-05	0.00011
nucleoside phosphate metabolic process GO:0006753	45	69	1.5	5.70E-05	0.00049
ribose phosphate metabolic process GO:0019693	35	55	1.6	8.70E-05	0.00069
purine nucleotide metabolic process GO:0006163	32	52	1.6	8.80E-05	0.00069
nucleoside phosphate binding GO:1901265	2.70E+02	323	1.2	0.00012	0.00087
male anatomical structure morphogenesis GO:0090598	26	42	1.6	0.00017	0.0011
biosynthetic process GO:0009058	4.70E+02	536	1.1	0.00036	0.0023
response to endogenous stimulus GO:0009719	38	56	1.5	0.00067	0.004
response to topologically incorrect protein GO:0035966	42	60	1.4	0.0011	0.0062
modification-dependent macromolecule catabolic process GO:0043632	51	71	1.4	0.0012	0.0063
ribonucleotide binding GO:0032553	2.40E+02	282	1.2	0.0018	0.0092
vesicle GO:0031982	73	95	1.3	0.0023	0.011
purine nucleotide binding GO:0017076	2.40E+02	278	1.2	0.0024	0.011
IRE1-mediated unfolded protein response GO:0036498	25	36	1.4	0.0078	0.036
ribonucleoprotein granule GO:0035770	21	31	1.4	0.0097	0.043
response to nitrogen compound GO:1901698	27	37	1.4	0.01	0.044
nucleoside binding GO:0001882	47	60	1.3	0.014	0.056
extracellular space GO:0005615	52	66	1.3	0.015	0.06
regulation of neuron differentiation GO:0045664	26	36	1.4	0.015	0.06
cation binding GO:0043169	3.60E+02	393	1.1	0.017	0.063
nuclear outer membrane-endoplasmic reticulum membrane network GO:0042175	54	68	1.3	0.017	0.063
amide transport GO:0042886	98	116	1.2	0.019	0.066
ion homeostasis GO:0050801	30	40	1.3	0.019	0.066
reproduction GO:0000003	1.70E+02	189	1.1	0.021	0.069
identical protein binding GO:0042802	26	35	1.3	0.022	0.072
cellular aromatic compound metabolic process GO:0006725	4.50E+02	485	1.1	0.028	0.088
protein catabolic process GO:0030163	75	89	1.2	0.03	0.09
heterocycle metabolic process GO:0046483	4.50E+02	483	1.1	0.032	0.096
envelope GO:0031975	72	86	1.2	0.034	0.098

Tissue enrichment analysis

Term	Expected	Observed	Enrichment	P value	Q value
epithelial system WBbt:0005730	1.00E+03	1317	1.3	3.60E-28	1.00E-25
midbody WBbt:0005740	1.60E+02	267	1.7	2.40E-20	3.50E-18
sex organ WBbt:0008422	3.30E+02	488	1.5	5.10E-20	5.00E-18
PVD WBbt:0006831	5.60E+02	731	1.3	9.00E-15	6.50E-13
hermaphrodite WBbt:0007849	4.50E+02	590	1.3	4.00E-14	2.30E-12
outer labial sensillum WBbt:0005501	5.70E+02	738	1.3	4.20E-14	2.30E-12
touch receptor neuron WBbt:0005237	3.80E+02	517	1.3	1.10E-13	4.70E-12
lateral nerve cord WBbt:0006769	81	142	1.8	4.80E-13	1.70E-11
gonadal primordium WBbt:0008366	5.70E+02	712	1.2	9.80E-11	3.10E-09
dorsal nerve cord WBbt:0006750	1.20E+02	180	1.5	6.40E-10	1.90E-08
somatic gonad WBbt:0005785	1.10E+02	164	1.5	7.60E-09	2.00E-07
nerve ring WBbt:0006749	1.80E+02	240	1.4	6.70E-08	1.60E-06
excretory system WBbt:0005736	1.90E+02	251	1.3	2.20E-07	5.00E-06
excretory secretory system WBbt:0006850	1.90E+02	251	1.3	2.40E-07	5.10E-06
anal depressor muscle WBbt:0004292	93	137	1.5	4.10E-07	8.00E-06
Ealp WBbt:0006546	8.2	22	2.7	4.20E-07	8.00E-06
Eala WBbt:0006104	7.8	21	2.7	6.80E-07	1.20E-05
spermatheca WBbt:0005319	1.20E+02	162	1.4	1.10E-06	1.80E-05
Epla WBbt:0006661	8.6	21	2.4	5.50E-06	8.50E-05
Epra WBbt:0006321	9.4	22	2.3	8.50E-06	0.00012
Earp WBbt:0006646	8.4	20	2.4	1.40E-05	0.0002
excretory duct cell WBbt:0004540	9.2	21	2.3	2.10E-05	0.00028
Eplp WBbt:0006496	9.2	21	2.3	2.10E-05	0.00028
P11 WBbt:0004410	11	24	2.1	3.50E-05	0.00042
P5 WBbt:0006774	6.4	16	2.5	3.90E-05	0.00045
hyp6 WBbt:0004679	12	24	2.1	4.90E-05	0.00055
tail WBbt:0005741	3.90E+02	455	1.2	5.40E-05	0.00058
Eara WBbt:0006161	8.4	19	2.3	5.60E-05	0.00058
P10 WBbt:0006779	7.8	18	2.3	6.00E-05	0.0006
P6 WBbt:0006775	6.6	16	2.4	6.40E-05	0.00062
Eprp WBbt:0006507	9.8	21	2.2	6.90E-05	0.00065
P4.p WBbt:0006892	12	25	2	8.20E-05	0.00075
P3.p WBbt:0006891	12	25	2	8.20E-05	0.00075
excretory cell WBbt:0005812	1.70E+02	214	1.3	9.20E-05	0.00078
P7 WBbt:0006776	6.8	16	2.3	0.0001	0.00085
P6.p WBbt:0006894	14	27	1.9	0.00011	0.0009
P5.p WBbt:0006893	14	27	1.9	0.00011	0.0009
P8.p WBbt:0006896	13	25	2	0.00011	0.0009
P7.p WBbt:0006895	14	27	1.9	0.00015	0.0011
P1 WBbt:0006770	6.4	15	2.3	0.00017	0.0012
P8 WBbt:0006777	6.4	15	2.3	0.00017	0.0012
vulA WBbt:0006762	11	22	2	0.00023	0.0016
P2 WBbt:0006771	6.6	15	2.3	0.00026	0.0017
ABarpppap WBbt:0006251	6.6	15	2.3	0.00026	0.0017
somatic cell WBbt:0008378	12	23	1.9	0.00027	0.0018
gonad arm WBbt:0008629	46	67	1.5	0.00037	0.0023
ABarppapp WBbt:0006240	6.8	15	2.2	0.00039	0.0024
P4 WBbt:0006773	6.8	15	2.2	0.00039	0.0024
P9 WBbt:0006778	6.8	15	2.2	0.00039	0.0024
ABpraappp WBbt:0006270	7.6	16	2.1	0.00052	0.003
ABarpaapa WBbt:0005844	7	15	2.1	0.00058	0.0033
MSpppp WBbt:0006409	6.4	14	2.2	0.00064	0.0036
Cpaaa WBbt:0006212	7.8	16	2.1	0.00074	0.004
Cpaap WBbt:0006594	7.8	16	2.1	0.00074	0.004
Epr WBbt:0006547	7.2	15	2.1	0.00083	0.0044
vulB2 WBbt:0006764	11	21	1.9	0.00087	0.0045
Psub1 WBbt:0006874	47	66	1.4	0.00091	0.0046
ABarpaapp WBbt:0006620	6.6	14	2.1	0.00093	0.0047
Ear WBbt:0006370	6.6	14	2.1	0.00093	0.0047

Appendix

Term	Expected	Observed	Enrichment	P value	Q value
spermathecal-uterine junction WBbt:0006756	12	22	1.8	0.00096	0.0047
hyp4 WBbt:0004687	11	20	1.9	0.001	0.0048
P12 WBbt:0004409	9.4	18	1.9	0.001	0.0049
hermaphrodite distal tip cell WBbt:0006863	38	55	1.5	0.0011	0.005
vulB1 WBbt:0006763	12	21	1.8	0.0011	0.0051
amphid socket cell WBbt:0008379	8.2	16	2	0.0014	0.0063
ABpraappa WBbt:0006035	8.2	16	2	0.0014	0.0063
vulC WBbt:0006765	12	22	1.8	0.0016	0.0068
Epl WBbt:0006000	7	14	2	0.0019	0.0081
ABpraapaa WBbt:0006108	7	14	2	0.0019	0.0081
ABplappaa WBbt:0006371	6.4	13	2	0.0022	0.009
ABarpppaa WBbt:0006041	6.4	13	2	0.0022	0.009
Caaap WBbt:0006267	8.6	16	1.9	0.0026	0.01
Caaaa WBbt:0005899	7.2	14	1.9	0.0026	0.01
vulF WBbt:0006768	13	22	1.7	0.0038	0.015
ABprappaa WBbt:0006350	6.8	13	1.9	0.0042	0.016
vulD WBbt:0006766	14	23	1.6	0.0048	0.018
ABplappap WBbt:0006067	7	13	1.9	0.0056	0.021
ABprappap WBbt:0006220	7	13	1.9	0.0056	0.021
Eal WBbt:0006441	7	13	1.9	0.0056	0.021
Cppa WBbt:0006168	9.2	16	1.7	0.0057	0.021
ABprapaap WBbt:0006624	6.4	12	1.9	0.0066	0.024
ABalaaap WBbt:0005982	7.2	13	1.8	0.0074	0.026
ABplpappa WBbt:0006232	7.2	13	1.8	0.0074	0.026
vulE WBbt:0006767	12	20	1.6	0.0082	0.028
Cpapa WBbt:0005962	8.8	15	1.7	0.0086	0.029
ABalaapa WBbt:0006130	6.6	12	1.8	0.0088	0.03
ABpraapap WBbt:0006062	6.6	12	1.8	0.0088	0.03
excretory gland cell WBbt:0005776	9	15	1.7	0.011	0.036
MSaaap WBbt:0006160	6.8	12	1.8	0.012	0.038
ABprapapp WBbt:0006290	6.8	12	1.8	0.012	0.038
ABalaaaa WBbt:0006427	7.6	13	1.7	0.012	0.04
Cppp WBbt:0006268	7.6	13	1.7	0.012	0.04
ABprappp WBbt:0006702	8.4	14	1.7	0.013	0.041
Cpapp WBbt:0005897	9.2	15	1.6	0.014	0.042
MSaapp WBbt:0006425	7	12	1.7	0.015	0.045
ABprpapaa WBbt:0006446	7	12	1.7	0.015	0.045
ABprapapa WBbt:0006510	7	12	1.7	0.015	0.045
ABarppaa WBbt:0006465	6.4	11	1.7	0.018	0.053
ABplpaapa WBbt:0006115	6.4	11	1.7	0.018	0.053
MSpaap WBbt:0005878	6.4	11	1.7	0.018	0.053
thermosensory neuron WBbt:0005838	4.40E+02	481	1.1	0.019	0.053
ABaraaaa WBbt:0006360	7.2	12	1.7	0.019	0.054
ABprppaa WBbt:0005984	7.2	12	1.7	0.019	0.054
Capa WBbt:0006444	8	13	1.6	0.02	0.055
anchor cell WBbt:0004522	15	22	1.4	0.022	0.062
ABalappa WBbt:0006157	6.6	11	1.7	0.023	0.062
ABprpapa WBbt:0006259	6.6	11	1.7	0.023	0.062
MSppaa WBbt:0006531	6.6	11	1.7	0.023	0.062
ABplaaapp WBbt:0006136	6.6	11	1.7	0.023	0.062
ABplaaapa WBbt:0006680	8.2	13	1.6	0.024	0.064
ABalppaa WBbt:0005944	8.2	13	1.6	0.024	0.064
ABplapapp WBbt:0006413	6.8	11	1.6	0.029	0.074
Capp WBbt:0006098	7.6	12	1.6	0.029	0.076
ABplpappp WBbt:0006390	7.6	12	1.6	0.029	0.076
hyp5 WBbt:0004685	10	15	1.5	0.035	0.089
ABplppppp WBbt:0006574	7	11	1.6	0.036	0.089
ABalpapa WBbt:0006573	7.8	12	1.5	0.036	0.089
uterine seam cell WBbt:0006789	8.6	13	1.5	0.036	0.089

B.17 List of transcription factors expressed in the seam cells but not the hypodermis

WormBase Gene ID	Public Name	Sequence Name	WormBase Gene ID	Public Name	Sequence Name
WBGene00019218	madf-3	H20J04.3	WBGene00001085	dpy-26	C25G4.5
WBGene00009174	F26H9.2	F26H9.2	WBGene00001820	ham-1	F53B2.6
WBGene00016162	crh-2	C27D6.4	WBGene00001186	egl-18	F55A8.1
WBGene00001821	ham-2	C07A12.1	WBGene00022795	ZK686.5	ZK686.5
WBGene00008195	ceh-88	C49C3.5	WBGene00016620	dhhc-9	C43H6.7
WBGene00022608	madf-9	ZC416.1	WBGene00004024	php-3	Y75B8A.1
WBGene00020708	dmd-8	T22H9.4	WBGene00017535	atf-8	F17A9.3
WBGene00003663	nhr-73	C27C7.4	WBGene00003703	nhr-113	ZK1025.9
WBGene00001210	egl-46	K11G9.4	WBGene00006970	zag-1	F28F9.1
WBGene00009899	efl-3	F49E12.6	WBGene00000439	ceh-16	C13G5.1
WBGene00016927	nhr-172	C54F6.9	WBGene00020062	nhr-270	R13D11.8
WBGene00006547	tbx-11	F40H6.4	WBGene00014189	nhr-245	ZK1025.10
WBGene00009608	nhr-265	F41D3.3	WBGene00005011	F26F4.8	F26F4.8
WBGene00018189	nhr-181	F38H12.3	WBGene00003719	nhr-129	C50B6.14
WBGene00021931	Y55F3AM.14	Y55F3AM.14	WBGene00007105	znf-207	B0035.1
WBGene00006492	let-391	C27A12.3	WBGene00003037	lin-54	JC8.6
WBGene00020093	R144.3	R144.3	WBGene00021816	Y53G8AR.9	Y53G8AR.9
WBGene00001061	dpl-1	T23G7.1	WBGene00003717	nhr-127	T13F3.3
WBGene00010770	K11D2.4	K11D2.4	WBGene00001249	elt-1	W09C2.1
WBGene00017687	ets-4	F22A3.1	WBGene00001438	fkh-6	B0286.5
WBGene00003592	nfi-1	ZK1290.4	WBGene00012277	ccch-3	W05B10.2
WBGene00016888	znf-598	C52E12.1	WBGene00004857	sma-3	R13F6.9
WBGene00017326	dmd-5	F10C1.5	WBGene00003626	nhr-32	K08H2.8
WBGene00001253	elt-6	F52C12.5	WBGene00003606	nhr-7	F54D1.4
WBGene00008417	D2030.7	D2030.7	WBGene00006554	tbx-35	ZK177.10
WBGene00000222	atf-6	F45E6.2	WBGene00003649	nhr-59	T27B7.1
WBGene00015397	nhr-149	C03G6.12	WBGene00012988	ztf-22	Y48C3A.4
WBGene00011661	ztf-27	T09F3.1	WBGene00003664	nhr-74	C27C7.3
WBGene00011206	R10E4.11	R10E4.11	WBGene00000895	dac-1	B0412.1

B.18 List of chromatin factors expressed in the seam cells but not the hypodermis

WormBase Gene ID	Public Name	Sequence Name
WBGene00014240	htas-1	ZK1251.1
WBGene00011636	cec-3	T09A5.8
WBGene00015501	C06A5.3	C06A5.3
WBGene00007433	swsn-7	C08B11.3
WBGene00001835	hda-2	C08B11.2
WBGene00004806	skp-1	T27F2.1
WBGene00001977	hmg-12	Y17G7A.1
WBGene00009180	nurf-1	F26H11.2
WBGene00001946	his-72	Y49E10.6
WBGene00001831	hcp-3	F58A4.3
WBGene00015938	anat-1	C17H12.13
WBGene00009025	phf-34	F21G4.4
WBGene00008547	F07A11.4	F07A11.4
WBGene00007256	swsn-9	C01H6.7
WBGene00010036	cpar-1	F54C8.2
WBGene00001974	hmg-4	T20B12.8
WBGene00006391	taf-9	T12D8.7
WBGene00009672	F43G9.12	F43G9.12
WBGene00001877	his-3	T10C6.12
WBGene00008206	set-6	C49F5.2
WBGene00001976	hmg-11	T05A7.4
WBGene00016061	hpo-15	C24G6.6
WBGene00007953	hda-11	C35A5.9
WBGene00001834	hda-1	C53A5.3
WBGene00001916	his-42	F08G2.3
WBGene00017423	F13C5.2	F13C5.2
WBGene00000482	chd-3	T14G8.1
WBGene00017757	bra-2	F23H11.1
WBGene00002169	isw-1	F37A4.8
WBGene00000262	bra-1	F54B11.6
WBGene00010369	chd-1	H06O01.2
WBGene00017993	cec-5	F32E10.6
WBGene00001971	hmg-1.1	Y48B6A.14
WBGene00001470	baz-2	ZK783.4
WBGene00000275	bub-1	R06C7.8

B.19 GO-term and tissue enrichment analysis results on genes associated only with *srf-3i1* CATaDa sites at the L4 stage

GO terms

Term	Expected	Observed	Enrichment	P value	Q value
supramolecular polymer GO:0099081	14	35	2.5	2.90E-07	3.50E-05
cell surface GO:0009986	4.7	15	3.2	1.70E-05	0.001
embryo development ending in birth or egg hatching GO:0009792	16	33	2.1	3.20E-05	0.0013
actin binding GO:0003779	5.3	14	2.7	0.00026	0.0077
post-embryonic development GO:0009791	23	38	1.7	0.00091	0.022
reproduction GO:0000003	36	54	1.5	0.0013	0.026
regulation of neuron differentiation GO:0045664	5.7	12	2.1	0.0047	0.08

Tissue enrichment analysis

Term	Expected	Observed	Enrichment	P value	Q value
striated muscle WBbt:0005779	68	115	1.7	2.00E-08	5.90E-06
sex organ WBbt:0008422	80	123	1.5	1.10E-06	0.00015
hermaphrodite WBbt:0007849	1.10E+02	142	1.3	0.00025	0.024
ABprpapppp WBbt:0005847	1.7	7	4.1	0.00025	0.024
ABprpappa WBbt:0006088	1.8	7	4	0.0003	0.024
ABpraapppp WBbt:0006270	1.8	7	3.9	0.00037	0.024
ABpraappa WBbt:0006035	2	7	3.6	0.00062	0.026
ABplpaapa WBbt:0006115	1.5	6	3.9	0.00069	0.026
ABprpaapa WBbt:0006047	1.5	6	3.9	0.00069	0.026
ABpraapap WBbt:0006062	1.6	6	3.8	0.00083	0.026
anal depressor muscle WBbt:0004292	22	37	1.7	0.001	0.027
ABpraapaa WBbt:0006108	1.7	6	3.6	0.0012	0.029
ABarpaapa WBbt:0005844	1.7	6	3.6	0.0012	0.029
ABplpappa WBbt:0006232	1.7	6	3.5	0.0014	0.029
ABplpappp WBbt:0006390	1.8	6	3.3	0.0019	0.038
anal sphincter muscle WBbt:0005798	6.1	13	2.1	0.0035	0.063
ABplpppaa WBbt:0006222	1.5	5	3.3	0.0038	0.065
ABprapaap WBbt:0006624	1.5	5	3.3	0.0038	0.065
uterine muscle WBbt:0005342	6.3	13	2.1	0.0043	0.065
ABarpaapp WBbt:0006620	1.6	5	3.2	0.0044	0.065
ABplppppa WBbt:0006352	1.6	5	3.2	0.0044	0.065
MSappa WBbt:0006717	1.6	5	3.2	0.0044	0.065
ABarpppap WBbt:0006251	1.6	5	3.2	0.0044	0.065
MSappp WBbt:0006125	1.6	5	3.2	0.0044	0.065
ABplaaaap WBbt:0006625	1.6	5	3.2	0.0044	0.065
ABarppapp WBbt:0006240	1.6	5	3.1	0.0051	0.065
ABprppppp WBbt:0005983	1.6	5	3.1	0.0051	0.065
ABprapapp WBbt:0006290	1.6	5	3.1	0.0051	0.065
ABprappaa WBbt:0006350	1.6	5	3.1	0.0051	0.065
ABplppppp WBbt:0006574	1.7	5	3	0.0059	0.065
SIA WBbt:0005361	1.7	5	3	0.0059	0.065
ABprapapa WBbt:0006510	1.7	5	3	0.0059	0.065
ABprpapaa WBbt:0006446	1.7	5	3	0.0059	0.065
MSaapp WBbt:0006425	1.7	5	3	0.0059	0.065
ABprpppap WBbt:0006237	1.7	5	3	0.0059	0.065

Term	Expected	Observed	Enrichment	P value	Q value
ABprappap WBbt:0006220	1.7	5	3	0.0059	0.065
ABplaapap WBbt:0005887	1.7	5	3	0.0059	0.065
ABplaapaa WBbt:0005866	1.7	5	3	0.0059	0.065
ABplpppap WBbt:0006665	1.7	5	3	0.0059	0.065
Caaaa WBbt:0005899	1.7	5	2.9	0.0068	0.065
ABprpppaa WBbt:0006552	1.8	5	2.8	0.0078	0.065
Caapa WBbt:0006123	1.8	5	2.8	0.0078	0.065
ABplaappa WBbt:0006519	1.8	5	2.8	0.0088	0.065
linker cell WBbt:0005062	1.9	5	2.7	0.01	0.066
Cpaap WBbt:0006594	1.9	5	2.7	0.01	0.066
Cpaaa WBbt:0006212	1.9	5	2.7	0.01	0.066
head muscle WBbt:0006761	9.2	16	1.7	0.011	0.068
Z2 WBbt:0004576	2.5	6	2.4	0.011	0.068
Z3 WBbt:0004575	2.5	6	2.4	0.011	0.068
Capa WBbt:0006444	1.9	5	2.6	0.011	0.068
midbody WBbt:0005740	38	51	1.4	0.012	0.071
ABplaaapa WBbt:0006680	2	5	2.6	0.013	0.071
gonad arm WBbt:0008629	11	18	1.6	0.015	0.082
Caaap WBbt:0006267	2	5	2.4	0.016	0.084
Epla WBbt:0006661	2	5	2.4	0.016	0.084
vulA WBbt:0006762	2.7	6	2.3	0.016	0.084
Psub1 WBbt:0006874	11	18	1.6	0.017	0.086
Cpapa WBbt:0005962	2.1	5	2.4	0.017	0.087
ABarpppaa WBbt:0006041	1.5	4	2.6	0.017	0.087
ABplappaa WBbt:0006371	1.5	4	2.6	0.017	0.087
ABplapppp WBbt:0006656	1.5	4	2.6	0.017	0.087
P11 WBbt:0004410	2.7	6	2.2	0.018	0.087
ABplaaapp WBbt:0006136	1.6	4	2.5	0.02	0.09
Cpapp WBbt:0005897	2.2	5	2.3	0.021	0.096
Cppa WBbt:0006168	2.2	5	2.3	0.021	0.096
Eplp WBbt:0006496	2.2	5	2.3	0.021	0.096
ABplapapp WBbt:0006413	1.6	4	2.5	0.022	0.096
Epra WBbt:0006321	2.2	5	2.2	0.023	0.099

B.20 GO-term and tissue enrichment analysis results on genes associated only with *dpy-7syn1* CATaDa sites at the L4 stage

GO-terms

| Term	Expected	Observed	Enrichment	P value	Q value
molting cycle GO:0042303	6.4	24	3.7	3.40E-09	4.00E-07
structural constituent of cuticle GO:0042302	10	27	2.7	9.30E-07	5.60E-05
organic acid metabolic process GO:0006082	27	50	1.8	1.70E-05	0.00068
response to endogenous stimulus GO:0009719	11	23	2.1	0.00026	0.0077
nuclear outer membrane-endoplasmic reticulum membrane network GO:0042175	16	26	1.7	0.0041	0.098

Tissue enrichment analysis

Term	Expected	Observed	Enrichment	P value	Q value
epithelial system WBbt:0005730	3.20E+02	428	1.3	9.20E-10	2.70E-07
touch receptor neuron WBbt:0005237	1.20E+02	169	1.4	1.90E-05	0.0028
hyp6 WBbt:0004679	3.7	11	3	0.00025	0.024
excretory duct cell WBbt:0004540	2.9	9	3.1	0.00056	0.04
ABarppaa WBbt:0006465	2.1	7	3.4	0.00079	0.046
excretory system WBbt:0005736	60	84	1.4	0.00097	0.047
excretory secretory system WBbt:0006850	60	84	1.4	0.001	0.047
PVD WBbt:0006831	1.80E+02	220	1.2	0.0015	0.054
midbody WBbt:0005740	51	71	1.4	0.0018	0.058
P3.p WBbt:0006891	4	10	2.5	0.0019	0.058
P4.p WBbt:0006892	4	10	2.5	0.0019	0.058
outer labial sensillum WBbt:0005501	1.80E+02	222	1.2	0.0022	0.058
P8.p WBbt:0006896	4.1	10	2.5	0.0022	0.058
Eala WBbt:0006104	2.5	7	2.8	0.003	0.061
Ealp WBbt:0006546	2.6	7	2.7	0.0041	0.079
amphid socket cell WBbt:0008379	2.6	7	2.7	0.0041	0.079
ABalaapa WBbt:0006130	2.1	6	2.8	0.0045	0.079
ABalappa WBbt:0006157	2.1	6	2.8	0.0045	0.079
ABprappp WBbt:0006702	2.7	7	2.6	0.0047	0.079
Earp WBbt:0006646	2.7	7	2.6	0.0047	0.079
Eara WBbt:0006161	2.7	7	2.6	0.0047	0.079
sex organ WBbt:0008422	1.10E+02	133	1.2	0.0048	0.079
P5.p WBbt:0006893	4.5	10	2.2	0.005	0.079
P6.p WBbt:0006894	4.5	10	2.2	0.005	0.079
Epla WBbt:0006661	2.8	7	2.5	0.0055	0.079
P7.p WBbt:0006895	4.6	10	2.2	0.0056	0.079
ABplaappp WBbt:0005948	2.3	6	2.7	0.0063	0.079
hyp4 WBbt:0004687	3.4	8	2.3	0.0068	0.079
pm6 WBbt:0003724	2.9	7	2.4	0.0072	0.079
ABprppaa WBbt:0005984	2.3	6	2.6	0.0073	0.079
Epr WBbt:0006547	2.3	6	2.6	0.0073	0.079
ABalaaap WBbt:0005982	2.3	6	2.6	0.0073	0.079
Eplp WBbt:0006496	2.9	7	2.4	0.0082	0.079
ABplppaa WBbt:0006170	2.4	6	2.5	0.0085	0.079
vulA WBbt:0006762	3.6	8	2.2	0.0086	0.079
Epra WBbt:0006321	3	7	2.3	0.0093	0.079
vulB2 WBbt:0006764	3.6	8	2.2	0.0096	0.079
ABplaappa WBbt:0006519	2.4	6	2.5	0.0098	0.079
ABalaaaa WBbt:0006427	2.4	6	2.5	0.0098	0.079
uterine muscle WBbt:0005342	8.4	15	1.8	0.011	0.079
vulB1 WBbt:0006763	3.7	8	2.2	0.011	0.079
ABalpapa WBbt:0006573	2.5	6	2.4	0.011	0.079
Cpaaa WBbt:0006212	2.5	6	2.4	0.011	0.079
pm7 WBbt:0003721	2.5	6	2.4	0.011	0.079
Cpaap WBbt:0006594	2.5	6	2.4	0.011	0.079
thermosensory neuron WBbt:0005838	1.40E+02	168	1.2	0.011	0.079
Eprp WBbt:0006507	3.1	7	2.2	0.012	0.079
AIM WBbt:0006814	2.6	6	2.3	0.015	0.088
CEP WBbt:0005244	2.6	6	2.3	0.015	0.088
hyp5 WBbt:0004685	3.3	7	2.2	0.015	0.088
ABalaapp WBbt:0006553	2.1	5	2.4	0.015	0.088
ABprpaap WBbt:0006253	2.1	5	2.4	0.015	0.088
Ear WBbt:0006370	2.1	5	2.3	0.018	0.097
ABprapap WBbt:0006678	2.1	5	2.3	0.018	0.097
ABaraaap WBbt:0005861	2.1	5	2.3	0.018	0.097

Appendix C: Additional graphical data

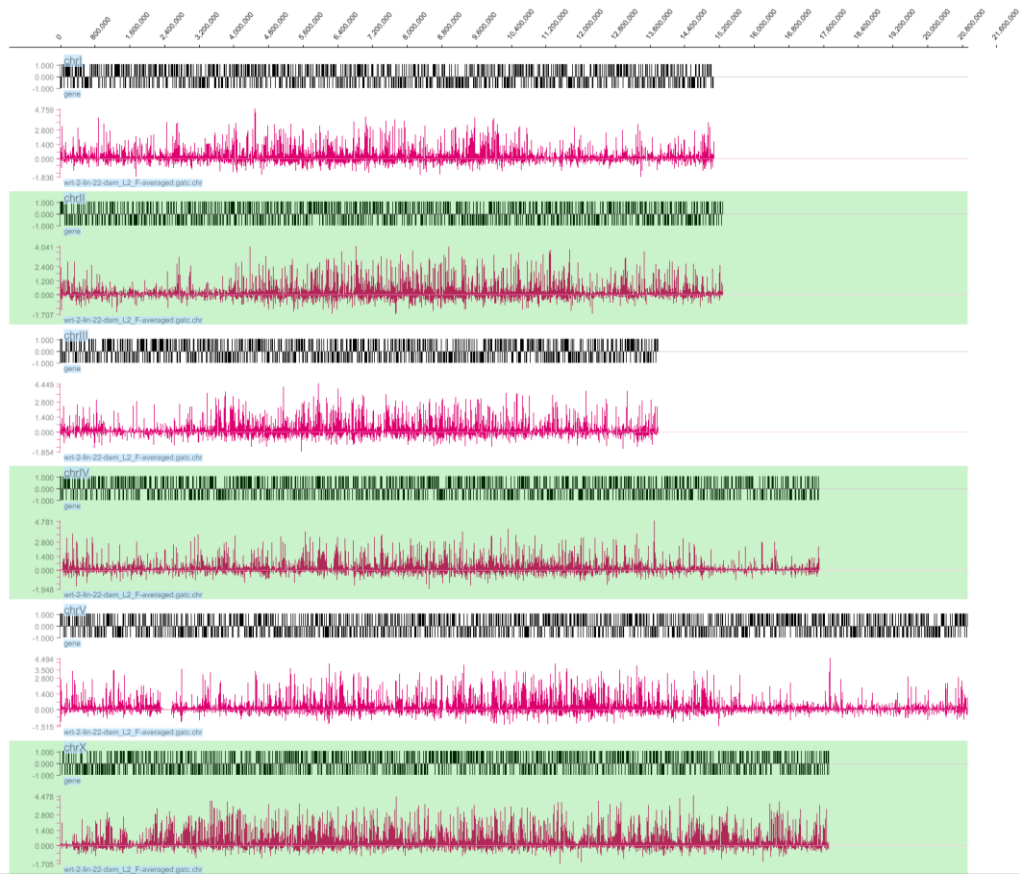## C.1 Representative profiles for LIN-22 and NHR-25 replicate reproducibility
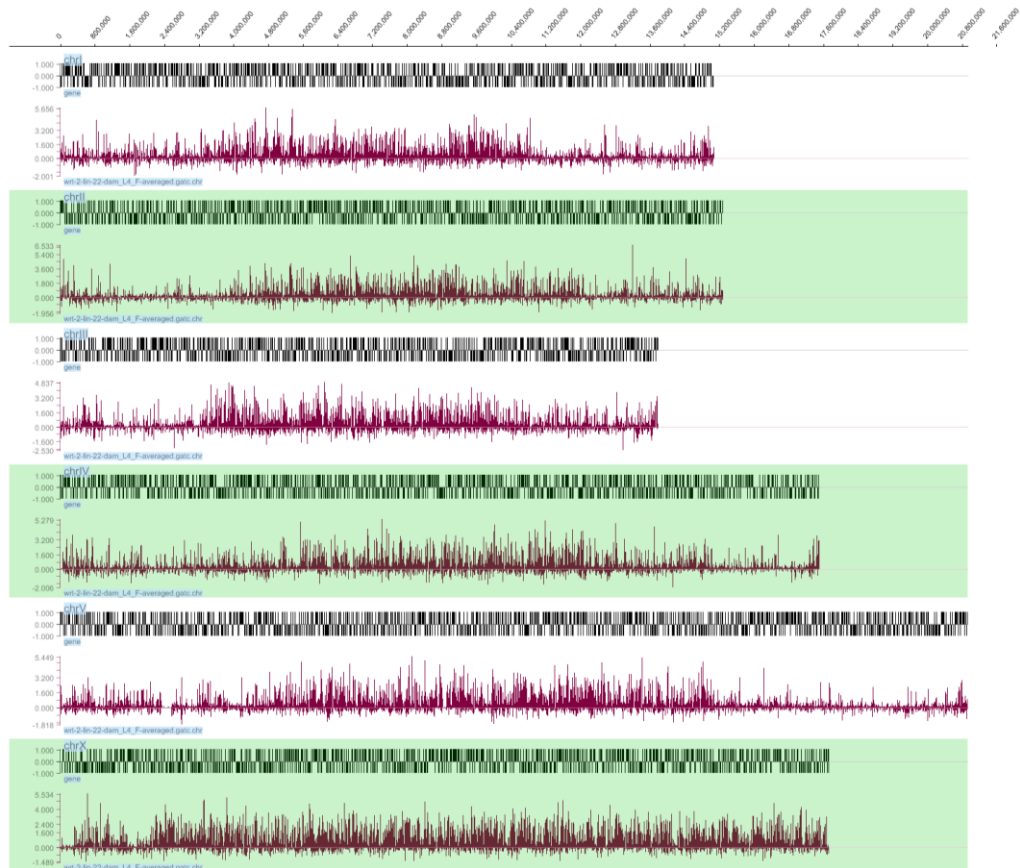
LIN-22
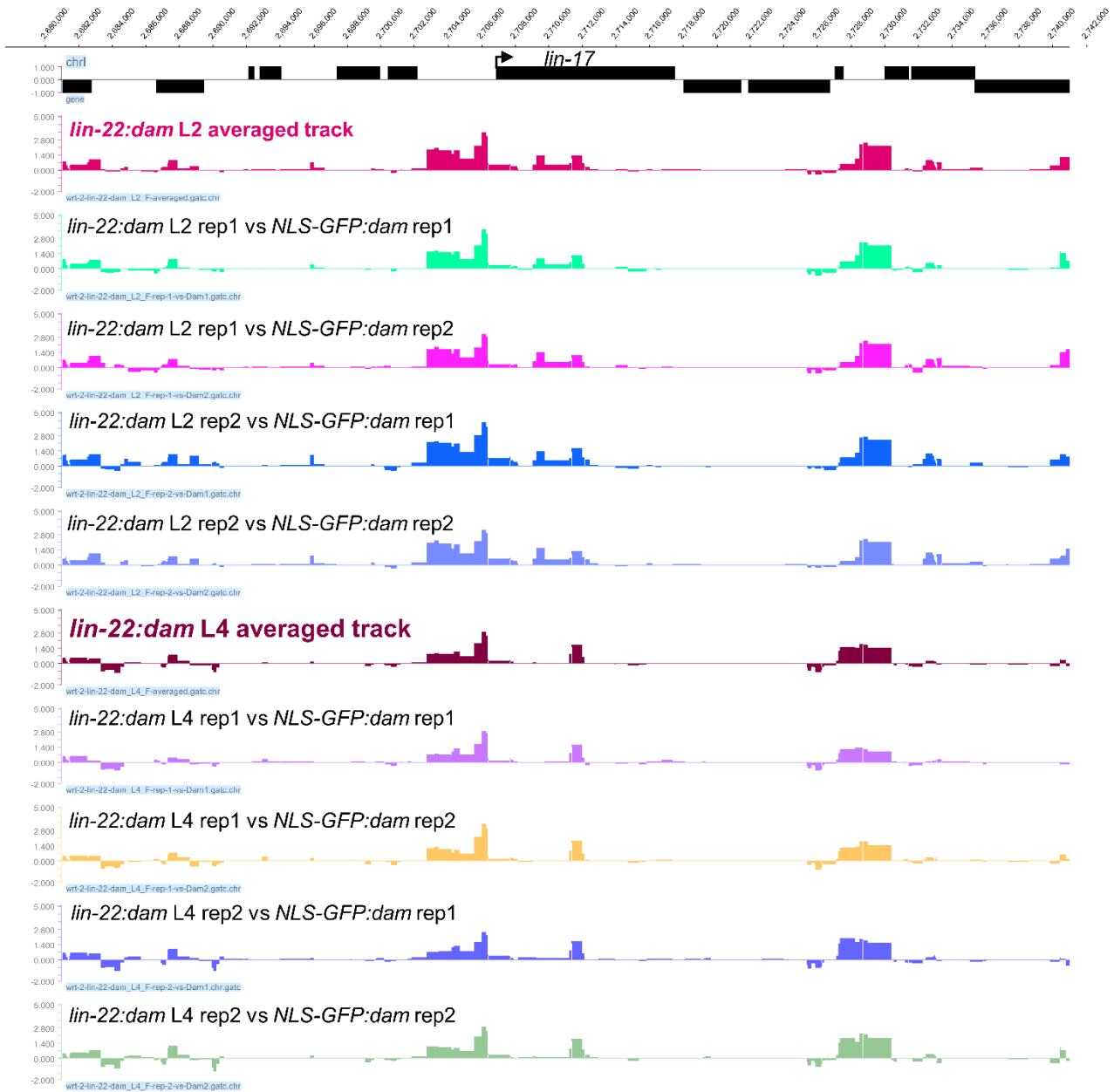
NHR-25

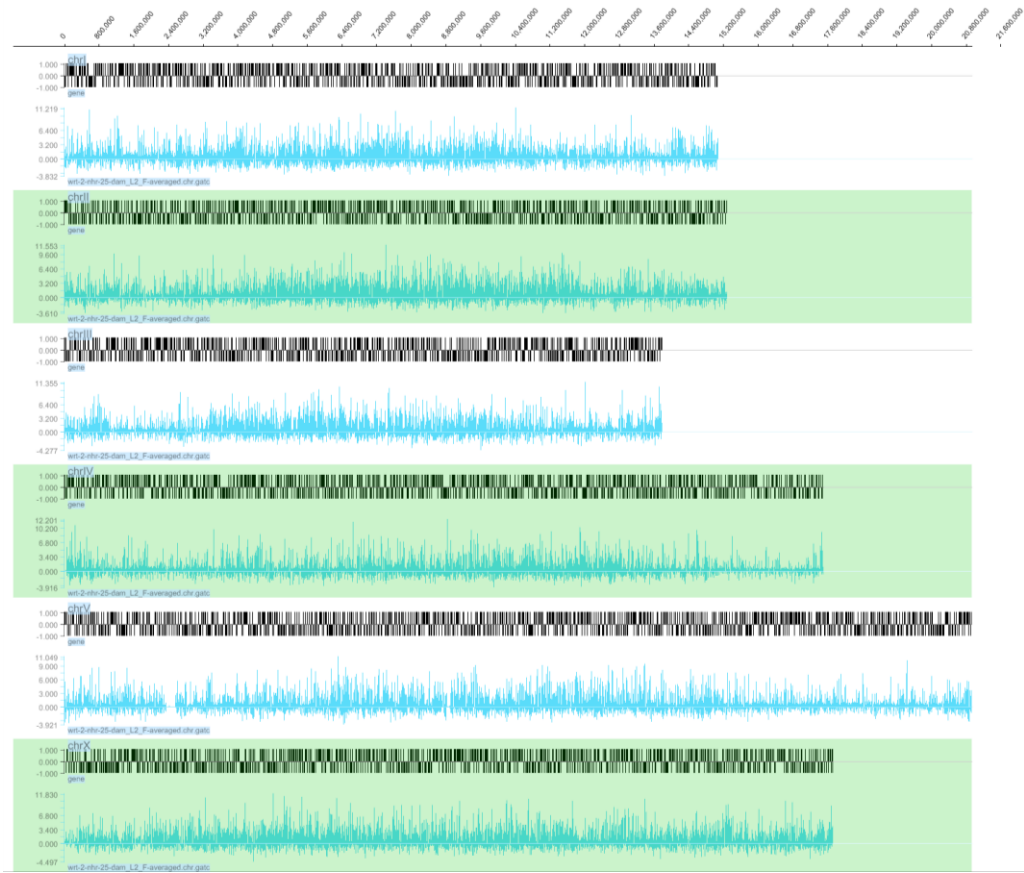## C.2 LIN-22 complete genome-wide averaged signal tracks

L2

L4

C.3 NHR-25 complete genome-wide averaged signal tracks
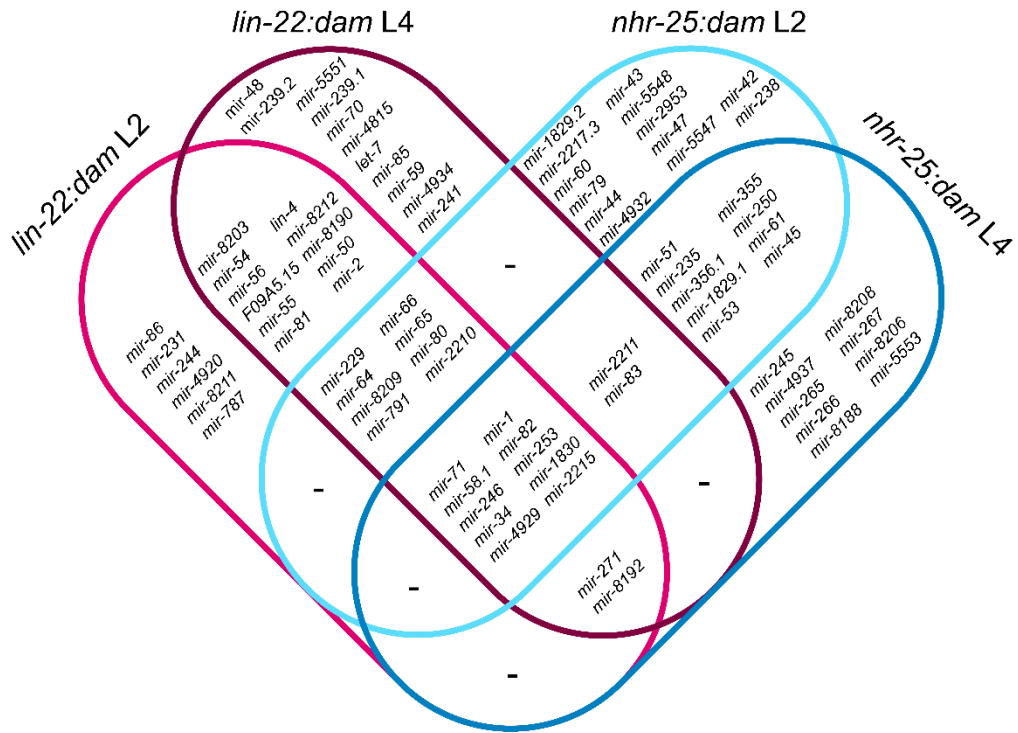
L2

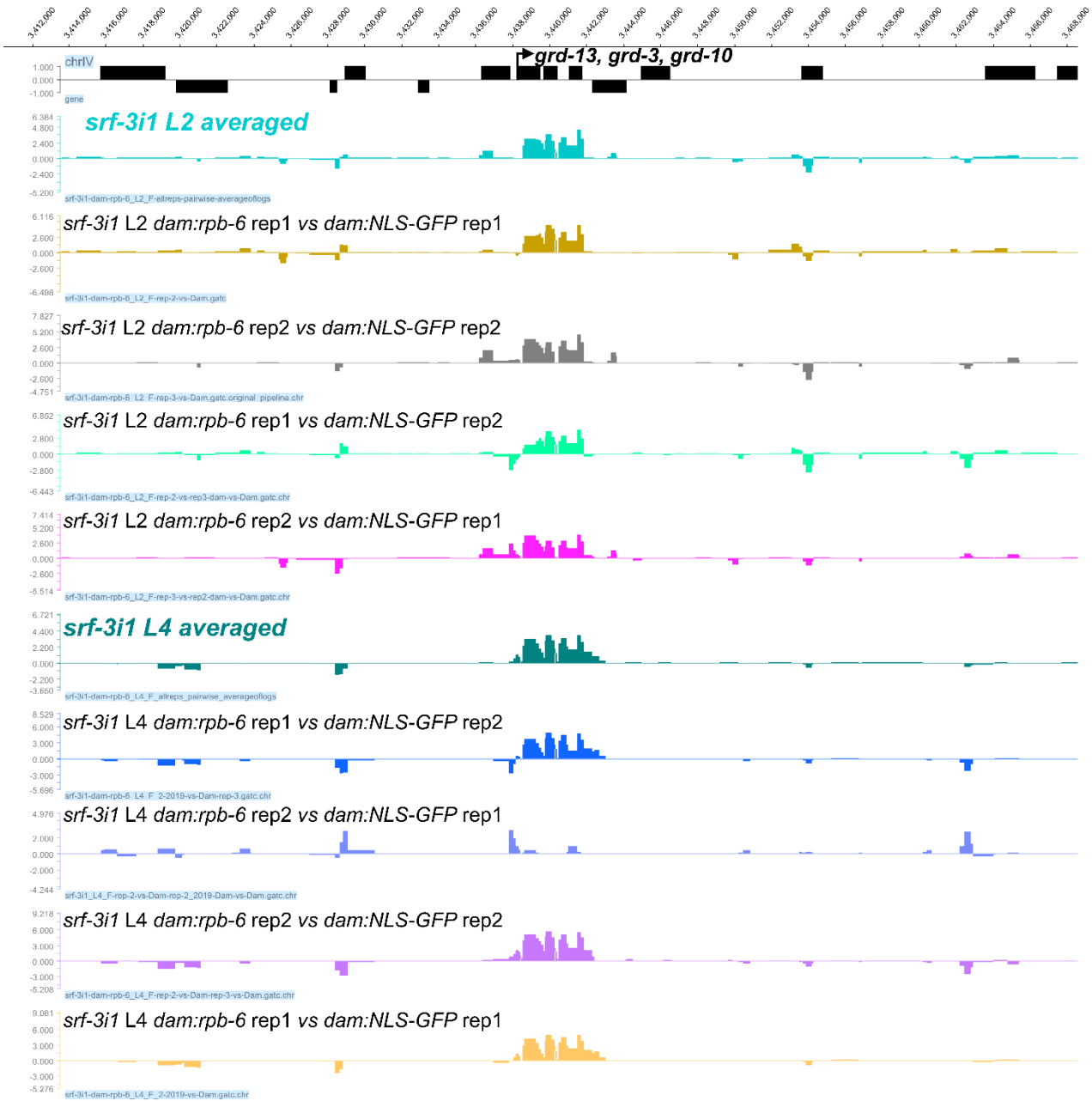
L4

C.4 Criteria selection for NHR-25 DNA motif identification by TaDa
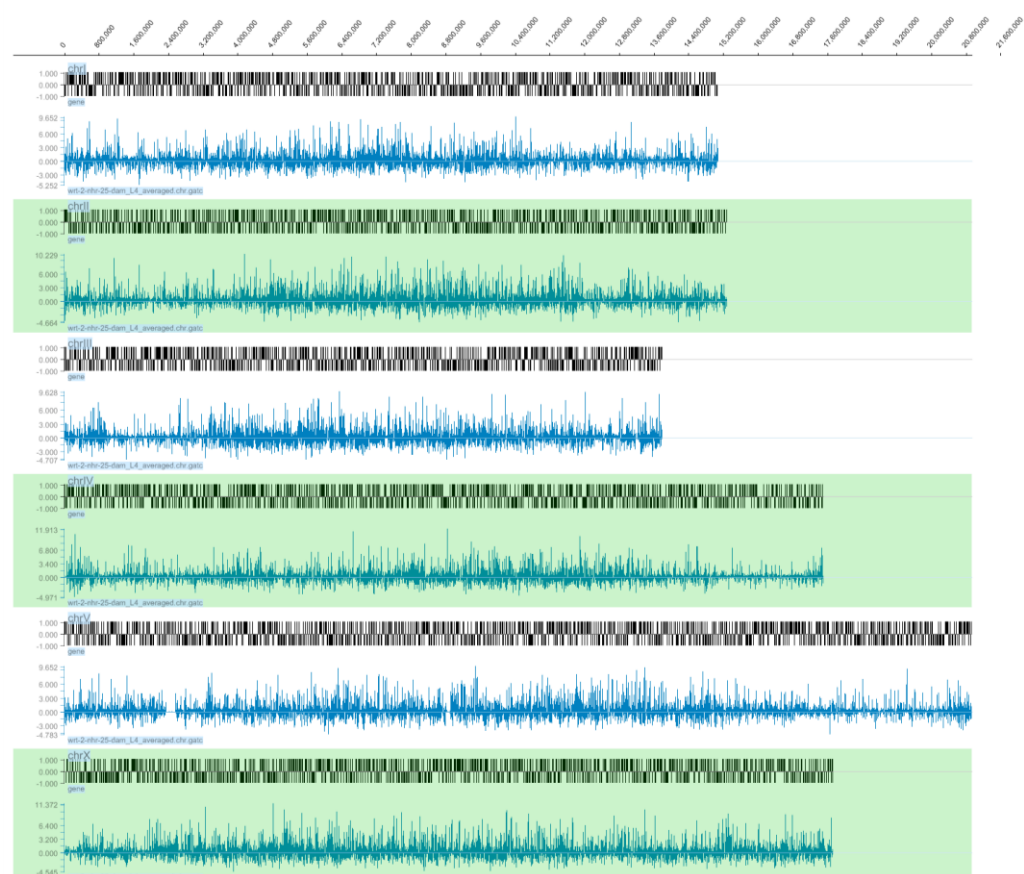
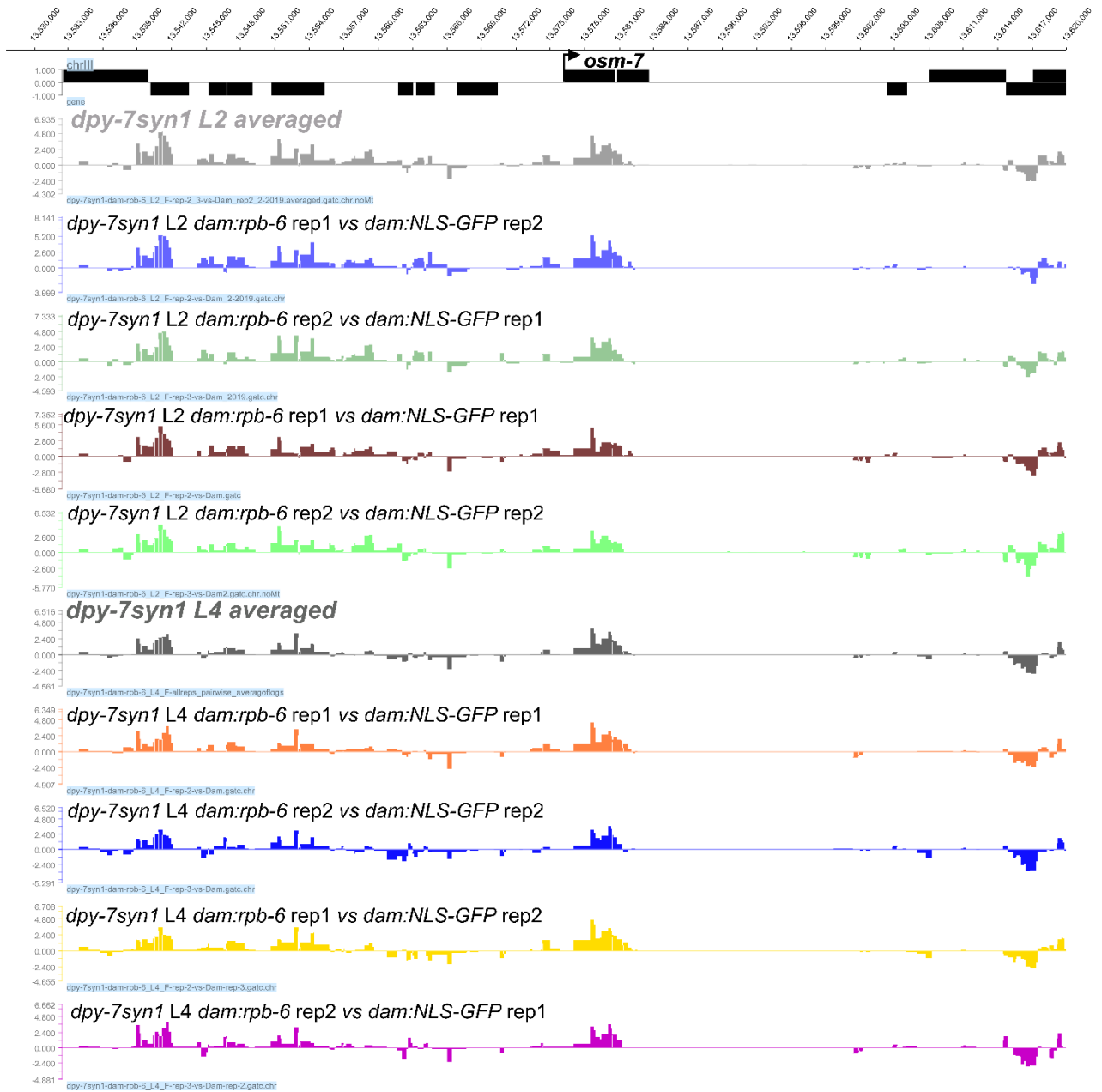## C.5 Venn diagram of putative miRNA targets of LIN-22 and NHR-25 by TaDa

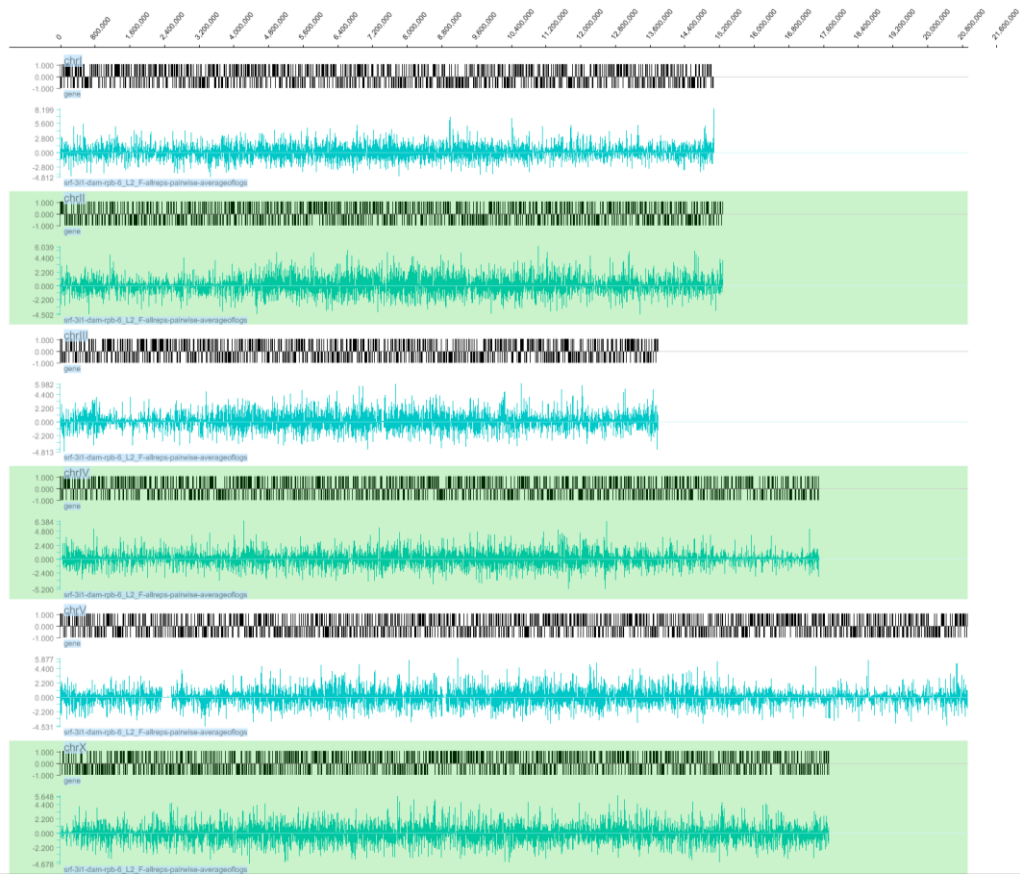# C.6 Representative profiles for *srf-3i1* and *dpy-7syn1* replicate reproducibility
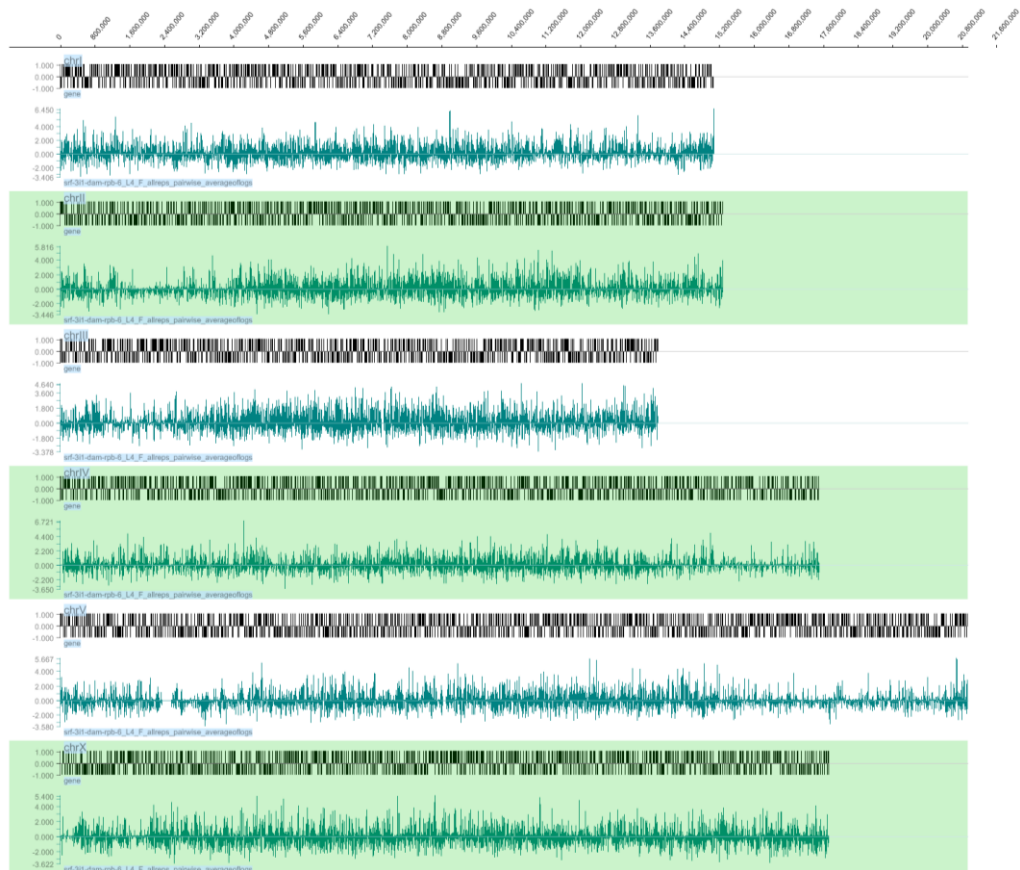
srf-3i1

dpy-7syn1

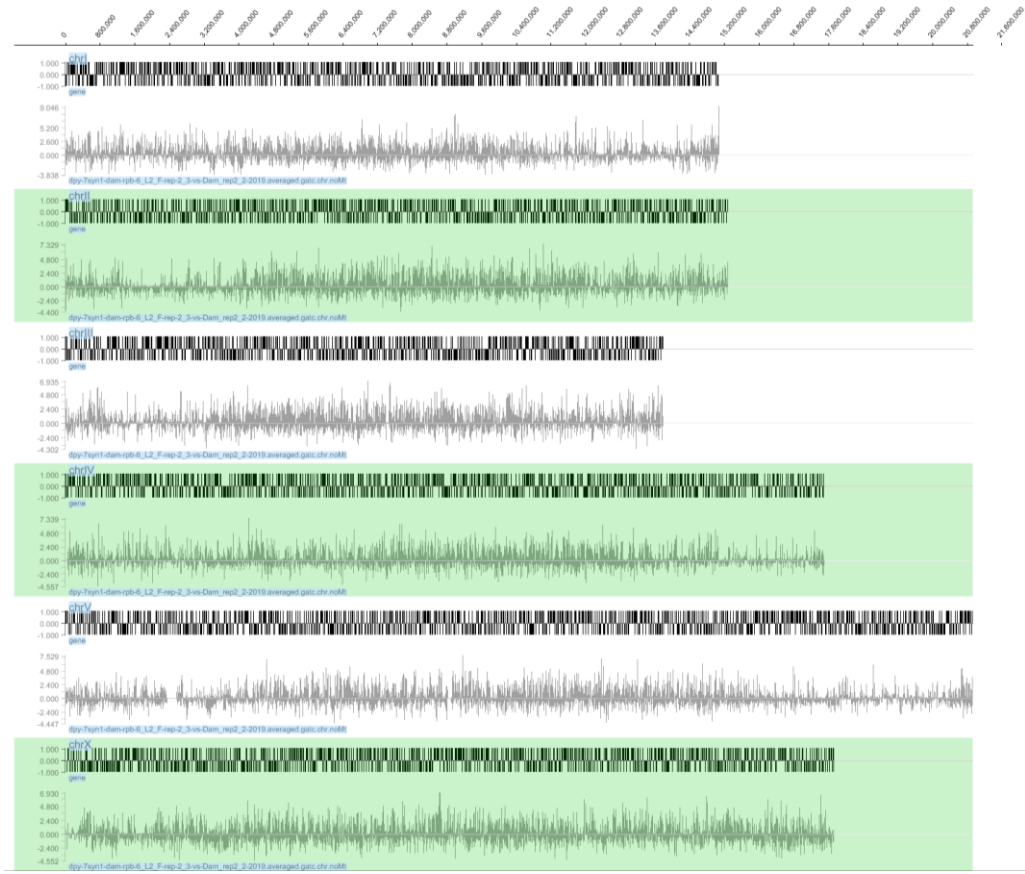## C.7 *srf-3i1* RPB-6 complete genome-wide averaged signal tracks

L2

L4

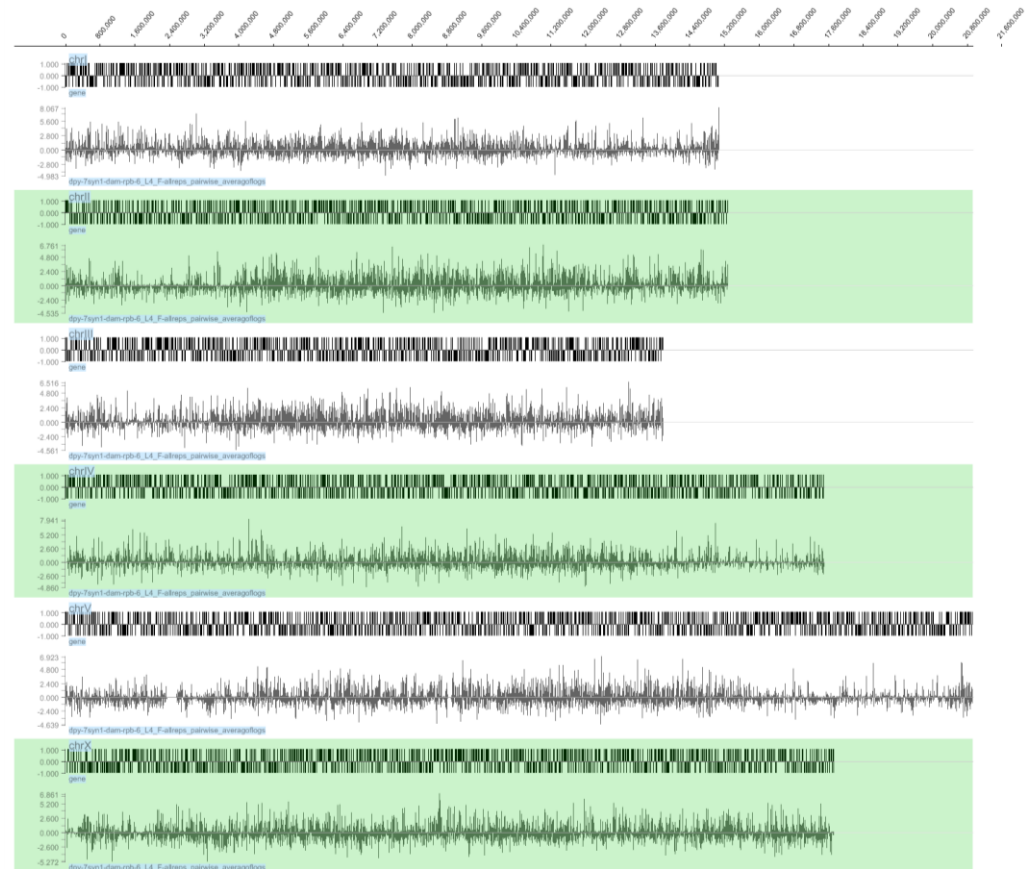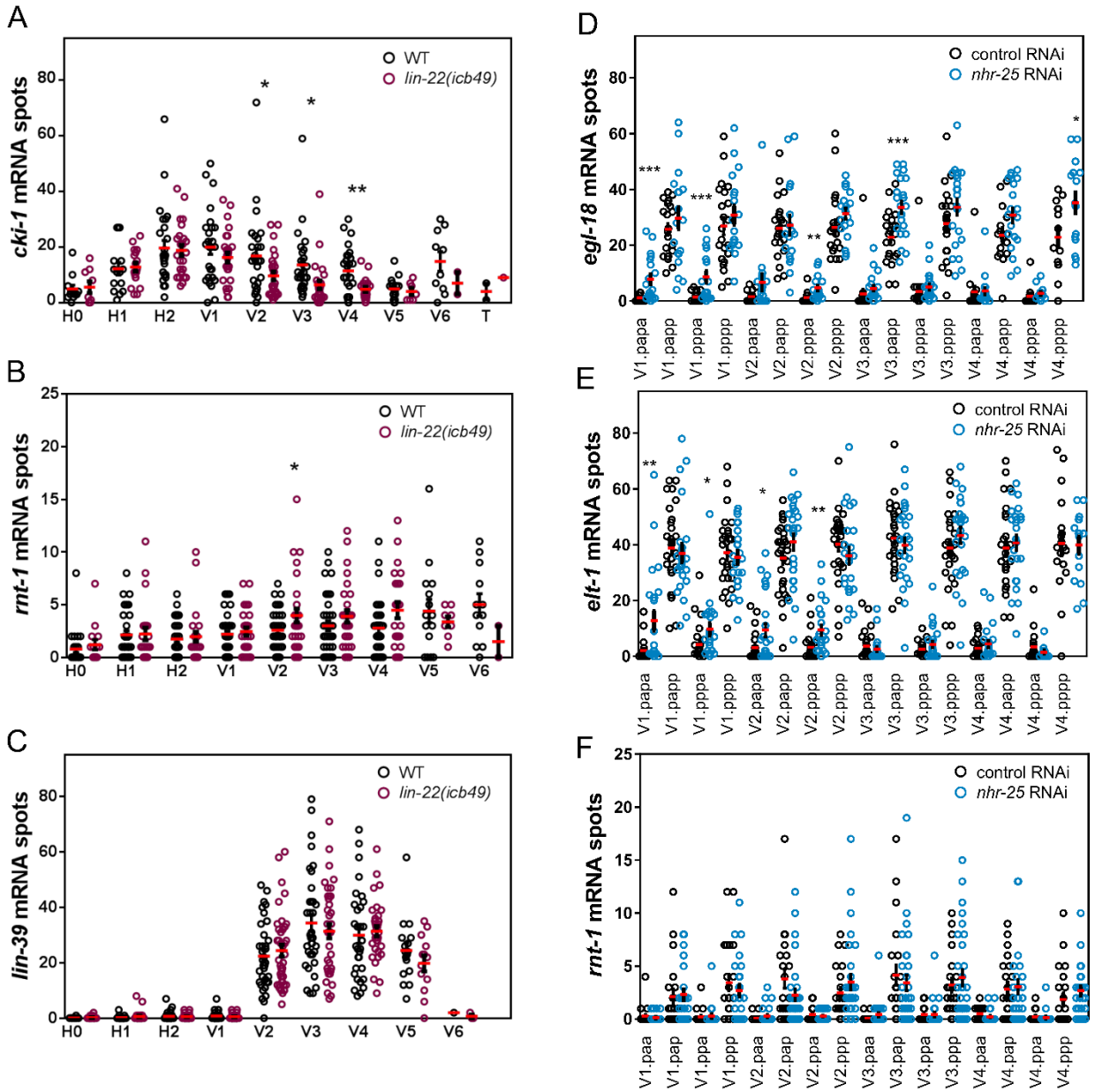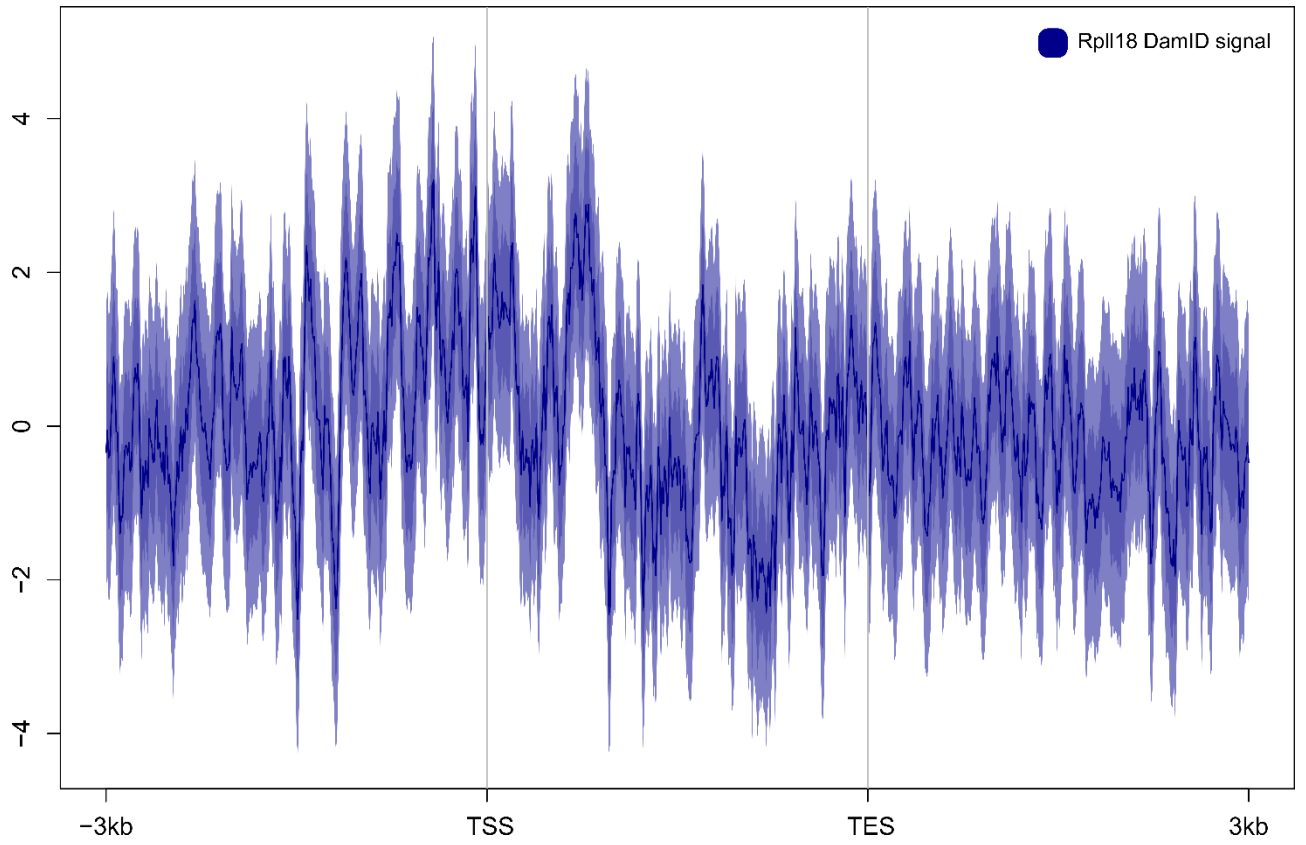## C.8 *dpy-7syn1* RPB-6 complete genome-wide averaged signal tracks

L2

L4

C.9 Seam cell lineage-specific smFISH quantifications

A-C are at the L1 stage; D, E are at the L3 asymmetric division, F is at the L2 asymmetric division

C.10 RpII18 signal aggregation plot across *Drosophila* genes

C.11 All enriched motifs in promoters of genes expressed in *srf-3i1* at both L2 and L4 stages

Homer *de novo* Motif Results (promoter_motifs/)

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 907
Total background sequences = 39045
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1	GAGAGAGAGA	1e-22	-5.199e+01	66.81%	50.34%	582.6bp (772.1bp)	eor-1/MA0543.1/Jaspar(0.738) More Information | Similar Motifs Found	motif file (matrix)
2	CTCGCTGCGA	1e-20	-4.613e+01	22.71%	11.57%	584.3bp (778.4bp)	POL013.1_MED-1/Jaspar(0.582) More Information | Similar Motifs Found	motif file (matrix)
3	AACGAAAAGA	1e-17	-3.972e+01	56.45%	42.08%	591.7bp (712.7bp)	blmp-1/MA0537.1/Jaspar(0.665) More Information | Similar Motifs Found	motif file (matrix)
4	TCTCTCGTTG	1e-16	-3.824e+01	28.11%	16.75%	610.6bp (743.8bp)	blmp-1/MA0537.1/Jaspar(0.632) More Information | Similar Motifs Found	motif file (matrix)
5	CGCCTCTCCT	1e-16	-3.814e+01	13.67%	5.92%	570.7bp (769.3bp)	POL006.1_BREu/Jaspar(0.566) More Information | Similar Motifs Found	motif file (matrix)
6	GTGCGTCGCG	1e-16	-3.804e+01	15.33%	7.05%	564.3bp (778.1bp)	DPL-1(E2F)/cElegans-Adult-ChIP-Seq(modEncode)/Homer(0.508) More Information | Similar Motifs Found	motif file (matrix)
7	ACACACACAC	1e-14	-3.336e+01	32.64%	21.28%	589.9bp (760.9bp)	daf-12/MA0538.1/Jaspar(0.833) More Information | Similar Motifs Found	motif file (matrix)
8	GACGGTCA	1e-14	-3.330e+01	40.02%	27.81%	596.4bp (750.6bp)	MF0004.1_Nuclear_Receptor_class/Jaspar(0.672) More Information | Similar Motifs Found	motif file (matrix)
9	GATACGCCGC	1e-14	-3.256e+01	7.17%	2.29%	533.1bp (730.4bp)	PL0018.1_hlh-25/Jaspar(0.648) More Information | Similar Motifs Found	motif file (matrix)
10	AGAGAAGAAA	1e-13	-3.198e+01	42.56%	30.37%	618.3bp (717.2bp)	elt-2/MA1701.1/Jaspar(0.644) More Information | Similar Motifs Found	motif file (matrix)
11	CAACGCGC	1e-13	-3.165e+01	45.76%	33.41%	596.5bp (755.0bp)	DPL-1(E2F)/cElegans-Adult-ChIP-Seq(modEncode)/Homer(0.756) More Information | Similar Motifs Found	motif file (matrix)
12	AAAGATAGGG	1e-13	-3.025e+01	34.40%	23.35%	626.5bp (708.3bp)	SD0003.1_at_AC_acceptor/Jaspar(0.599) More Information | Similar Motifs Found	motif file (matrix)
13	AGATATCAGA	1e-12	-2.880e+01	20.18%	11.68%	604.3bp (682.8bp)	elt-6/MA1439.1/Jaspar(0.642) More Information | Similar Motifs Found	motif file (matrix)
14 *	CCTCCGTT	1e-11	-2.546e+01	36.16%	25.85%	605.6bp (744.2bp)	MF0001.1_ETS_class/Jaspar(0.612) More Information | Similar Motifs Found	motif file (matrix)
15 *	GGGCGC	1e-9	-2.188e+01	69.02%	58.91%	597.1bp (770.0bp)	POL006.1_BREu/Jaspar(0.771) More Information | Similar Motifs Found	motif file (matrix)
16 *	GACAACGAGG	1e-9	-2.154e+01	24.26%	16.18%	605.7bp (746.4bp)	PL0007.1_mxl-3/Jaspar(0.595) More Information | Similar Motifs Found	motif file (matrix)
17 *	CGACGTTG	1e-9	-2.152e+01	36.16%	26.71%	637.2bp (757.4bp)	nhr-6/MA1451.1/Jaspar(0.494) More Information | Similar Motifs Found	motif file (matrix)
18 *	ATTGAGAT	1e-9	-2.119e+01	25.47%	17.28%	601.2bp (695.1bp)	lim-7/MA1441.1/Jaspar(0.598) More Information | Similar Motifs Found	motif file (matrix)
19 *	CGTCTTTATG	1e-9	-2.076e+01	4.74%	1.58%	572.8bp (739.0bp)	MF0005.1_Forkhead_class/Jaspar(0.610) More Information | Similar Motifs Found	motif file (matrix)
20 *	ATTCGCTTAT	1e-8	-2.044e+01	13.67%	7.75%	594.9bp (685.9bp)	blmp-1/MA0537.1/Jaspar(0.610) More Information | Similar Motifs Found	motif file (matrix)
21 *	GGCCCCAAGT	1e-8	-1.849e+01	3.20%	0.87%	641.1bp (794.3bp)	POL006.1_BREu/Jaspar(0.546) More Information | Similar Motifs Found	motif file (matrix)
22 *	GTTGACCC	1e-7	-1.733e+01	30.43%	22.50%	590.0bp (723.0bp)	MF0004.1_Nuclear_Receptor_class/Jaspar(0.841) More Information | Similar Motifs Found	motif file (matrix)
23 *	TCACATGA	1e-4	-1.049e+01	10.14%	6.51%	546.5bp (673.5bp)	hlh-30/MA1449.1/Jaspar(0.820) More Information | Similar Motifs Found	motif file (matrix)
24 *	GATTAA	1e-4	-1.024e+01	68.80%	62.36%	590.5bp (662.6bp)	unc-30/MA1443.1/Jaspar(0.880) More Information | Similar Motifs Found	motif file (matrix)

C.12 All enriched motifs in promoters of genes expressed in *dpy-7syn1* at both L2 and L4 stages

Homer *de novo* Motif Results (promoter_motifs/)

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 1315
Total background sequences = 39309
* - possible false positive

| Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-31	-7.349e+01	67.22%	50.78%	595.1bp (740.3bp)	eor-1/MA0543.1/Jaspar(0.748) More Information | Similar Motifs Found	motif file (matrix)
2		1e-31	-7.293e+01	25.02%	12.72%	590.4bp (736.6bp)	mab-3/MA0262.1/Jaspar(0.593) More Information | Similar Motifs Found	motif file (matrix)
3		1e-27	-6.319e+01	71.63%	56.74%	583.2bp (700.0bp)	blmp-1/MA0537.1/Jaspar(0.759) More Information | Similar Motifs Found	motif file (matrix)
4		1e-23	-5.339e+01	61.06%	47.03%	572.2bp (739.3bp)	eor-1/MA0543.1/Jaspar(0.628) More Information | Similar Motifs Found	motif file (matrix)
5		1e-19	-4.580e+01	29.13%	18.41%	606.7bp (679.7bp)	ELT-3(Gata)/cElegans-L1-ELT3-ChIP-Seq(modEncode)/Homer(0.953) More Information | Similar Motifs Found	motif file (matrix)
6		1e-18	-4.197e+01	9.51%	3.84%	569.3bp (774.3bp)	POL011.1_XCPE1/Jaspar(0.699) More Information | Similar Motifs Found	motif file (matrix)
7		1e-16	-3.827e+01	66.24%	54.63%	594.2bp (708.0bp)	lin-14/MA0261.1/Jaspar(0.699) More Information | Similar Motifs Found	motif file (matrix)
8		1e-16	-3.820e+01	20.53%	12.16%	563.7bp (758.7bp)	POL006.1_BREu/Jaspar(0.654) More Information | Similar Motifs Found	motif file (matrix)
9		1e-15	-3.683e+01	3.42%	0.69%	561.6bp (711.9bp)	POL013.1_MED-1/Jaspar(0.564) More Information | Similar Motifs Found	motif file (matrix)
10		1e-15	-3.636e+01	52.78%	41.36%	604.3bp (710.4bp)	cebp-1/MA1444.1/Jaspar(0.623) More Information | Similar Motifs Found	motif file (matrix)
11		1e-14	-3.425e+01	52.85%	41.79%	587.8bp (740.7bp)	efl-1/MA0541.1/Jaspar(0.603) More Information | Similar Motifs Found	motif file (matrix)
12		1e-14	-3.348e+01	30.34%	20.97%	579.1bp (732.8bp)	MF0004.1_Nuclear_Receptor_class/Jaspar(0.619) More Information | Similar Motifs Found	motif file (matrix)
13		1e-13	-3.078e+01	11.25%	5.74%	604.6bp (758.2bp)	MF0002.1_bZIP_CREB/G-box-like_subclass/Jaspar(0.539) More Information | Similar Motifs Found	motif file (matrix)
14		1e-12	-2.992e+01	17.79%	10.83%	614.5bp (748.4bp)	ceh-28/MA1445.1/Jaspar(0.547) More Information | Similar Motifs Found	motif file (matrix)
15		1e-12	-2.953e+01	48.59%	38.48%	577.8bp (728.6bp)	daf-12/MA0538.1/Jaspar(0.767) More Information | Similar Motifs Found	motif file (matrix)
16		1e-12	-2.900e+01	52.62%	42.51%	606.5bp (699.3bp)	MF0008.1_MADS_class/Jaspar(0.654) More Information | Similar Motifs Found	motif file (matrix)
17 *		1e-11	-2.577e+01	18.40%	11.81%	621.1bp (668.9bp)	blmp-1/MA0537.1/Jaspar(0.597) More Information | Similar Motifs Found	motif file (matrix)
18 *		1e-9	-2.274e+01	72.40%	64.03%	609.5bp (743.3bp)	PL0007.1_mrl-3/Jaspar(0.586) More Information | Similar Motifs Found	motif file (matrix)
19 *		1e-8	-2.042e+01	16.73%	11.11%	583.8bp (692.5bp)	che-1/MA0260.1/Jaspar(0.550) More Information | Similar Motifs Found	motif file (matrix)
20 *		1e-8	-2.030e+01	5.78%	2.66%	560.9bp (713.3bp)	POL002.1_INR/Jaspar(0.703) More Information | Similar Motifs Found	motif file (matrix)
21 *		1e-8	-1.899e+01	13.69%	8.78%	660.5bp (719.1bp)	POL009.1_DCE_S_II/Jaspar(0.729) More Information | Similar Motifs Found	motif file (matrix)
22 *		1e-5	-1.235e+01	50.72%	44.47%	590.8bp (699.2bp)	unc-30/MA1443.1/Jaspar(0.794) More Information | Similar Motifs Found	motif file (matrix)
23 *		1e-4	-1.055e+01	9.58%	6.56%	550.3bp (666.2bp)	hlh-30/MA1449.1/Jaspar(0.850) More Information | Similar Motifs Found	motif file (matrix)
24 *		1e-3	-7.789e+00	5.25%	3.39%	627.7bp (762.9bp)	POL004.1_CCAAT-box/Jaspar(0.550) More Information | Similar Motifs Found	motif file (matrix)
25 *		1e-1	-3.321e+00	95.74%	94.59%	599.6bp (734.4bp)	POL009.1_DCE_S_II/Jaspar(0.598) More Information | Similar Motifs Found	motif file (matrix)

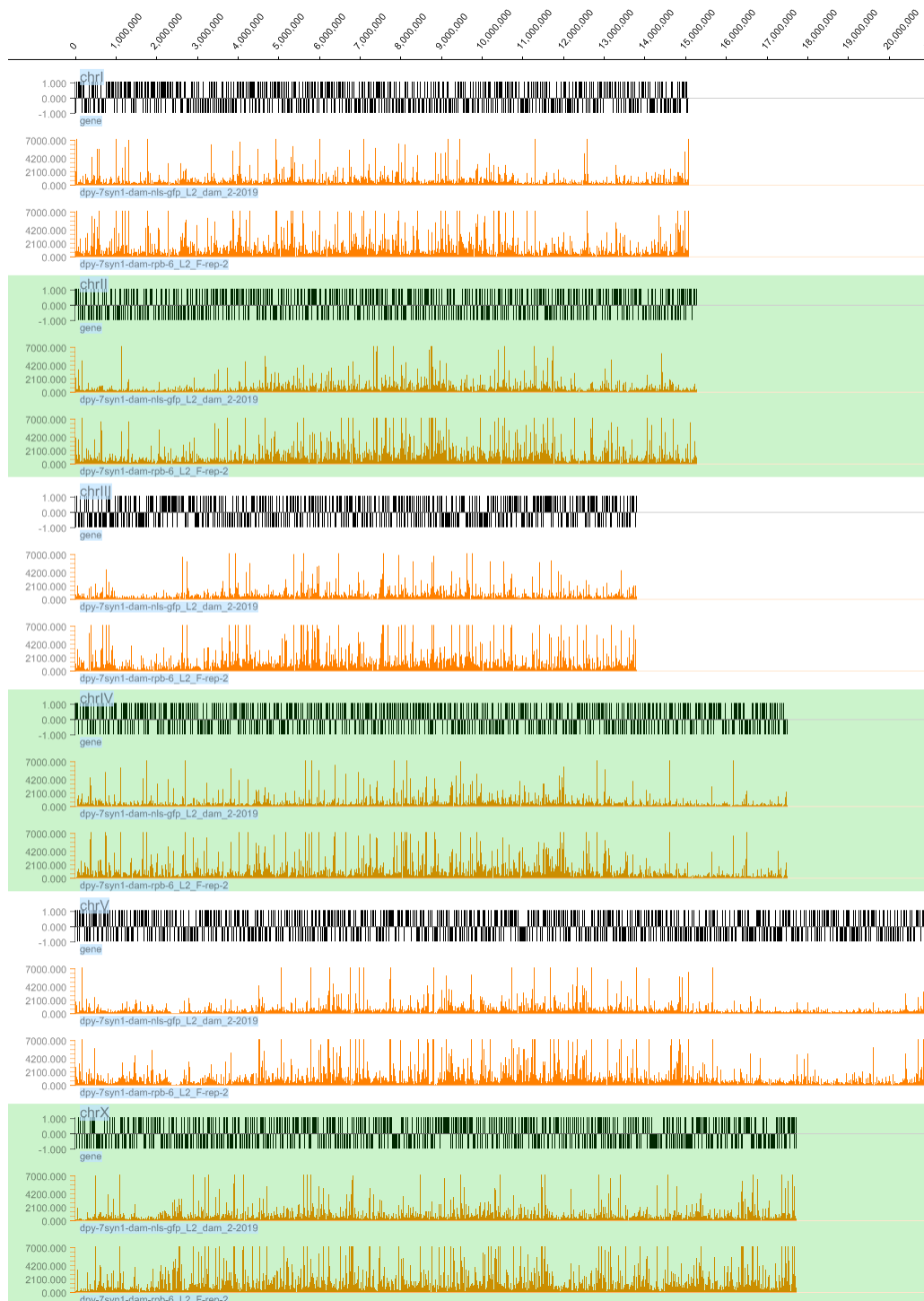## C.13 *srf-3i1* CATaDa complete genome-wide averaged signal tracks

L2
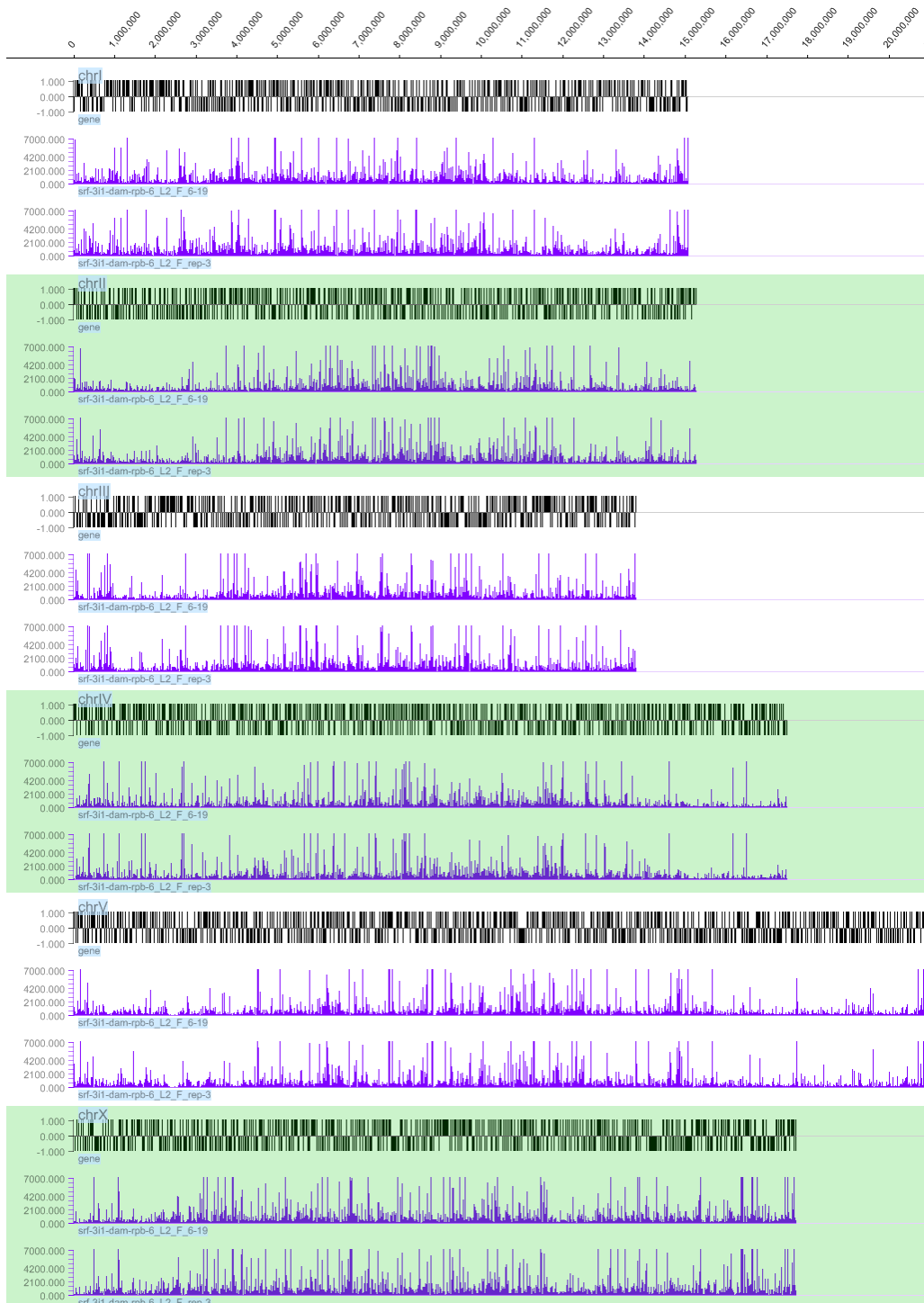
L4
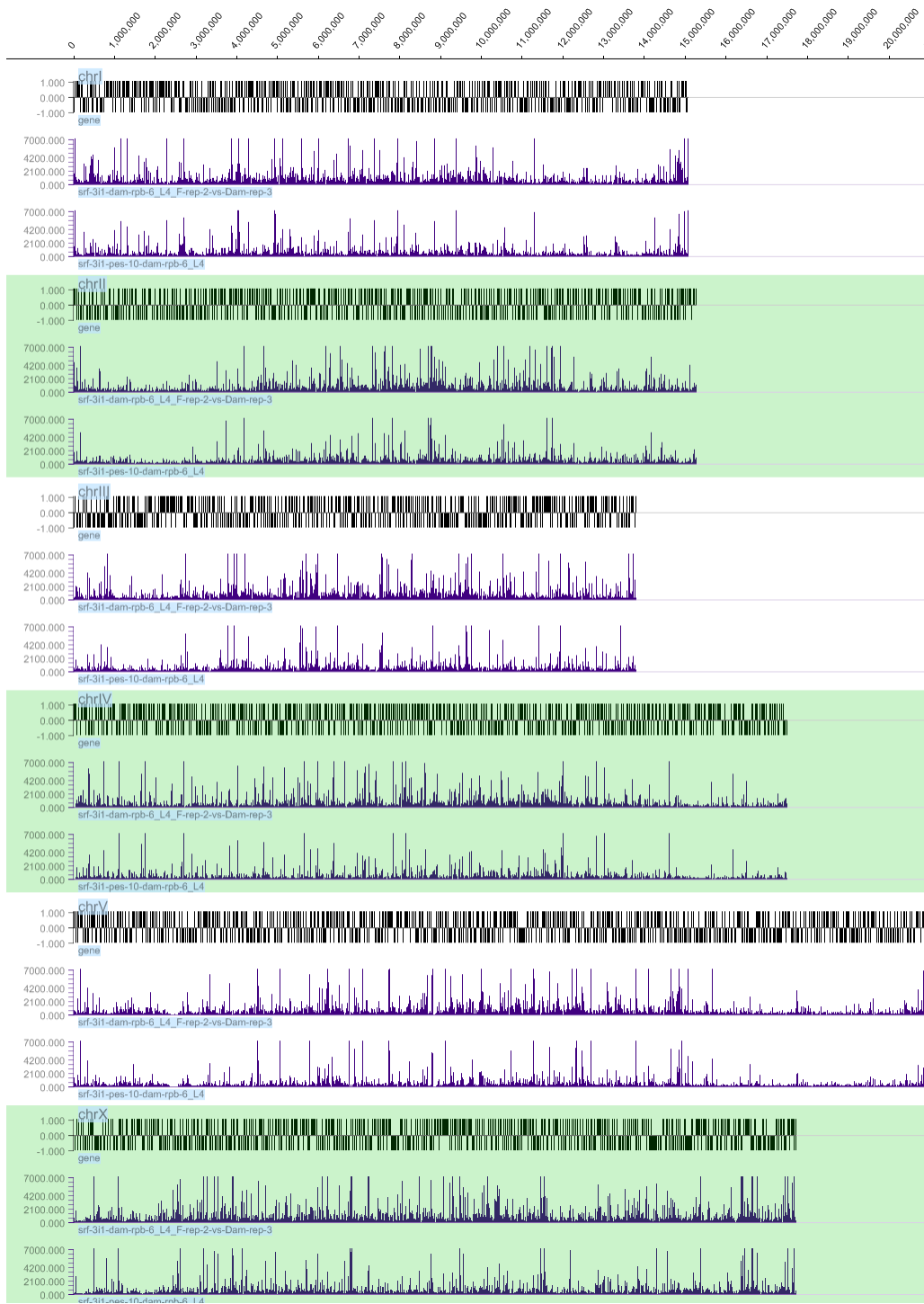
C.14 *dpy-7syn1* CATaDa complete genome-wide averaged signal tracks

L4