An Epigenetic Analysis of Planarian Stem Cells

Anish Dattani

St Anne's College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas Term 2018

Declaration

I hereby declare that this thesis entitled "An Epigenetic Analysis of Planarian Stem Cells" has been originally carried out by me under the supervision of Professor Aziz Aboobaker. This work has not formed the basis for award of any degree or diploma previously. The particulars given in the thesis are true to the best of my knowledge.

Anish Dattani
St Anne's College
University of Oxford

Thesis Abstract

Planarian flatworms possess somatic pluripotent stem cells, called neoblasts (NBs), which are able to differentiate into all cell types that constitute the adult body plan. Consequently, planarians possess remarkable regenerative capacities, and have become an invertebrate model system to study stem cell responses during regeneration. Transcriptomic studies have revealed the genes needed to both maintain NB pluripotency and ensure correct lineage specification during differentiation. However, these studies have not elucidated how this regulation of expression is controlled at the epigenetic level, and in particular by diverse histone modifications. In this thesis, we present a case for elevating planarians as a model system for studying the epigenetic regulation of stem cell pluripotency and differentiation.

Firstly, we describe an expression-based annotation of the asexual *Schmidtea mediterranea* genome. For each annotated locus, we allocate proportional values for a gene's expression in either S/G2/M NBs (X1), G1 NBs + post-mitotic progeny (X2), or differentiated cells (Xins) – the three broadly-defined cellular compartments that can be isolated from planarians using Fluorescence Activated Cell Sorting (FACS). The production of a well-annotated genome incorporating transcriptomic information serves as the basis for correlating the presence of particular histone modifications with underlying gene expression. We have developed an optimized ChIP-seq protocol for planarian NBs and show that the active histone marks H3K4me3 and H3K36me3 and suppressive H3K4me1 and H3K27me3 marks correlate with the transcriptional output of genes. We also show that genes with little transcriptional activity in NBs, but which switch on in post-mitotic progeny during differentiation are bivalent, being marked by both H3K4me3 and H3K27me3 at the promoter-proximal region. Bivalent histone modifications in mammalian embryonic and germline stem cells enable transcriptional poising of genes, and consistent with this we find that bivalent genes in planarian NBs are marked by paused RNA Pol II at the promoter-proximal region.

In addition to histone modifications, enhancers and their associated transcription factors can also directly influence gene expression. Consequently, we elucidate the potential TF repertoire of planarians, and identify those enriched in NBs. We also present preliminary evidence to suggest that ATAC-seq on planarian cells can identify both transcriptionally permissive gene promoters, at least in differentiated cells, as well as putative enhancers that correlate with expression of neighbouring genes. In the future, we hope to be able to build gene regulatory networks by identifying TF-binding sites in open chromatin regions and elucidating enhancer targets in a range of planarian cell types.

Acknowledgements

First, and foremost, I would like to thank my supervisor Aziz Aboobaker for his help and guidance over the course of my time in his lab. I am grateful for his patience and encouragement and, most importantly, the stimulating scientific discussions we have had.

I would also like to thank Damian Kao for introducing me to bioinformatics, and being my collaborator on this project. Damian made me understand the value of using both publicly available tools and basic python scripts to effectively analyse NGS data, and was always available to offer feedback on problems.

Thank you to Yuliana Mihaylova who, in her last few months in the lab, patiently supervised my first ChIP-seq experiments. Without her efforts to optimize this protocol, none of this work in this thesis would have been possible.

I would also like to thank Sounak Sahu, who has been a great colleague, and with whom I have enjoyed discussing and working through problems. Also many thanks to Nobuyoshi Kosaka, Prasad Abnave, and Divya Sridhar who have been supportive colleagues throughout my time in the lab.

Thank you to my parents and Alice whose love and support have kept me both sane and happy.

Contents Page

Thesis outline

The aim of this thesis is to present a case for utilizing planarians as a useful and potentially rewarding organism for research into the epigenetic mechanisms underlying stem cell maintenance and differentiation.

In **Chapter I** we provide the reader with a primer of planarian biology, and summarize the various transcriptomic studies that have identified the genes involved in neoblast (NB) maintenance and commitment to the various differentiated cell lineages. In **Chapter II** we present a case for why planarians present an advantage over conventional mammalian cell-culture based systems in epigenetics studies of stem cell pluripotency and differentiation. We summarise the limited studies in planarians which describe a function for conserved components of the histone modification machinery, and we suggest how new tools such as ChIP-seq and ATAC-seq can be applied to these investigations to yield greater molecular and evolutionary insights. Given that ChIP-seq data requires a well-annotated genome to correlate the presence of histone modifications with underlying genes, in **Chapter III** we outline our expression-based annotation of the asexual *Schmidtea mediterranea* genome. We also incorporate transcriptomic data to allocate a percentage of total expression to each annotated locus in the S/G2/M NBs (X1), stem cell progeny + G1 NBs (X2), and differentiated cell (Xins) compartments. The generation of this resource helps in **Chapter IV** where we conduct ChIP-seq for conserved histone marks in planarian NBs and show that a gene's epigenetic environment correlates with our transcriptomic categorization. We find that specific genes with little transcriptional activity in the X1 NB compartment, but which are enriched for expression in X2 post-mitotic progeny, are marked with the opposing histone modification marks H3K4me3 and H3K27me3 at promoter proximal regions. This provides evidence for the existence of bivalent promoters in stem cells outside of vertebrates. In **Chapter V** we present an annotation of the newly released sexual genome of *S. mediterranea* (Grohme et al. 2018) that improves on contiguity, and utilize this to identify transcription factors genome-wide as well as those that have potential NB functions. We present preliminary data to show that ATAC-seq on planarian FACS sorted cells can

be used to identify open chromatin regions of the genome such as transcriptionally permissive promoters and putative cis-regulatory elements. In **Chapter VI,** we discuss strategies that will enable us to dissect out the regulatory landscape of planarian stem cells and the plasticity of the cistrome that allows for well-coordinated differentiation.

Chapter I

General Planarian Introduction

Abstract

Planarian flatworms are capable of profound feats of regeneration fuelled by a population of adult stem cells called neoblasts (NBs). This chapter provides a primer to planarian biology and synthesizes various studies that have increased our understanding of molecular and cellular principles that enable NBs to both maintain their identity as well as differentiate precisely under homeostatic and regenerative scenarios. In particular, we review the various transcriptomic approaches that have uncovered substantial NB heterogeneity and efforts to identify the individually pluripotent clonogenic neoblast (cNeoblast). These transcriptomic studies provide a framework with which to understand the epigenetic regulation of neoblasts and their commitment to various lineages.

1.1 Planarian phylogeny and anatomy

Planarians or triclads are an order of triploblastic, unsegmented, free-living worm acoelomates within the phylum Platyhelminthes (*Platy*, flat; *helminth*, worm), which additionally includes parasitic clades such Cestoda (tapeworms), Trematoda (flukes) and Monogenea (fish gill parasites) **(Figure 1A).** Molecular phylogenetic analyses have classified Platyhelminthes within the superphylum Lophotrochozoa, a largely neglected taxonomic group in molecular and cellular research (Ruiz-Trillo et al. 1999; Aguinaldo et al. 1997; Struck et al. 2014; Egger et al. 2015). Currently, planarians are classified according to three suborders, Maricola (salt-water dwelling), Cavernicola (cave-dwelling) and Continenticola (comprising of the old orders of Terricola (land) and Paludicola (freshwater) planarians) (Hallez 1892; Carranza et al. 1998). Most relevant to researchers investigating the regenerative capabilities of these worms, is the Dugesiidae family within the Continenticola whose members include *Dugesia japonica*, *Giardia tigrina,* and *Schmidtea mediterranea.*

Planarians are devoid of a coelom, with all internal organs separated from the body wall by mesenchymal tissue commonly known as parenchyma. However, planarians do have derivatives of all three germ layers (ectoderm, mesoderm, and endoderm). They possess a nervous system consisting of a bi-lobed cephalic ganglion, connected to two ventral longitudinal nerve cords interconnected by commissural neurons extending to the tail-end of the animal (Cebrià et al. 2002; Robb and Alvarado 2002). Sensory structures, such as photoreceptors (Carpenter et al. 1974), chemoreceptors (MacRae 1967), and pressure receptors (Hyman 1951) are found at the anterior end of the animal and send projections to the cephalic ganglia. A centrally located pharynx is used both as mouth and anus, and is connected to a three-branched (triclad) digestive system consisting of one anterior and two posterior gut branches (Newmark and Alvarado 2002) (**Figure 1B and 1C**). Waste expulsion and osmoregulation are facilitated by a network of protonephridia, which consist of ciliated 'flame cells' that work to filter out water and small molecules from the mesenchyme and expel them via long tubules ending at the surface of the planarian (McKanna 1968; Ishii 1980; Rink

et al. 2011; Thi-Kim Vu et al. 2015). Ventral ciliated epithelial cells are responsible for their gliding locomotion, whereas muscle fibres being used to orient the direction of movement, maintain the integrity of the planarian body, and guide regeneration (Hyman 1951; Cebrià et al. 1997; Orii et al. 2002; Scimone et al. 2017). Planarians can reproduce sexually as cross-fertilizing hermaphrodites, or, in the case of a number of species (including *D. japonica*, *G. tigrina* and *S. mediterranea*) an asexual biotype exists that reproduces by undergoing transverse fission posterior to the pharynx (Hyman 1951; Sahu et al. 2017) (**Figure 1D**). The asexual biotype evolved from sexual animals and can be karyotypically distinguished by a Robertsonian translocation (fusion of chromosome 1 to chromosome 3) that may be responsible for their lack of ability to differentiate a germline and somatic copulatory organs (Baguñà et al. 1999; Newmark and Alvarado 2002) .

1.2 Schmidtea mediterranea: filling the experimental gap in regeneration studies

S. mediterranea has received extensive attention in biological studies of regeneration. These animals are able to regenerate all their tissues irrespective of whether they are derived from endoderm, ectoderm or mesoderm, yet their evolutionary position means that they share with vertebrates the developmental pathways responsible for the bilaterian Bauplan. Furthermore, planarians are relatively easy to culture in the laboratory, exhibit developmental plasticity (includes both sexual and asexual forms of reproduction), have a diploid genome that has been fully sequenced (Robb et al. 2007; Grohme et al. 2018), and various *de novo* assembled transcriptomes are publicly available via an online, mineable repository (Brandl et al. 2016; Rozanski et al. 2019). Cellular and molecular techniques such as *in situ* hybridization (King and Newmark 2013; Pearson et al. 2009), immunohistochemistry (Ross et al. 2015), and RNA interference (RNAi) (Sanchez Alvarado and Newmark 1999; Reddien et al. 2005a) have also been developed in *S. mediterranea* facilitating studies of the animal's stem cells and regenerative capacity. As of yet, however, reproducible genetic manipulation via transgenic methods is currently lacking in *S. mediterranea*, which would greatly assist in-depth studies i.e., gene overexpression, dissection of regulatory elements, real-time imaging, and lineage-tracing.

1.3 Planarian regeneration, the role of NBs, and their identification

A fundamental feature of planarians is that they contain a population of self-renewing pluripotent stem cells, called neoblasts (NBs), which divide to replace any cell type during homeostasis or following tissue injury. Indeed, before his seminal work on the problem of inheritance in Drosophila, T.H. Morgan observed that a fragment as small as 1/279 of a planarian could regenerate a complete animal (Morgan 1898). Upon injury, regeneration begins with rapid closing of a wounded surface, achieved by the protective spreading of existing epithelial cells. Subsequently, NBs divide rapidly throughout the worm, with mitotic numbers peaking 6 hours post wounding (Wenemoser and Reddien 2010). If the wound requires the replacement of missing tissue, a second peak of NB proliferation occurs at 48 hours concentrated at the wound site and their progeny form an unpigmented bud of regenerated tissue called the blastema. The NB progeny will then differentiate into the missing structures of the animal.

Cell division is an essential criterion for defining NBs. Cells that have passed through S-phase or are in mitosis can be easily experimentally labelled and visualized either with the thymidine analogue bromodeoxyuridine (BrdU), or an antibody to a mitosis-specific histone modification (Newmark and Sánchez Alvarado 2000). Some studies also claim the use of RNA probes to genes that are specifically present during S phase (e.g., *pcna*) (Orii et al. 2005). Moreover, as with all cycling cells (Becker et al. 1963; Till and McCulloch 1961), irradiation sensitivity can also be used as a feature to distinguish cycling NBs from differentiated post-mitotic cells, and doses over 30Gy ablate all dividing cells within 24 hours of exposure (Eisenhoffer et al. 2008). The ablation of the NB compartment leads to a loss of regenerative ability, failure of homeostatic tissue maintenance, and eventual death. Bulk transplantation of un-irradiated NBs from a healthy donor worm to an NB-ablated irradiated worm rescues tissue maintenance and regenerative ability. Remarkably, the transplantation of a single healthy NB into an irradiated host can result, albeit with low efficiency, in the reconstitution of the entire NB population, and complete restoration of homeostatic tissue maintenance and regenerative ability (Wagner et al. 2011). Consequently, these lines of evidence

serve as proof that the NB population as a whole is pluripotent, and that at least some of these individual NBs are truly pluripotent stem cells.

Non-cell-cycle related features can also distinguish the NB population. NBs are small in size (7-12 μm in diameter), have a large nucleus and scant cytoplasm, and reside in the planarian parenchyma. NBs are scattered broadly throughout the parenchyma, being absent only from the anterior tip and the pharynx – the only two regions that cannot support regeneration in isolation. Moreover, NBs have large cytoplasmic perinuclear ribonucleoprotein (RNP) granules called chromatoid bodies (CBs) that resemble the germ granules, or 'nuage', found in the germline stem cells of other metazoans (Coward 1974; Morita and Best 1984; Morita et al. 1969).

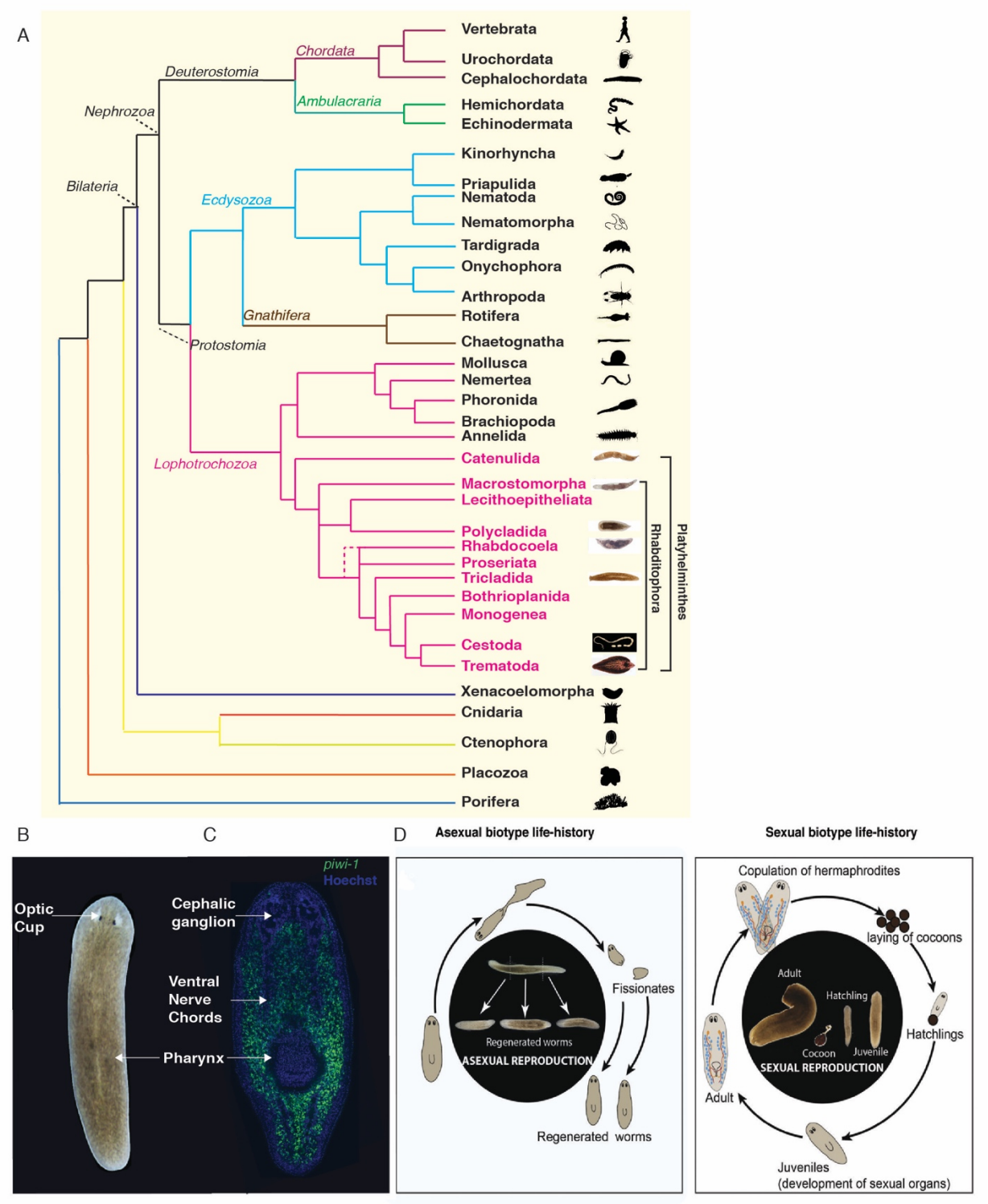


Figure 1: **A.** Current consensus phylogenetic relationships of major animal phyla. The 11 Platyhelminth orders are highlighted and consist of two clades: Catenulida and Rhabditophora (according to Egger et al. (2015)). Flatworm pictures taken from Duran and Egger (2012). **B.** Brightfield image of the triclad flatworm *Schmidtea mediterranea*, commonly used as a regeneration model system. **C**. *In situ* hybridisation against *piwi-1*, a canonical marker for neoblasts (NBs) and nuclear staining with Hoechst. **D.** Asexual and sexual life-histories of *S. mediterranea*. In the asexual biotype, worms replicate by clonal transverse fission. In the sexual biotype, worms are cross-fertilizing hermaphrodites and lay eggs. These eggs hatch, and germline development occurs in the juvenile to form an adult with ovaries and testes. Figure 1D taken from Sahu et al. (2017).

1.4 Methodologies for identifying genes associated with NBs

The irradiation sensitivity of NBs has made it possible to determine NB-specific genes, through observation of the expression patterns of individual genes (assayed through RNA-seq and *in situ* hybridisations) both prior to, and following NB ablation. One caveat of this approach is that irradiation likely induces secondary transcriptome-wide changes in gene expression that are not simply as a result of NB removal. Consequently, RNAi of the NB-specific histone isoform *histone-2B* (*H2B*) has been used in order to genetically ablate NBs (Solana et al. 2012). The transcriptomic profiles of wild-type worms were then compared to that of *H2B* RNAi and lethally irradiated worms to ascertain genes specific to the NBs whilst also circumventing the possibility of obtaining false positives introduced by the upregulation of DNA damage response (DDR) associated gene repertoire.

Another commonly used methodology for isolating NBs is Fluorescence Activated Cell Sorting (FACs) (Hayashi et al. 2006; Romero et al. 2012). FACs analysis of animals stained with Hoechst enables the isolation of two cell populations that are lost in irradiated animals: the 'X1' gate, which represents cells in the cell cycle with >2C DNA; and the 'X2' gate, which registers as a <=2C DNA population due to Hoechst efflux. The remaining irradiation-insensitive (Xins) cells are assumed to be post-mitotic differentiated cell types possessing 2C DNA. Consequently, RNA-seq has been performed on the X1 compartment to successfully identify genes specific to the NB compartment and which are not transcriptionally active in the post-mitotic Xins category (Önal et al. 2012; Labbé et al. 2012).

1.5 RNA-binding proteins are highly enriched in planarian neoblasts

Many genes are specifically expressed in cells that share NB features (irradiation sensitivity, morphology, division, localization), and those encoding RNA-binding proteins (RBPs) are

significantly over-represented in this repertoire (Önal et al. 2012; Labbé et al. 2012; Solana et al. 2012). Significantly, RBPs are known to be expressed in both germline stem cells and multipotent/pluripotent stem cells of all animals and are key constituents of the conserved Germline Multipotency Program (GMP) (Juliano et al. 2010; Solana 2013; Lai and Aboobaker 2018). Evidence for this conservation comes from a reconstruction of the ancestral stem cell gene repertoire by comparison of the transcriptomes of totipotent archeocytes from the demosponge *Ephydatia fluviatilis,* NBs from *Schmidtea mediterranea*, and multipotent interstitial cells (I-cells) from *Hydra vulgaris* (Alié et al. 2015). Within the reconstructed set of 180 conserved genes, transcription factors (TFs) were relatively underrepresented (3/180) compared to RBPs (44/180) (Alié et al. 2015). One explanation for the preponderance of RBPs is that the post-transcriptional regulation of transcripts involved in the maintenance of multi/pluripotency and differentiation allows for the ability to rapidly transition from a stem cell to differentiated state, without the need for time-consuming genetic cascades. More broadly, the presence of GMP and nuage components such as RBPs in planarian NBs has prompted a revision of their classification from 'somatic stem cells' to 'Primordial stem cells' (PriSCs) that are a part of the germline cycle, and which can drive both asexual reproduction by fission as well as the regeneration of the sexual germline (Solana 2013).

A number of planarian studies have established functional roles for RPBs in NBs. For example, homologs of classical germline associated RBPs such as *Vasa* (Wagner et al. 2012)*, Pumilio* (Salvetti 2005)*, Tudor* (Solana et al. 2009)*, Piwi* (Reddien et al. 2005b; Palakodeti et al. 2008)*, Bruno* (Guo et al. 2006), are all expressed in NBs and knockdown by RNAi of many of these genes results in an abrogation of regeneration and affects the stem cell compartment to varying degrees. Indeed, the RBP *piwi-1* (or *smedwi-1* when referred to in the species of *S. mediterranea*) is used a common NB *in situ* hybridisation marker in many planarian studies. Moreover, RBPs have also been shown to play an essential role in the regulation of differentiation. For example, mRNA deadenylation is a precursor to degradation, and in eukaryotes this is carried out by the CCR4-NOT complex. RNAi of the gene encoding the largest subunit of this complex, *not-1*, resulted in NBs failing to differentiate owing to an accumulation of stem cell-related mRNAs (Solana et al. 2013). Although not a RBP

itself, NOT1 as part of the CCR4-Not complex, may be recruited to particular mRNAs by specific RBPs whose identity is unresolved in planarians. For instance, it has been suggested that the RBP *mex-3-1* which, when knocked down gives a very similar differentiation-related phenotype as *not-1,* may in fact tether its RNA targets to the CCR4-Not complex (Zhu et al. 2015; Krishna et al. 2019). This would uncover the mechanism by which *mex3* enables translational repression of target loci in other model systems (Pereira et al. 2013). Other RBPs such as the homologs to *Pumilio* and *Tristetraprolin* may also play a role in interacting with planarian CCR4-NOT, similar to their function in other organisms (Wahle and Winkler 2013; Webster et al. 2019) .

Although RBPs are enriched in planarian NBs and, more broadly, play an ancestral role in animal stem cells, it is likely that there are a number of clade-specific TFs that are also responsible in maintaining multi/pluripotency and self-renewal. Of the five genes that have been shown to induce pluripotency in mammalian somatic cells (*Myc, Nanog, Klf4, Oct4* and *Sox2*) (Takahashi and Yamanaka 2006), homologues to *Nanog* are clearly lacking in the flatworm genomes of *S. mediterranea and Macrostomum lignano* as well as cnidarians such as *Hydra magnipapillata* and *Nematostella vectensis*, and definitive homologues to the other 4 genes are not apparent in these organisms (Önal et al. 2012; Wasik et al. 2015; Hemmrich et al. 2012; Putnam et al. 2007; Chapman et al. 2010; Steele et al. 2011). For instance, the NBs of *S. mediterranea* are enriched for the expression of two Sox/HMG box TF family members (*soxP-1* and *soxP-2),* and may be analogous in function to mammalian *Sox2* (Önal et al. 2012; Van Wolfswinkel et al. 2014; Wagner et al. 2012). Likewise, 4 out of 11 Sox genes in the hydrozoan *Clytia hemisphaerica* were shown to be expressed in multipotent I-cells (Jager et al. 2011), and a Myc gene in *Hydra* is necessary for the balance of stem cell self-renewal and differentiation of multipotent I-cells (Ambrosone et al. 2012; Gold and Jacobs 2013). Other TF families may play a role in the regulation of multi/pluripotency networks: for instance, in *Hydra*, a group of 29 zinc-finger (ZNF) genes have been shown to be enriched in multipotent I-cells (Hemmrich et al. 2012). Such genes may represent clade-specific TFs responsible for the maintenance of stem cell populations in invertebrates, and their molecular functions are yet to be uncovered in the stem cell or different organisms.

1.5 Neoblast heterogeneity: individually pluripotent cNeoblasts and lineage-committed progenitors

The prevailing 'NB specialisation' model suggests that NBs are a heterogeneous population containing both individually pluripotent stem cells, as well as NB subtypes that express distinct tissue-specific transcription factors (Reddien 2013). As such, during regeneration, blastema cells would have their fate pre-determined by the specialised NB classes, which are in turn products of individually pluripotent NBs (called clonogenic or cNeoblasts) (Wagner et al. 2011).

To test the hypothesis that the fate of regenerating cells is specified in the NBs themselves, FACS was used to separate X1 NBs from the pre-pharyngeal area of animals 48 hours following head or trunk amputation (Scimone et al. 2014). RNA-sequencing analysis identified 33 transcription factors (TFs) expressed in sorted X1 NBs that are *piwi-1*+ around the wound site, which when knocked down by RNAi ablated the regeneration of specific tissues. For example, *FoxA* was identified as being important for pharynx regeneration, *pax3/7* for the differentiation of dopamine-B-hydroxylase expressing neurons, and *klf* for *cintillo*-expressing sensory neurons. Moreover, TFs expressed together in the same differentiated tissues (e.g. the pharyngeal markers *FoxA* and *meis*) were also expressed in the same isolated X1 NBs (Scimone et al. 2014).

Intriguingly, these specialised NB subtypes seem to be distributed broadly throughout the homeostatic animal, in regions that are distant from their eventual differentiated site. For example, specialized stem cells exhibiting *FoxA* are present throughout the trunk of planarians, but play a role in pharynx development (Adler et al. 2014; Scimone et al. 2014). Similarly, eye-specialized NBs are present in the anterior trunk region between the eyes and pharynx. These progenitors then divide and are incorporated into the eye during homeostatic turnover. Moreover, during regeneration, a trail of non-dividing eye progenitor cells (or 'trail cells') emerge from the wound area as they migrate to their target sites (Lapan and Reddien 2011, 2012). These lines of evidence suggest that lineage-

committed NBs can be located at considerable distances away from their target sites, and can effectively 'home in' to replace lost or damaged tissues.

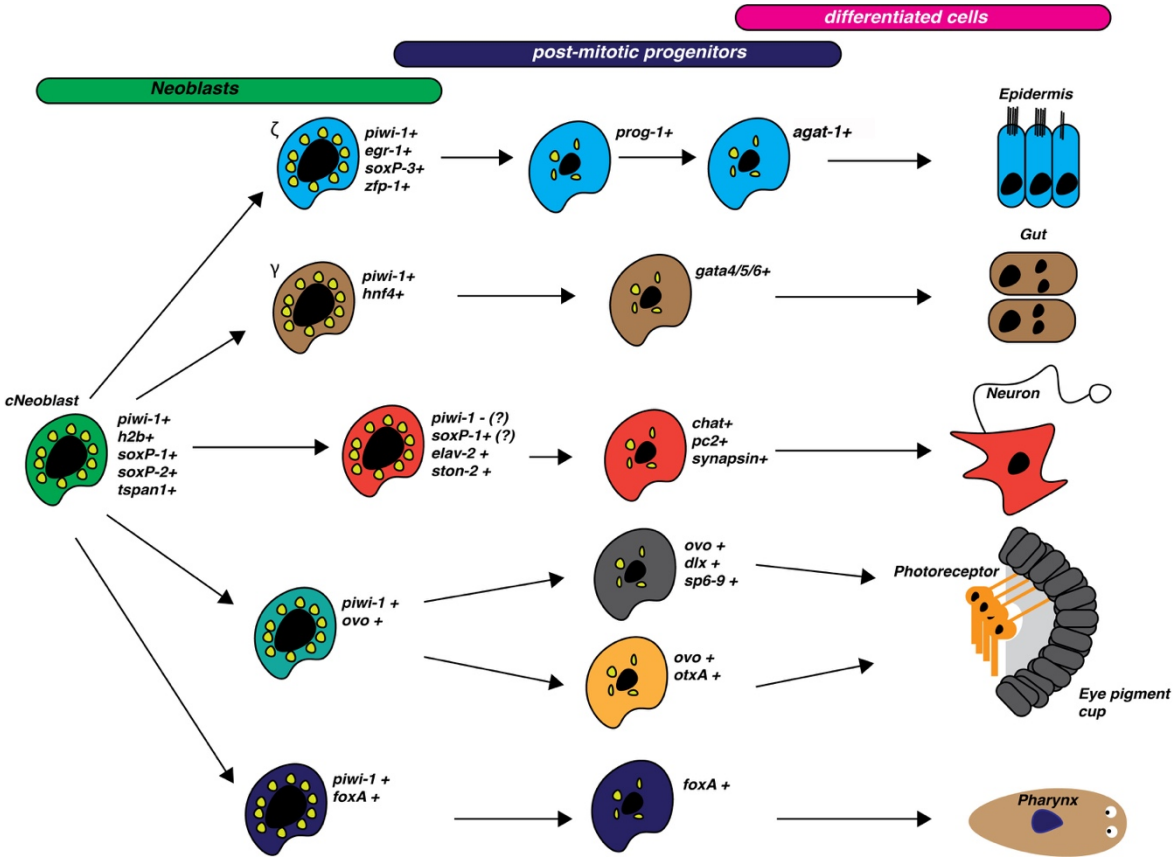
Single-cell RNA quantification methodologies have also revealed extensive functional heterogeneity in the NB population. Initially, systematic single-cell multiplexed qPCR analyses separated NBs into two major categories of stem cells: (1) the zeta (ζ) NBs, distinguished by the expression of *soxP-3, egr-1*, *zfp-1*, *fgfr-1*, are the pre-cursor stem cells for the epidermis, and (2) the sigma (σ) class of NBs, distinguished by *soxP-1*, *soxP-2* expression. A sub-class of σ NBs was also identified as the gamma (γ) NBs, that express *gata4/5/6*, *nkx2.2-like*, *hnf4*, and *prox-1*, and which serve as progenitors for the intestine (Van Wolfswinkel et al. 2014) **(Figure 2)**. Importantly, the σ NBs were shown to be able to give rise to the ζ NBs by using stem cell grafts from *zfp-1* RNAi worms into irradiated hosts, and subsequently observing the generation of *zfp-1*+ cells from σ NB precursors. Consequently, it has been hypothesized that the σ NBs harbour individually pluripotent cNeoblasts that are capable reconstituting all cellular lineages following single-cell transplantation (Van Wolfswinkel et al. 2014). Importantly, the functions of genes shown to be enriched in the σ-class compared to γ NBs, in particular TFs, are still relatively understudied (e.g. *soxP-1, soxP-2, pbx-1, smad6-7)* and may be essential for NB pluripotency, analogous to mammalian ESC 'Yamanaka' factors.

Single-cell RNA-seq (scRNA-seq) of X1 stem cells in the head region revealed an additional NB class, dubbed the nu (ν) NBs, that exhibit transcripts for defining the neuronal lineage (Molinaro and Pearson 2016). These ν NBs, however, are unusual in that unlike the ζ NBs or γ NBs, the level of *piwi-1* transcript in these cells is very low or absent, but the levels *piwi-2* (another stem cell marker) remain comparable with other NBs. As very few cells in this study were sampled for sequencing (96 X1 and 72 X2 cells), the low *piwi-1* expression in these 17/96 X1 ν NBs is most likely an artefact in library-making procedures (Dattani A, Evans D, Aboobaker in prep). Nevertheless, in theory ν NBs may exist as a population of late neural committed stem cells with low *piwi-1* that will transition to neural post-mitotic progenitors following one round of cell division.

More recently, Drop-Seq has allowed for the generation of transcriptomes from thousands of individual cells from whole worms. In addition to identifying previously unknown tissue types, many more specialized NB subsets were uncovered, in addition to ζ and γ classes, and included precursors to protonephridia, muscle, neurons, parenchymal cells, and other differentiated cell types (Fincher et al. 2018; Plass et al. 2018). Moreover, the results from both studies indicate that the molecular markers for σ NBs are also expressed in other lineage-committed progenitors, such as neural progenitors, and expression of markers such as *soxP-1* are not necessarily unique to cNeoblasts, but may only enrich for them (see Chapter V) (Plass et al. 2018; Fincher et al. 2018).

In order to identify the NB population enriched for cNeoblasts, scRNA-seq has been used on X1 cells, to reveal 12 transcriptionally distinct NB clusters. One of these clusters did not exhibit markers for specific lineages, and unlike the other 11 clusters expressed high-levels of a conserved cell-surface marker protein-coding gene, *tetraspanin-1* (*tspan-1*) (Zeng et al. 2018). An antibody against TSPAN-1 protein allowed for FACS isolation of this specific NB population, and single-cell transplantation into irradiated hosts improved recovery efficiency compared to single cell transplantation from a pool of isolated bulk X1 cells (Zeng et al. 2018). However, it remains to be tested whether the recovery efficiency using cells sorted with TSPAN-1 also improves when compared to the transplantation of cells from the other 11 X1 cell clusters, which may or may not also contain truly pluripotent cNeoblasts. Nevertheless, this result does suggest that the TSPAN-1 NB population is enriched for cNeoblasts, and the ability to prospectively isolate this population provides exciting opportunities for molecular analysis, and potential genetic manipulation of individually pluripotent cells.

Figure 2: Lineage markers involved in the differentiation of stem cells to different terminal tissue types. Note for the neuronal lineage we have questioned the absence of *piwi-1* in neural-committed ν NBs (Molinaro and Pearson 2016), but have included it as a late NB population. Moreover, we have also included the potential expression of *soxP-1* in this population as observed by recent single-cell RNA-seq studies (Fincher et al. 2018; Plass et al. 2018).



1.6 Conclusion

It is clear that in the last 20 years, a lot has been uncovered regarding the cellular and genetic basis for planarian NB maintenance and differentiation, which in turn has allowed us to better understand the regenerative process. Given the notable absence of transgenic manipulation of NBs, many ultimate questions will be difficult to address utilising novel RNA-seq approaches and RNAi alone. However, the dissection of epigenetic mechanisms controlling NB pluripotency and fate choice remain relatively explored, and much can still be done to investigate underlying processes without the need for genetic manipulation. In Chapter II we present the case for using planarians as a model system for stem cell epigenetics research, and argue that this will allow for the discovery of both evolutionarily conserved and novel epigenetic processes in planarian stem cells compared to conventional mammalian model systems.

Chapter II

Planarian flatworms as a new model system for understanding the epigenetic regulation of stem cell pluripotency and differentiation

Chapter II has been reproduced as a publication as *"Planarian flatworms as a new model system for understanding the epigenetic regulation of stem cell pluripotency and differentiation"* in the journal *Seminars in Cell and Developmental Biology*. The reproduction of text and figures is in line with Copyright terms documented on the journal website: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights. *Author contributions:* Anish Dattani and Aziz Aboobaker conceived and wrote manuscript; Divya Sridhar contributed to discussions.

Abstract

Transcriptomic studies in planarian flatworms have revealed that gene expression is precisely coordinated to maintain neoblast pluripotency and ensure correct lineage specification during differentiation. However, as yet, these studies have not revealed how this regulation of expression is controlled. In this chapter, we propose that planarians represent an effective system to study the epigenetic regulation of these processes in an *in vivo* context. We consolidate evidence suggesting that although DNA methylation may be present in some flatworm lineages, it does not regulate neoblast function in *Schmidtea mediterranea*. A number of phenotypic studies have documented the role of histone modification and nucleosome remodelling complexes in regulating distinct neoblast processes and we focus on four of these epigenetic regulators: the Nucleosome Remodelling Deacetylase Complex (NuRD complex), Polycomb Repressive Complex (PRC), SET1/MLL methyltransferases, and the nuclear PIWI/piRNA complex. Given the advent of ChIP-seq and planarian-tailored analyses methodologies, we propose future avenues of research that will identify the genomic targets of these complexes allowing for a clearer picture of how NB processes are coordinated at the epigenetic level. These insights may ultimately be relevant to mammalian biology and disease. Moreover, the unique biology of planarians, compared with other conventional invertebrate model systems, will also allow us to investigate how extracellular signals feed into epigenetic regulatory networks to govern concerted NB responses during regenerative polarity, tissue patterning, and remodelling.

2.1 Introduction

During development in mammals, the precise coordination of transcription ensures the transition of early embryonic pluripotent stem cells and their progeny into a vast array of cell types that constitute the entire adult organism. The control and maintenance of gene expression during ongoing lineage commitment and differentiation is dependent on pervasive epigenetic control – the heritable modulation of gene activity that is independent of the underlying DNA sequence. Epigenetic modifications can be broadly classified into 3 groups: (1) DNA methylation, (2) histone modification and nucleosome positioning, (3) small RNA mediated transgenerational inheritance (not discussed further in this paper, but reviewed in (Rechavi and Lev 2017; Anava et al. 2014)).

The methylation of cytosine at the C5 position in CpG di-nucleotide islands, usually located upstream of gene promoters, acts as a beacon to attract the repressive epigenetic modifying complexes in order to robustly block the transcriptional machinery from accessing these sites altogether (Bird 2002). The DNA methylation process has been mainly studied in vertebrates and plants owing to a ubiquitous presence in these taxa. DNA methylation has a patchier distribution among invertebrates; for example, it is absent altogether in *Caenorhabditis elegans* and still contentious in *Drosophila melanogaster* (Dunwell and Pfeifer 2014; Tweedie et al. 1999; Lyko et al. 2000) . Relatively little is known about how DNA methylation is involved in regulating gene expression across the breadth of the Animal Kingdom or when this has evolved.

The presence of covalent modifications on the four nucleosome core histone proteins (H2A, H2B, H3, and H4) around which DNA is wrapped is, in contrast to DNA methylation, ubiquitous across eukaryotes (Luger et al. 1997). These nucleosomes are organised in a higher order chromatin structure through Histone 1 (H1) linker proteins between nucleosomes, and each genetic locus of expression can have its own unique structure and set of histone modifications that can vary between cell states and types, optimised to transcriptional need. Modifications to histone proteins include acetylation, methylation, phosphorylation, and ubiquitination. These either directly affect the

chromatin structure by altering DNA/nucleosome interactions or by attracting effector complexes that contain modification-specific binding domains. These complexes act on chromatin in various ways to influence whether genes are upregulated, downregulated, or silenced (Strahl and Allis 2000; Kouzarides 2007; Bhaumik et al. 2007).

One general feature of the epigenome, which can be exploited experimentally by DNA sequencing library technologies, is that 'open' chromatin correlates with actively transcribed genes and active enhancers. Conversely, 'closed' chromatin, or heterochromatin, tends to correlate with gene silencing. Consequently, chromatin configuration can be seen as the interface between transcriptional and epigenetic regulation. Thus, an important role of some histone modifications is to recruit chromatin remodelling complexes that reposition nucleosomes, expel nucleosomes entirely, or exchange histone variants, thereby affecting the accessibility of the transcriptional machinery to gene promoters and regulatory enhancers involved in providing temporal and spatial specificity. The location of these histone modifications can be ascertained by the use of Chromatin Immunoprecipitation followed by sequencing (ChIP-seq), whereby chromatin is harvested from cells and antibodies specific to particular conserved histone marks are used to enrich for DNA bound by these marks. Following sequencing of these DNA fragments, the genomic location of particular histone marks can be elucidated for specific cell types. Easier, albeit cruder, methodologies for ascertaining the location of open chromatin include DNase-seq and ATAC-seq, which use enzymatic digestion and Tn5 transposition respectively of bulk DNA followed by sequencing to identify hypersensitive sites that correspond to open chromatin. Compared with ChIP-seq, ATAC-seq/DNase-seq require less starting material and do not need an *a priori* knowledge of histone marks.

The bulk of our knowledge on the epigenetic regulation of stem cell pluripotency and differentiation comes from work using Embryonic Stem Cells (ESCs). ESCs are derived from the inner cell mass (ICM) of the blastocyst stage embryo and can produce progenitors that contribute to any type of adult tissue. In culture, ESCs retain an indefinite capacity for self-renewal when differentiation is inhibited by a variety of media conditions that mimic aspects of the microenvironment of the ICM.

These conditions include culture with: (1) cytokine leukaemia inhibitory factor in the presence of serum (serum/LIF) (Smith et al. 1988; Williams et al. 1988), (2) serum-free medium with 2 small molecule inhibitors (2i/LIF) (Ying et al. 2008), (3) or with knockout serum replacement (KOSR/LIF) (Martin Gonzalez et al. 2016). Different culture conditions give rise to different ESCs: 2i/LIF-grown ESCs are unrestricted and highly plastic and reflective of the early blastocyst, whereas serum-grown ESCs are more heterogeneous and restricted in their cell potential, reflective of the late blastocyst (Marks et al. 2012; Martin Gonzalez et al. 2016). However, it has not been demonstrated whether the transcriptomic and epigenetic states of 2i/LIF cultured ESCs are stable over long-term culture. A recent study showed that 2i/LIF ESCs lose DNA methylation at imprinted loci, which leads to an impaired developmental potential and karyotypic abnormality (Yagi et al. 2017a, 2017b).

Induced pluripotent stem cells (iPSCs) represent another burgeoning stem cell study system, and are cells that have been reprogrammed to a pluripotent state from somatic cells usually by the over expression of Yamanaka transcription factors (Takahashi and Yamanaka 2006). This leads to broad epigenetic changes that are heterogeneous between reprogramming events, and often somatic epigenetic remodelling is incomplete (Hawkins et al. 2010; Bar-Nur et al. 2011). Both ESCs and iPSCs represent *in vitro* study systems and it is likely that significant regulatory differences exist between these systems and stem cells in their *in vivo* contexts. Given the importance of understanding pluripotency for biomedical research, it is very surprising that other animal models where pluripotent somatic cells are present are still relatively poorly supported. These models are simpler and more accessible, and can allow for the study of pluripotent cells *in vivo*. At the very least it is likely that a comparative study of epigenetic control mechanisms will be broadly informative, regardless of whether mechanisms are conserved or divergent.

In this review chapter, we propose that the NBs of *S. mediterranea* can be used to study the epigenetic regulation of stem cells in an *in vivo* context, thereby representing a useful non-mammalian system. Early experiments in this model system have shown that knockdown of orthologues of mammalian epigenetic regulators by RNA interference (RNAi) can lead to different stem cell defects and errors

in lineage commitment of stem cell progeny, culminating in a loss of regenerative capacity. The recent advent of ChIP-seq in planarians will allow for the investigation of these defects in greater detail (Duncan et al. 2015; Dattani et al. 2018; Mihaylova et al. 2018), enable assessment of the conservation of epigenetic programs, and potentially identify important functions and targets of epigenetic complexes that may have been either overlooked or difficult to study in mammalian ESC or iPSC culture based systems. In this review chapter, we consolidate the existing planarian studies of epigenetic regulators. We describe and synthesize our understanding of the phenotypic defects of RNAi of genes involved in epigenetic complexes and propose avenues of exploration to understand how planarian NBs respond to extracellular signals to coordinate differentiation under homeostatic and regenerative conditions.

The case for Planarian flatworms as an *in vivo* model system to study stem cell epigenetics

2.2 Transcriptional profiling of planarian somatic neoblasts reveals similarity with ESCs

Planarians represent a promising system for characterizing the epigenome of stem cells. Importantly, there is some evidence to suggest that planarian NBs have a transcriptional pluripotency program that is conserved with mammalian ESCs as well as the pluri- and multipotent adult stem cells and ESCs of other animals. Independent studies and methods have uncovered genes with enriched expression in NBs, and which also have homologs expressed in ESCs that are involved in the balance between self-renewal and differentiation. These include regulators and targets of Oct4, RNA splicing factors, epigenetic modifiers, and RNA binding proteins (Önal et al. 2012; Solana et al. 2012; Labbé et al. 2012). The planarian NB transcriptome broadly also broadly reflects that of the multipotent stem cells of Hydra and totipotent archeocytes of the demosponge *Ephydatia fluviatilis,* together suggesting the existence of an ancestral stem cell expression repertoire, rich in RNA regulatory factors and poor in transcription factors (Alié et al. 2015; Solana 2013). While more work is required to assess the extent and nature of this conservation, these early findings lend credence to the use of

planarian NBs as a model system for stem cell epigenetic studies, owing to the fact that discoveries in planarians may be directly relevant to mammalian ESC biology.

2.3 Planarians provide the opportunity to study the role of epigenetic mechanisms controlling stem cells in vivo

One main advantage planarians have over ESCs is that they represent an *in vivo* system whereby the epigenetic response to extracellular signals can be explicitly tested. Planarian studies have identified many key signalling pathways that regulate regenerative polarity (Blassberg et al. 2013; Iglesias et al. 2008; Petersen and Reddien 2008, 2009; Yazawa et al. 2009; Rink et al. 2009; Gaviño and Reddien 2011; Scimone et al. 2016; Lander and Petersen 2016), found where these signals originate from in the animal (Witchley et al. 2013; Wurtzel et al. 2015; Scimone et al. 2016, 2017), and in some cases the transcriptional changes the signals control in NBs responding correctly to injury (Kao et al. 2013; Wurtzel et al. 2015; Scimone et al. 2014). We have also made progress in understanding how the dynamic process of tissue homeostasis in planarians is controlled (Reuter et al. 2015). So far, however, we do not know the role of epigenetic mechanisms in regulating regenerative polarity, tissue patterning, or tissue homeostasis because the limited studies that do exist to date have investigated epigenetics only in the context of stem cell maintenance and differentiation (Hubert et al. 2013; Duncan et al. 2015; Mihaylova et al. 2018; Dattani et al. 2018). However, with the application of epigenomic techniques, such as ChIP-seq and, in the future, ATAC-seq, these processes can be studied in the context of the whole regenerative response.

We can, for instance, ask how the epigenome of NBs responds to different signals and conditions and how this then impacts on changes in gene expression that allow the regeneration of the correct structures. For example, we know that NBs respond very differently to anterior and posterior facing wounds that produce different positional signals (Owlarn and Bartscherer 2016), and respond dynamically to starvation conditions resulting in regulated de-growth (González-Estévez et al. 2012; Mangel et al. 2016). These represent fundamental *in vivo* stem cell responses for which epigenetic

responses are not yet described. By combining specific regenerative or environmental scenarios with RNAi of key signalling pathways (e.g. Wnt (De Robertis 2010; Petersen and Reddien 2009; Scimone et al. 2016; Yazawa et al. 2009), Hh (Rink et al. 2009; Yazawa et al. 2009), TOR (Tu et al. 2012), JNK (Almuedo-Castillo et al. 2014; Tejada-Romero et al. 2015)) it should be possible to assess the importance of epigenetic mechanisms in the NB response to these conditions and control signal. While all current examples of using ChIP-seq in planarians have been in the context of studying the role of enzymes that mediate specific histone marks (Duncan et al. 2015; Mihaylova et al. 2018), these studies have also established the basis for a much broader investigation in the plethora of exciting experimental paradigms offered by planarians.

Loss of DNA methylation in invertebrates and flatworm lineages

2.4 The patchy distribution of DNA methylation in invertebrates

DNA methylation, the transfer of a methyl group to the cytosine ring of DNA (5mC), typically occurs in the context of CpG dinucleotides, and is responsible for the silencing of the underlying DNA segment. DNA methylation-based silencing is responsible for both long-term transposable element (TE) suppression, preventing these selfish genetic elements from disrupting genomic integrity, and genomic imprinting, which allows for the monoallelic expression of a subset genes dependent on parental origin (Law and Jacobsen 2010). DNMT1 and DNMT3 are the two generally accepted families of DNA methyltransferases ancestral to both animals and plants (Jurkowski and Jeltsch 2011; Zemach and Zilberman 2010). DNMT2 has been dismissed as a misnomer because it has an exclusive role in tRNA methylation (Tuorto et al. 2015; Goll et al. 2006), although earlier studies suggested it had a role in retroelement silencing (Phalke et al. 2009; Schaefer and Lyko 2010). 5mC is a substrate for methyl-binding domain containing proteins (MBDs) that attract nucleosome remodelling and histone modification complexes to the DNA segment. The ancestral group of MBD genes in animals is MBD2/3 (or less commonly, but correctly referred to as MBD1/2/3) and MBD4/MeCP2 which following two rounds of whole genome duplication (2R) resulted in the

paralogs MBD1, MBD2, MBD3, MBD4 and MeCP2 in vertebrates (Albalat 2008; Albalat et al. 2012).

The overall pattern of DNA methylation is variable between organisms. Vertebrates, in particular mammalian genomes, are heavily methylated at most CpG sites (i.e. global DNA methylation) which correlates with transcriptional silencing. Only a small number of CpGs are lowly methylated, localized to short genomic regions, and usually located in proximity to annotated gene promoters and enhancers (Schultz et al. 2015; Mendizabal and Yi 2016). These hypo-methylated regions are typically characterized by a high GC content and are referred to as CpG islands (Deaton and Bird 2011; Bird et al. 1985). Conversely, invertebrate genomes tend to be sparsely methylated and in many cases DNA methylation and the machinery for producing this mark are absent entirely **(Figure 1)** (Tweedie et al. 1997). Loss of methylation has happened independently within many groups, including nematodes, arthropods, and flatworms, which have members with and without DNA methylation and/or DNA methyltransferases. Although the ultimate reason as to why many invertebrates have lost DNA methylation is unclear (Zemach et al. 2010; Zemach and Zilberman 2010), the proximate cause is most likely associated with the mutagenic load associated with DNA methylation (such as GT mis-match from 5mC to thymine deamination (Duncan and Miller 1980; Britten et al. 1988; Sved and Bird 1990; Bulmer 1986) and alkylation damage (Rošić et al. 2018).

2.5 Planarians most likely lack endogenous DNA methylation and lack cognate DNA methyltransferases

DNA methylation studies in flatworms have been controversial and contradictory. Given that NB-like cells are a conserved feature of the phylum Platyhelminthes, it is important to know whether DNA methylation plays a role in epigenetic regulation of these cells. Early studies argued that DNA methylation was not present in the Platyhelminthes on the basis of methylation-based restriction endonucleases followed by amplification of restriction fragments (Methylation Sensitive Amplified Polymorphism – MSAP) (Rosado Fantappié et al. 2001). While DNMT1 and DNMT3 are absent in

specific flatworm lineages, including in representatives of the Trematoda, Cestoda, Monogenea, Macrostomorpha and Tricladia, it was argued that having DNMT2 and MBD2/3 orthologues could be indicative of a propensity for DNA methylation in these organisms (Geyer et al. 2013). However DNMT2 is a tRNA methyltransferase with no convincing role in DNA methylation having ever been established (Raddatz et al. 2013; Goll et al. 2006; Schaefer and Lyko 2010). One study with the parasitic flatworm *Schistosoma mansoni* claimed the presence of methylated DNA in precise locations of the genome and a role for cytosine methylation in the regulation of oviposition. This work utilized 5-azacytidine (AzaC) to inhibit DNA methylation in adult mating pairs, but this drug is also known to inhibit RNA methylation by DMNT-2 with high efficiency and as such the phenotypic effects could simply reflect inhibition of this process (Geyer et al. 2011; Schaefer et al. 2009). Indeed, another study utilizing whole-genome bisulfite sequencing showed that the *S. mansoni* genome was not methylated and that incompletely converted cytosines following bisulfite treatment likely accounted for why DNA methylation was found in an earlier study (Raddatz et al. 2013). It therefore seems probable that the *S. mansoni* genome is not endogenously methylated, and perhaps previous positive data actually reflects methylated cytosine scavenged from the host or culture environment (Marsit 2015; Zauri et al. 2015). In contrast to parasitic genomes, the genome of the more basal flatworm *Macrostomum lignano* has been shown to have both DNMT1 and DNMT3 and low levels of DNA methylation (Wasik et al. 2015; Wudarski et al. 2017).

The genome of *S. mediterranea* was definitively shown to lack cytosine-dependent methylation on the combined basis of MSAP, a lack of antibody staining against 5mC, and undetectable levels of 5mC in High Performance Liquid Mass Chromatography coupled Mass Spectrometry (HPLC-MS) (Jaber-Hijazi et al. 2013). Moreover, neither DNMT1 nor DNMT3 have been found in the genome of *S. mediterranea*, and like other closely related Platyhelminthes it does not contain an MBD4/MeCP2 (**Figure 1**). Additionally, the MBD2/3 protein in *S. mediterranea* does not have the highly conserved ARG22 involved in forming hydrogen bonds with guanine in methylated CpG islands (Zou et al. 2012; Ohki et al. 2001; Jaber-Hijazi et al. 2013). As a consequence of these different lines of evidence we can suggest that the function of *mbd2/3* is independent of DNA

methylation and that DNA methylation is not involved in the epigenetic control of stem cells in *S. mediterranea*. However, given the presence of a complete set of DNA methylation machinery in *Macrostomum lignano,* it is possible that DNA methylation is present and may be involved in the epigenetic regulation of stem cells in this basal flatworm species. Further studies, in *M. lignano*, another pertinent regenerative model (Wudarski et al. 2017; Wasik et al. 2015), will address this and help to understand whether DNA methylation is a basal characteristic of flatworms.

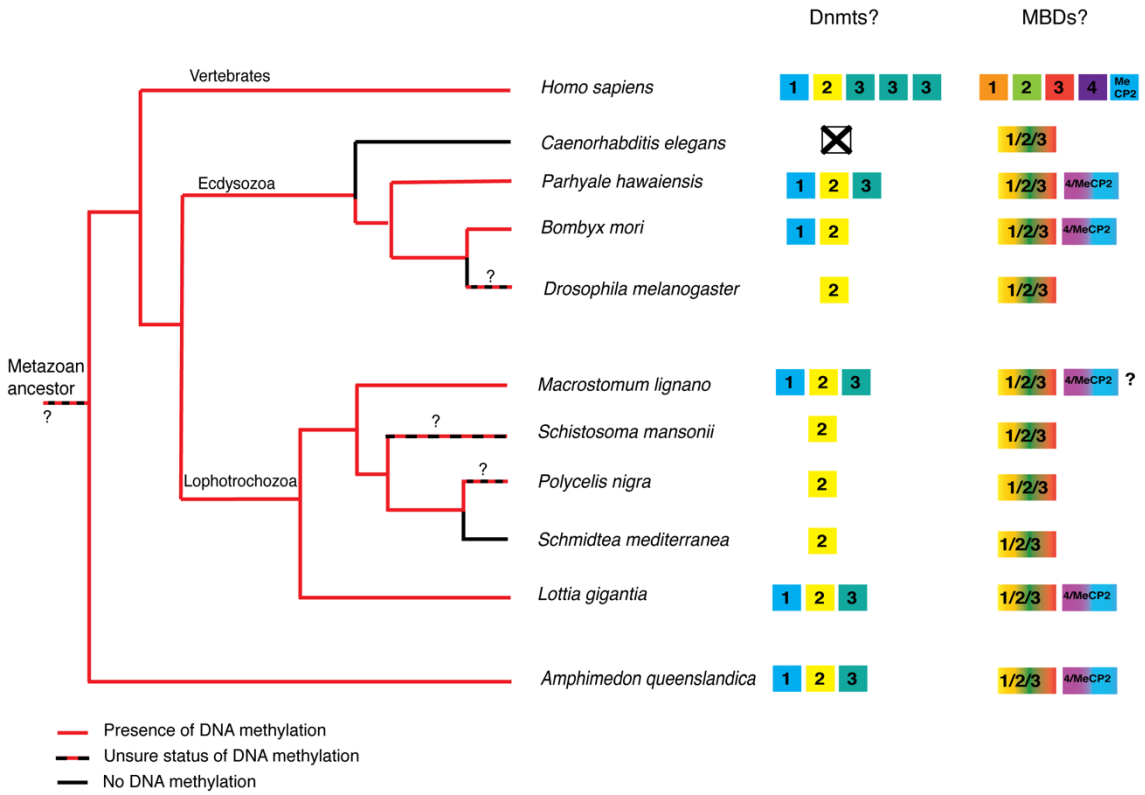


Figure 1. Presence/absence of DNA methylation in flatworms and other species. The methylation status of the metazoan ancestor is unknown, but DNA methylation is not present in many invertebrates. The presence of DNMT and MBD genes in metazoan species were confirmed by tBLASTn in NCBI against species genomes/transcriptomes as well as cross-referencing with the findings of Albalat et al. 2012. Although all three DNMT orthologues have been identified in the *Macrostomum lignano* genome (Wasik et al. 2015), we cannot confirm the presence of MBD4/MeCP2. The presence of an MBD1/2/3 orthologue is based on a tBLASTn search using http://www.macgenome.org/, but we also find two additional MBD genes, which may be a *Macrostomum*-specific MBD1/2/3 paralogues.

2.6 Connecting histone modification changes at genes with whole organismal phenotypes

Planarian NB and differentiation defects resulting from the knockdown of histone epigenetic complexes can be effectively assayed with *in situ* hybridization using a growing list of markers specific to different stem cell and differentiated lineages. The following sub-sections review studies that have adopted this approach in studying the role of major histone and nucleosome modifying complexes in planarian stem cell biology and differentiation during both regeneration and homeostasis. Given the recent advent of ChIP-seq on planarian NBs, we offer possibilities for how these phenotypes can be better characterized at a molecular level in order to uncover both unappreciated and conversed functions of epigenetic complexes when compared to their mammalian counterparts.

2.7 The multipronged NuRD complex and the role of MBD2/3 in 5mC-free planarians

Studies from different animals have shown that the nucleosome remodelling and deacetylase (NuRD) complex is essential for embryonic development. The NuRD protein complex is relatively unique in that it has at least three distinct enzymatic activities involved in chromatin directed gene regulation: deacetylation, ATP-dependent chromatin remodelling, and lysine-specific demethylation. A lack of DNA methylation in planarians means that genomic targets and biological effects of conserved histone modifiers and chromatin remodelers in NBs can be studied without consideration of an interplay with DNA methylation. This simplification in comparison to mammals may be particularly useful for studying the methylation independent roles of the NuRD complex.

Firstly, NuRD has histone deacetylase activity through the activities of the HDAC subunit, which is highly conserved and present in all eukaryotes. HDAC activity leads to a loss of the active H3K27ac on specific genes and provides a substrate for PCR2-mediated tri-methylation (H3K27me3) leading

to transcriptional silencing (Reynolds et al. 2012b). NuRD targets in ESCs include developmental genes that are transcriptionally 'poised' or bivalent genes that harbour both active H3K4me3 and repressive H3K27me3 marks (Reynolds et al. 2012b). This suggests that HDAC/NuRD activity regulates the balance between the acetylation and methylation state of H3K27 such that genes can be released for transcription upon ESC differentiation to defined lineages. Other NuRD targets in ESCs, somewhat counterintuitively, include pluripotency associated genes. The activity of histone acetyltransferases (HATs), which promote the transcription of pluripotency genes in ESCs, is dampened (but not lost) by the HDAC activity of NuRD. Once ESCs differentiate, HAT activity diminishes, and pluripotency genes are silenced in differentiating cells (Reynolds et al. 2012a) .

Secondly, NuRD is also implicated in ATP-dependent chromatin remodelling as a result of the mutually-exclusive chromodomain-helicase-DNA-binding paralogous subunits CHD3 (Mi-2α), CHD4 (Mi-2β) and CHD5. These subunits utilize the energy released from the hydrolysis of ATP to ADP to induce nucleosome sliding, which either enables the recruitment of transcriptional complexes or suppresses transcription entirely. Early studies in *Arabidopsis thaliana* (Ogas et al. 1999) and *C. elegans* (Unhavaithaya et al. 2002; von Zelewsky et al. 2000) indicated that the CHD subunits are involved in the silencing of embryonic genes during differentiation. Discoveries in mammalian systems have since shown that CHD, as part of the NuRD complex, also functions in guiding lineage-specific gene programs. For example, in mammals, the CHD3, CHD4, and CHD5 proteins regulate distinct and non-redundant aspects of gene regulation in three distinct stages of cortical differentiation (Nitarska et al. 2016).

NuRD also associates with the lysine-specific histone demethylase 1A (LSD1) to target the removal of active mono and di-methyl moieties from lysine 4 of histone 3 (H3K4) (Wang et al. 2009). NuRD complexes containing LSD1 associate with the promoters of genes involved in cell growth (including TGFβ signaling), survival, migration, and tissue invasion. Indeed, LSD1-NuRD complexes prevented breast cancer invasion *in vitro* and metastases *in vivo*, indicating that the loss of the LSD1-NuRD complex or reduction in activity may predispose to cancer (Wang et al. 2009). One hypothesis

is that LSD-1 targets H3K4me2 removal at promoters leading to gene silencing (Adamo et al. 2011). Moreover, LSD1-NuRD complexes localize to active ESC enhancers to decommission them via removal of the H3K4me1 active mark, resulting in increased differentiation (Whyte et al. 2012; Hu and Wade 2012).

In addition to these three enzymatic subunits, the NuRD complex also associates with two interchangeable methyl-CpG-binding domain (MBD) proteins, MBD2 and MBD3, in vertebrates. MBD2 has the capacity to selectively recognize 5mC, whilst MBD3 has lost the ability to bind to 5mC during vertebrate evolution (Hendrich and Bird 1998; Zhang et al. 1999). Whilst earlier studies suggested that MBD3/NuRD had a role independent of DNA methylation, MBD3 can bind to 5-hydroxymethylcytosine (5hmC) (Mellén et al. 2012; Yildirim et al. 2011) - the first oxidative product in the demethylation of 5mC by the enzyme TET1 (Lu et al. 2015; Tahiliani et al. 2009). A recent study proposed that MBD2/NuRD and MBD3/NuRD bind to the same genomic loci, and suggests a model by which the two MBD proteins are interdependent and form a regulatory loop to reinforce transcriptional silencing **(Figure 2A)** (Hainer et al. 2016): (1) following the conversion of 5mC to 5hmC by TET1, (2) MBD3/NuRD binds to 5hmC loci leading to DNMT1localization (3) enabling conversion of 5hmC back 5mC (4) leading to subsequent binding of MBD2. Occupation of MBD3/NuRD at 5hmC can be disrupted by further TET1 activity leading to CpG demethylation, which can precede transcription activator binding and gene expression. Importantly, MBD proteins function in coordinating crosstalk between DNA methylation and NuRD to produce a suppressive chromatin environment at target loci (Denslow and Wade 2007).

Some reports suggest that MBD2/NuRD and MBD3/NuRD may function independently of CpG methylation in mammalian stem cell systems and have a role in transcriptional activation of genes and enhancers (Baubec et al. 2013; Shimbo et al. 2013; Günther et al. 2013; Menafra and Stunnenberg 2014). However, data from these studies have since been re-analyzed and no evidence for MBD2 and MBD3 methylation-independent functions are supported (Hainer et al. 2016).

Most invertebrates contain an ancestral MBD2/3 gene that following 2R resulted in one MBD2 gene and one or two copies of the MBD3 gene in vertebrates. (Hendrich and Tweedie 2003). Nuclear Magnetic Resonance (NMR) showed that the MBD2/3 protein of the sponge *Ephydatia muelleri,* a basal metazoan, can bind to methylated DNA consistent with the presence of DNA methylation in this species (Cramer et al. 2017). However, the MBD2/3 protein of *Drosophila melanogaster* lacks DNA binding activity, but continues to associate with the NuRD complex (Cramer et al. 2017). Consequently, we can posit that the ancestral MBD2/3 did bind 5mC or 5mhC and this activity has been lost secondarily in some non-methylated invertebrate species. MBD2/3 in these cases may not necessarily require DNA methylation as a genomic reference to recruit the NuRD complex to target loci, and most likely has a DNA-methylation independent role.

Like Drosophila, *S. mediterranea* also has no detectable levels of endogenous cytosine methylation. RNAi of *mbd2/3* resulted in a loss of certain differentiated cell lineages (e.g. epidermis, gut and pharynx) without reducing NB number. Moreover, there was an accumulation of early epidermal NB progeny (*prog-1*+) but a reduction in late progeny (*agat-1*+) (Jaber-Hijazi et al. 2013). Given that *mbd2/3* mRNA is restricted to the stem cell (X1) and stem cell progeny compartments (X2), it is likely that *MBD2/3* protein influences the expression of genes involved in the terminal differentiation program. However, it remains to be addressed whether *mbd2/3* has a role as a part of the planarian NuRD complex. If MBD2/3 has an ancestral role in coordinating NuRD activity independently of methyl-binding, this mechanism may also function in mammals and would help to resolve long-standing disputes over whether MBD can function independently of CpG (Yildirim et al. 2011; Hainer et al. 2016; Baubec et al. 2013; Shimbo et al. 2013). It is possible that *MBD2/3* either directly or indirectly (via binding of an unknown DNA-binding gene) associates with pluripotency related genes and/or differentiation-related genes and recruits NuRD to modulate their transcription in NBs and NB post-mitotic progeny **(Figure 2B)**.

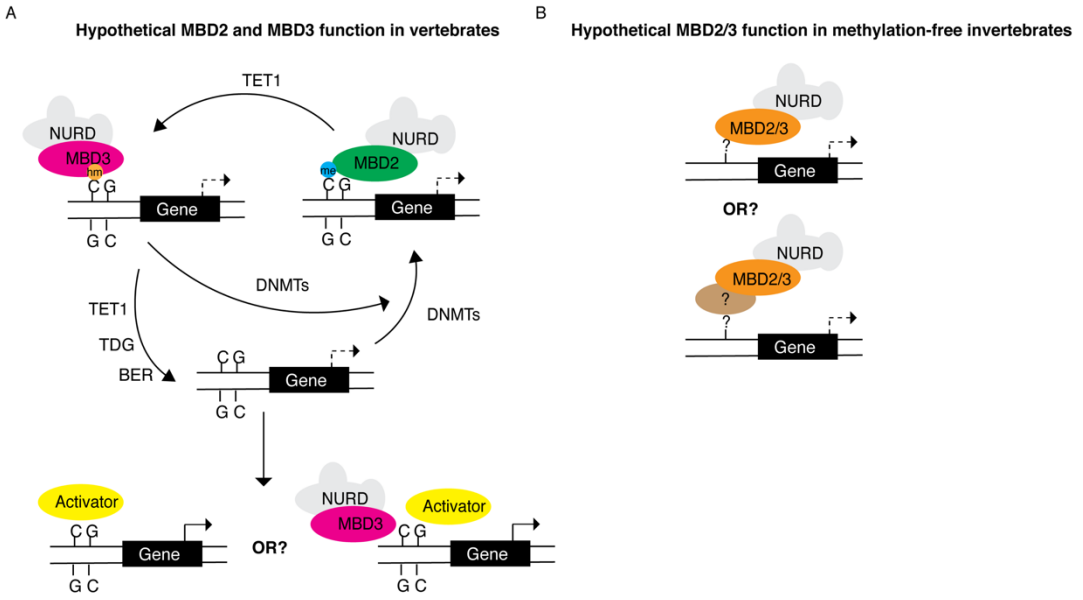


Figure 2: A. Hypothesis as to the function of MBDs in vertebrates adapted from Hainer et al (2016). MBD2 binds to 5 mC at promoters and recruits the NuRD complex to reinforce silencing or dampening of gene transcriptional activity (dashed arrow) of target promoters (e.g., pluripotency and differentiation-related genes). TET1 oxidises the 5methyl group of 5 mC to 5hmC. MBD3 binds to 5hmc with greater affinity than 5mC. DNMT1 localization depends on the presence of 5hmC, and this leads to the restoration of 5 mC at this site. Alternatively, following additional TET1-TDG-BER activity, CpG sites are fully de-methylated. This can lead to activator binding and transcriptional activity of the gene. The presence of MBD3/NuRD at these active de-methylated gene promoters is contentious, but is represented on the diagram. **B.** In methylation-free invertebrates, MBD2/3/NuRD can either bind gene promoters directly or indirectly via a DNA binding protein resulting in gene silencing or reduction in gene transcription.

The functions of four other NuRD complex components have also been investigated by RNAi knockdown in planarians: *CHD4* (Scimone et al. 2010), *HDAC1* (Eisenhoffer et al. 2008; Zhu and Pearson 2013; Robb and Alvarado 2014), the nucleosome interactor RbAp48 (Bonuccelli et al. 2010; Hubert et al. 2015), and the GATA-type zinc-finger domain-containing TF *p66* (Vásquez-Doorman and Petersen 2016). Similar to the *mbd2/3* RNAi phenotype, all of these genes led to an abrogation of stem cell differentiation, but do so in differing manners. For instance, *CHD4* and *HDAC1* RNAi animals lost both early epidermal progeny (*prog-1*+) and late epidermal progeny (*agat-1*+) cells whereas the *mbd2/3* RNAi phenotype led to an accumulation of *prog-1*+ cells but a reduction in *agat-1*+ cells (Zhu and Pearson 2013; Robb and Alvarado 2014; Jaber-Hijazi et al. 2013). This indicates that these genes have distinct functions in the lineage differentiation process in planarians

– *mbd2/3* is required for a later point in differentiation, whereas *CHD4* and *HDAC-1* are involved much earlier. Conversely, RNAi of *p66* had no effect on *prog-1* cell number, but like *mbd2/3* RNAi, led to a decrease in *agat-1*+ cells (Vásquez-Doorman and Petersen 2016). Intriguingly, following *p66* knockdown there was also an increase in photoreceptor neurons (PRNs), but no difference in eye pigment cup cell (PCC) production, indicating that p66 acts to suppress PRN production in wild-type worms (Vásquez-Doorman and Petersen 2016). Moreover, the NB proliferation responses in *HDAC1*, *CHD4* and *RbAp48* knockdown animals are reduced, whereas *mbd2/3* and *p66* RNAi worms had normal levels of proliferation and formed blastemas (Vásquez-Doorman and Petersen 2016; Bonuccelli et al. 2010; Hubert et al. 2013; Robb and Alvarado 2014).

The conflicting RNAi phenotypes for different NuRD components can be explained by most of them having roles in other complexes. For instance, both mammalian RbAp48 and HDAC-1 have been shown to be a member of the Sin3a deacetylase complex which, in a complex with *Nanog*, is involved in the activation of pluripotency factors and suppression of differentiation genes. Moreover, RbAp48 is also an important co-factor in the chromatin assembly factor (CAF-1) complex whose function is to initiate nucleosome assembly by adding histones H3 and H4 onto newly synthesized DNA (Smith and Stillman 1989; Marheineke and Krude 1998). Likewise, there is accumulating evidence that vertebrate CHD4 has multiple functions independent of the NuRD complex (O'Shaughnessy-Kirwan et al. 2015). **(Table 1).**

Table 1: Planarian orthologs of NuRD subunits. tBLASTn searches of human NuRD components were made against the dd_smed_v6 transcriptome and IDs are tabulated. FACS RNA-seq proportions (i.e. X1/X2/Xins) were obtained using the dataset from (Dattani et al. 2018). X1 refers to NBs, X2 to post-mitotic cells, and Xins differentiated cells. RNAi phenotypes, post-amputation (pa) or post-RNAi in homeostatic conditions, are documented for each gene where available.

Gene	Function	Dresden Transcriptome ID	X1 Prop	X2 Prop	Xins Prop	NBs?	Mitotic Activity?	Early Epidermal Progeny?	Late Epidermal Progeny?	Other Phenotypes?	Ref
MBD2/3	NuRD	dd_smed_v6_3054_0_2	43	52	5	No NB loss D7pa or 3 weeks post-RNAi homeostasis	No reduction D7pa or 3 weeks post-RNAi homeostasis	Increase D14pa No overall change, but anterior accumulation 3 weeks post-RNAi homeostasis	Decrease D7pa and 3 weeks post-RNAi homeostasis	Loss of gut branches, eyes, pharynx, brain neurons D7pa. Protonephridia and VNCs still form D7pa.	[Jaber-Hijazi et al. 2013]
HDAC-1	NuRD, CoREST/REST, NCoR/SMRT, Sin3, SHIP1	dd_smed_v6_695_0_1	46	49	5	NB loss Day 10 post-RNAi homeostasis	Reduction D4 post-RNAi homeostasis	Loss D10 post-RNAi homeostasis	Loss D12 post-RNAi homeostasis	No blastema formation following amputation	[Robb et al. 2008]
MTA1-like-1	NuRD, DNA Damage, Inflammation, EMT transition	dd_smed_v6_5860_0_2	44	49	6	-	-	-	-	-	-
MTA1-like-2		dd_smed_v6_3995_0_1	36	53	10	-	-		-	-	-
LSD-1	NuRD, CoREST/ CtBP, SIRT1	dd_smed_v6_7431_0_1	47	44	9	-	-	-	-	-	-
Rbap46/48–1	NuRD, Sin3	dd_smed_v6_5055_0_1	61	34	5	-	-	-	-	-	-
Rbap46/48–2		dd_smed_v6_2065_0_1	45	50	5	-	Reduction D14pa	-	-	-	[Hubert et al. 2015
Rbap46/48–3		dd_smed_v6_5609_0_1	53	44	4	-	-	-	-	-	-
CHD4	NuRD, Sin3	dd_smed_v6_2331_0_1	33	56	10	Reduction D16-18 post-RNAi homeostasis	Reduction D16-18 post-RNAi homeostasis	Reduction D16-18 post-RNAi homeostasis	Decrease D6 post-RNAi homeostasis	No blastema formation following amputation	[Scimone et al. 2010]
CHD3/CHD5		dd_smed_v6_9090_0_1	53	38	10	-	-	-	-	-	-
p66a	NuRD	dd_smed_v6_3115_0_1	28	62	11	Increase D14 post-RNAi homeostasis	Slight reduction D8pa	No change	Decrease D14 post-RNAi homeostasis	Loss of eye PRNs, brain neurons 7dpa. Protonephridia and VNCs still form D8pa.	[Vásquez-Doorman and Peterson 2016]

Overall, it is clear that future studies investigating the role of the NuRD complex in planarians should utilise the *mbd2/3* or *p66* phenotypes, as these subunits are specific to NuRD. Investigating the role of MBD2/3 in a DNA methylation-null organism like *S. mediterranea* has an important evolutionary significance, and may clarify an important DNA-methylation independent role. ChIP-seq will help to resolve whether genes are aberrantly marked by H3K27 acetylation and methylation in both NBs and NB-progeny following *mbd2/3* knockdown. Alternatively, development of a ChIP-grade planarian MBD2/3 antibody would help identify genomic targets of this protein.

2.8 SET1/MLL family of proteins - functional insights from planarian studies

Tri-methylation of lysine 4 on histone 3 (H3K4me3) is a major conserved mark of chromatin at nucleosomes immediately downstream of transcribed genes across metazoans. In yeast, the SET domain containing 1 gene (Set1) catalyses the mono-, di-, and tri-methylation of H3K4. The SET domain is a motif of ~130 amino acids that provide histone methyltransferase activity, and the SET1 protein forms a macromolecular complex called COMPASS (complex of proteins associated with SET1) (Miller et al. 2001; Krogan et al. 2002). In *Drosophila melanogaster,* there are three proteins homologous to Set1: dSET1, Trithorax (Trx) and Trithorax-related (Trr) which functions in a complex with LPT (lost plant homeodomains of Trr). In mammals, there are at least six Set1-related proteins: SetD1a and SetD1b that are orthologous to Drosophila dSet1; MLL1 and MLL2 orthologous to Drosophila Trx; and MLL3 and MLL4 that are orthologous to Drosophila LPT/Trr, with the N-terminus of MLL3/4 corresponding to LPT and the C-terminus for Trr (**Figure 3A and 3B**) (Eissenberg and Shilatifard 2010). Another homolog of Set1/MLL, called MLL5, is found in Drosophila and mammals, but lacks histone methyltransferase activity and has diverged in sequence and structure from other SET/MLL proteins (Emerling et al. 2002; Zhang et al. 2017).

Expansion of the COMPASS family evolutionary time implies diversification in H3K4 methylation function. Drosophila dSet1 and mammalian SetD1A and SetD1B complexes mediate the bulk of genomic H3K4me di- and tri-methylation indicating an involvement in global gene activation

(Ardehali et al. 2011; Mohan et al. 2011; Wu et al. 2008). Conversely, mammalian MLL1 and MLL2 are required for the methylation of a subset of developmentally important gene promoters. MLL2 is largely responsible for the methylation of H3K4 at bivalent genes in ESCs, whereas MLL1 is required for the H3K4 trimethylation of a smaller subset of genes and may be functionally redundant (Milne et al. 2002; Denissov et al. 2014).

Unlike the Set1/SetD1A/SetD1B and Trx/MLL1/2 complexes, the Trr/MLL3/MLL4 complexes are likely responsible for the deposition of H3K4me1 at promoters and, in particular, enhancers (Hu et al. 2013a; Cheng et al. 2014). Although active gene promoters are marked by H3K4me3 closest to the TSS, H3K4me1 at TSS-proximal regions is a mark of inactive genes and correlates with MLL3/4 occupancy at these regions. H3K4me1 spatially restricts H3K4me3 interactors on active genes, resulting in a bimodal ChIP-seq profile with H3K4me3 occupancy at the TSS but H3K4me1 signal both upstream and downstream of the TSS (Cheng et al. 2014). Conversely, H3K4me1 is also ubiquitous at active enhancers, but the functional relevance of this mark is not well understood. It has been proposed that MLL3/4 binds to enhancers, and recruits the coactivator p300, which acetylates H3K27 (Dorighi et al. 2017). Mouse ESC (mESC) knockouts for the catalytic domain of both MLL3 and MLL4 showed a loss of H3K4me1 and H3K27ac at enhancers, but the overall effect on enhancer RNA (eRNA) production and gene transcription was minimal. Conversely, complete MLL3/4 knockouts have a strong reduction in enhancer RNAs (eRNAs) and diminished transcription of target gene bodies. These results suggest that the function of MLL3/4 as a long-range coactivator is unrelated to methyltransferase activity (Dorighi et al. 2017). MLL3 and MLL4 are frequently mutated in a number of cancer types (Lee et al. 2009; Morin et al. 2011; Parsons et al. 2011; Jones et al. 2012; Pugh et al. 2012; Cleary et al. 2013), and changes in enhancer function may underlie tissue specific alterations in gene expression leading to cancer pathogenesis (Lee et al. 2013; Hu et al. 2013a).

Although these three groups of H3K4 methyltransferases have a well-documented role in promoter and enhancer activation, there is a substantial lack of knowledge concerning the exact loci these

marks affect and whether these targets are evolutionarily conserved. Two separate planarian studies have sought to understand the effects of these enzymes in the context of NB differentiation by knockdown of individual H3K4 methyltransferases, and have successfully related whole-organism phenotypes with epigenetic changes at target loci using ChIP-seq **(Figure 3C and 3D).**

RNAi of *set1* in planarians resulted in extensive neoblast loss and worms failed to produce a significant blastema upon amputation (Hubert et al. 2013; Duncan et al. 2015) **(Figure 3C)**. ChIP-seq identified a number of stem cell genes involved in RNA binding (i.e. *piwi-1*), transcription (*soxP-2*), and chromatin modification (MLL3) that were depleted for H3K4me3 following *set1* RNAi. Conversely, RNAi of the MLL1/2 orthologue resulted in a loss of epidermal cilia (Duncan et al. 2015) consistent with locomotory defects, and an earlier report also recorded a loss of ciliated protonephridia (Hubert et al. 2013). ChIP-seq and RNA-seq following MLL1/2 RNAi identified a much narrower set of genes that were depleted for H3K4me3 and transcriptionally downregulated, including a number of ciliogenesis related genes. Interestingly, these ciliogenesis related genes were shown to have a high H3K4me3 mark in NBs, comparable with that of NB-related genes, despite having low transcript expression in this compartment. Consequently, one conclusion of this study is that MLL1/2 specifically marks genes involved in ciliogenesis for later activation during differentiation, keeping them in a transcriptionally poised state in stem cells **(Figure 3D).** In this way, cilia related genes may in fact be bivalent – being marked with both H3K4me3 and H3K27me3, analogous to the role of MLL2 in establishing H3K4me3 at bivalent genes in ESCs (Denissov et al. 2014).

Three orthologues of mammalian MLL3 and MLL4 genes have been identified in the *S. mediterranea* genome. Two of these planarian orthologues, *Trr-1 and Trr-2*, are related to Drosophila Trr and the C-terminus of mammalian MLL3/4. RNAi knockdown of *Trr-1* led to a regenerative delay, with worms able to form a small blastema consistent with a reduction but not complete loss of mitotic neoblast activity. Conversely, *Trr-2* did not show any defects compared to wild-type worms. When double RNAi was performed with *Trr-1* and *Trr-2*, the phenotype was

enhanced with worms unable to form a regenerative blastema consistent with stem cell loss (Hubert et al. 2013). Moreover, animals began to form tissue outgrowths indicative of NB over-proliferation (Mihaylova et al. 2018). This strengthening of phenotype with double *Trr* knockdown is indicative of a degree of functional redundancy between the *S. mediterranea Trr* homologs.

The singular planarian homolog of Drosophila LPT and the N-terminus of MLL3/4 contains two PHD fingers and a singular PHD-like zinc finger binding proteins indicative of chromatin binding (**Figure 3A)**. RNAi of *LPT* resulted in failure of stem cell differentiation of some lineages including neuronal, epidermal, and pharynx regions (Mihaylova et al. 2018). Moreover, *LPT* RNAi worms showed an increase in neoblast mitotic activity before the formation of epidermal outgrowths that are populated with NBs, lineage-defined NBs and epidermal progeny. Consequently, *LPT* knockdown results in proliferation and differentiation defects that have cancer-like features, which is significant as both MLL3 and 4 are tumour suppressors in mammals (Lee et al. 2009; Morin et al. 2011; Parsons et al. 2011; Jones et al. 2012; Pugh et al. 2012; Cleary et al. 2013) . RNA-seq revealed a number of genes involved in cell proliferation and differentiation that were significantly upregulated including the serine-threonine kinase oncogenes *pim-2* and *pim-2-like*. Moreover, *p53* was significantly downregulated in this dataset, consistent with its well documented role as a tumour suppressor. For some genes, changes in transcription following *LPT* RNAi correlated with differences in H3K4me1 at the promoter region – for example, *pim-2-like* showed a decrease in H3K4me1 at the TSS indicative of its upregulation following *LPT* RNAi. However, for many genes correlations between RNA-seq and ChIP-seq data were not apparent. One likely explanation for this is that epigenetic changes at enhancers are acting to modulate changes in transcription, but currently enhancers are uncharacterised in planarians **(Figure 3D).** Future studies using ATAC-seq paired with available H3K4me1 data, would serve well to identify enhancers genome-wide, and changes to these regulatory elements following *LPT* RNAi can then be effectively assessed. Such a study may reveal novel enhancer targets that are conserved with mammalian MLL3 and MLL4, and more generally, would increase our knowledge of the effect enhancers have on transcriptional regulation during stem cell differentiation.

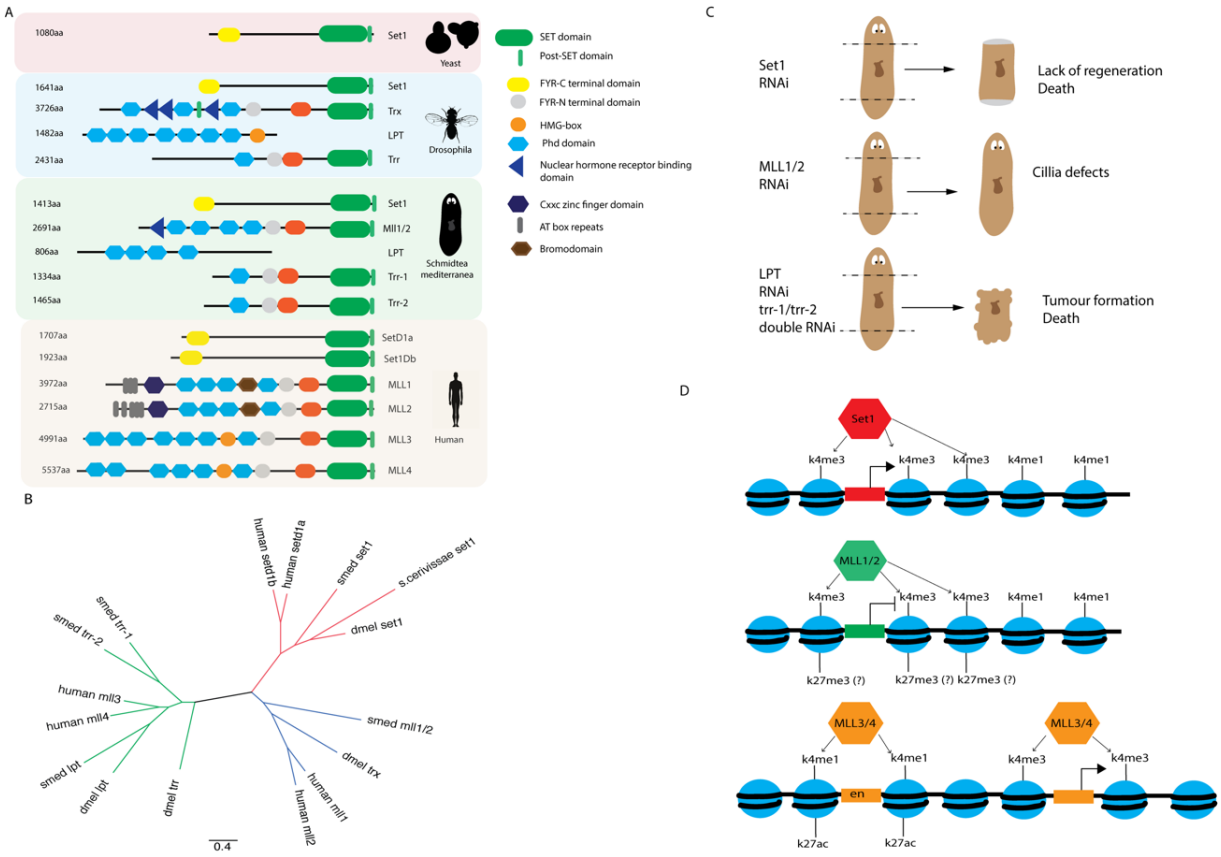


Figure 3. A. Domain architecture and duplication of SET1/MLL proteins in Yeast, Drosophila, *Schmidtea mediterranea*, and humans. **B.** Phylogenetic relationships SET1/MLL proteins in these four organisms. Proteins were aligned with MAFFT and RaxML was used for constructing a maximum likelihood phylogeny using the PROTGAMMA model, and branch supports were estimated with 100 pseudoreplicates. **C.** Cartoon showing planarian phenotypes after gene knockdown by RNAi. **D.** Diagram showing hypothetical mechanistic function relating to phenotypes and neoblast ChIP-seq analyses of Set1, MLL1/2 and LPT/Trr-1/Trr-2 genes. SET1 activates stem cell genes genome-wide by the addition of H3K4me3 mark at promoters. MLL1/2 regulates the transfer of H3K4me3 at differentiation-related gene promoters (i.e. ciliogenesis-related genes) that may be bivalent and marked by H3K27me3. MLL3/4 transfers H3K4me1 to active enhancers and inactive promoters, and regulate genes involved in cell proliferation and differentiation.

2.9 Polycomb repressive complex (PRC) and its role in maintaining bivalency and regulating stem cell differentiation

The term Polycomb originally referred to a mutant of *Drosophila* that displayed improper body segmentation, owing to a mis-regulation of homeotic genes (Jürgens 1985; Lewis 1978). The Polycomb Group of genes has since been used to define proteins that act as negative regulators of key developmental genes, and whose mutations result in phenotypes comparable to that of the

Polycomb. Polycomb mediated gene silencing works through the post-transcriptional modification of histones at two marks: H3K27me3 and H2AK119ub (Simon and Kingston 2013).

The PRC2 complex is responsible for the di- and tri-methylation of lysine 27 of histone 3 (H3K27me3) via its enzymatic subunit Ezh (Müller et al. 2002; Cao et al. 2002; Kuzmichev et al. 2002). This complex is broadly conserved in eukaryotes, but has been lost in some yeast species such as *S. pombe* and *S. cerevisiae* consistent with a genome-wide absence of H3K27me3 (Shaver et al. 2010). Conversely, the PRC1 complex has been traditionally thought of as being downstream of the PRC2 complex, with both complexes synergizing to recruit each other's modifying enzymes. The H3K27me3 mark, laid down by PRC2, is recognized and bound to by the chromodomain of the CBX protein component of the PRC1 complex. The E3-ligase protein (RING) of PRC1 ubiquitinates H2AK119, and this suppresses gene transcription by inhibiting RNA Poll II transcriptional elongation (Cao et al. 2005; Zhou et al. 2008). However, studies from vertebrates indicate that PRC1 also has a number of non-canonical roles independent of gene silencing owing to an existence of multiple paralogs of PRC1 components (reviewed in (Gil and O'Loghlen 2014; Aranda et al. 2015)).

Although high resolution microscopy experiments have suggested that both PRC1 and PRC2 mediated histone marks are important for chromatin compaction (Francis et al. 2004; Boettiger et al. 2016), sequential ChIP experiments have revealed the co-occurrence of PRC components (RING1B and Ezh2) with RNA Pol II at loci in ESCs, which is mirrored by the presence of both H3K4me3 and H3K27me3 (Lesch and Page 2014; Bernstein et al. 2006). These bivalent loci are defined by the presence of 'poised' RNA Pol II, which has a phosphorylated Serine 5 (Ser5p) at the heptapeptide repeat, YSTPSPS, of the C-terminal domain (CTD) of the largest subunit RPB1. Conversely, bivalent genes lack the elongating form of RNA Pol II, characterized by the presence of phosphorylated Serine 2 (Ser2p) at the heptapeptide (Brookes and Pombo 2009). The correlation between Pol II Ser5p and PRC repression means that bivalent genes are prepared for transcriptional activation upon differentiation. Indeed, knockouts of PRC1 and PRC2 components lead to the

aberrant expression of differentiation related genes, many of which are bivalent (Azuara et al. 2006; Boyer et al. 2006; Pasini et al. 2007).

The role of PRC genes in regulating planarian stem cell differentiation remains relatively unexplored, but the scope for investigation is considerable. Three planarian genes encoding homologs of PRC2 subunits, *ezh, suz12-,* and *eed-1*, were shown to be expressed in planarian stem cells by *in situ* hybridisation. RNAi knockdown of these genes in wild-type worms followed by amputation did not produce an observable phenotype, with no stem cell loss. In order to sensitize animals to any subtle RNAi defects, a dose of sublethal irradiation (12.5Gy) was used to reduce the stem cell number, such that surviving stem cells would have to proliferate in order to repopulate the stem cell compartment – a recovery process that takes 14 days (Wagner et al. 2012). Utilising this assay, the PRC2 genes were shown to be necessary for stem cell clonal expansion, with worms displaying epidermal lesions and eventual lysis as a consequence of stem cell loss. Although this phenotype is reflective of a general role of PRC2 in NB biology, genomics and transcriptomics based approaches would help in elucidating mis-regulated genes following RNAi.

It is clear that further analysis is needed to understand the effects of the canonical PRC complexes on the regulation of neoblast gene expression programs during both maintenance and differentiation. For example, bulk sequencing of the X1 NB compartment revealed genes involved in the differentiation process that were marked with both H3K4me3 and H3K27me3 at their promoters, indicative of bivalency (Dattani et al. 2018). In ESCs, CpG islands play an important role as PRC recruitment elements and are important in the assembly of bivalent chromatin at key developmental genes and restriction of elongating Pol II. (Tanay et al. 2007; Lee et al. 2017). Given that planarians do not have CpG islands understanding how PRC complexes localise to bivalent genes independently of this genomic reference could be relevant to mammalian biology.

2.10 PIWI, epigenetic silencing of transposable elements, and probable role in pluripotency gene regulation in planarian NBs.

A major selective force during the evolution of an organism's genome is the maintenance of genomic integrity over generations (Ernst et al. 2017; Sahu et al. 2017) . TEs are highly mutagenic, because they can insert into protein-coding genes, and contain repetitive sequences that can initiate ectopic recombination (Hedges and Deininger 2007). Given that TEs constitute a large proportion of eukaryotic genome, their repression is necessary for the maintenance of gene function and genomic stability. This is particularly true for multi- and pluripotent stem cells that must repress TE activity in order to maintain long-term proliferation over successive generations (Juliano et al. 2011; van Wolfswinkel 2014). In order to combat the invasion of TEs, metazoans have evolved a novel RNA class called PIWI-interacting RNAs (piRNAs). These small RNA molecules are 24-31nt long, and are transcribed from TE derived piRNA clusters in the genome. piRNAs bind to members of the PIWI (P-element induced wimpy testes) subclass of Argonaute superfamily of proteins (Aravin et al. 2006; Girard et al. 2006; Grivna et al. 2006; Watanabe et al. 2006). PIW-piRNA effector complexes can silence TEs either by epigenetic modifications at their genomic sites (transcriptional silencing TGS) or by cleaving TE transcripts directly (post-transcriptional silencing - PTGS) (**Figure 4a**).

Most animals typically have at least one nuclear expressed PIWI protein and one or two cytoplasmically expressed PIWI proteins that employ these distinct silencing modes (Weick and Miska 2014). For example, Drosophila germ cells express two cytoplasmic PIWI proteins, Argonaute 3 (Ago3) and Aubergine (Aub), and one nuclear called Piwi. The cytoplasmic Drosophila Ago3 and Aub bind to complementary TEs following transcription and directly cleave them using slicer activity. They also generate additional template piRNAs (secondary piRNAs) from the transposon debris thereby generating a piRNA self-amplification loop termed the 'ping-pong' cycle (Gunawardane et al. 2007; Brennecke et al. 2007; Huang et al. 2014). The nuclear Drosophila Piwi functions to silence TEs epigenetically by recruitment of the DNA methylation and/or histone

modifying complexes that lay down the H3K9me3 mark concomitant with the formation of heterochromatin (Le Thomas et al. 2014; Shpiz et al. 2011; Rozhkov et al. 2013).

The mechanism by which the PIWI-piRNA effectors guide the chromatin modifying machinery to the TE locus is beginning to be elucidated in animals. For example, independent RNAi screens in Drosophila identified the ovary specific nuclear protein CG9754/Panoramix that when eliminated leads to TE transcriptional increases similar to Piwi knockdown (Yu et al. 2015; Sienski et al. 2015). It is likely that CG9754/Panoramix acts as a protein scaffold between the nuclear PIWI-piRNA complex and the TGS machinery that includes the H3K9 methyltransferase SETDB1 and the heterochromatin protein HP1. No homologues for the Drosophila CG9754/Panoramix have been identified in mammals or other invertebrates. An evolutionary arms race between host organism and TE parasite means that proteins involved in the PIWI-piRNA pathway have diverged significantly between species, and different animals may have convergently evolved proteins with similar functions (Lewis et al. 2016).

The planarian genome has a repetitive content of -60%, far exceeding the 46% repeat content of the human genome, with many substantially large LTR members more than 30kb in length (Grohme et al. 2018). Previous studies have identified three major planarian PIWI proteins: SMEDWI-1, SMEDWI-2, and SMEDWI-3 in *Schmidtea mediterranea* and their respective orthologues DjPIWIA, DjPIWIB, and DjPIWIC in a sister species *Dugesia japonica.* All three of these genes are highly expressed in NBs and knockdown of *smedwi-2* and *smedwi-3* (but not *smedwi-1*) causes reductions in organismal piRNA levels, resulting in regenerative defects and lethality. One study in *D. japonica* revealed that following depletion of DjPIWIB, TEs continue to be silenced in NBs 7 days post RNAi, and NBs still retain the capacity for proliferation (Shibata et al. 2010). However, TEs were de-repressed at the onset of differentiation, and *in situ* hybridisation detected the up-regulation of a gypsy element in differentiated cells. Moreover, antibody staining revealed that DjPIWIA and DjPIWIC have cytoplasmic expression patterns restricted to the NB compartment, whereas DjPIWIB is expressed at the protein level in the nuclei of NBs and continues to be expressed

in post-mitotic progeny and differentiated cells (Shibata et al. 2016). We can hypothesise on the basis of protein expression and phenotype that DjPIWIB, like Drosophila Piwi, may function in the epigenetic silencing of TE loci, whereas DjPIWIA and DjPIWIB function in a ping-pong cycle, cleaving cytoplasmic TE mRNAs and generating piRNAs (Tharp and Bortvin 2016; Shibata et al. 2016; Sahu et al. 2017) (**Figure 4A**). Thus, when DjPIWIB is lost in the NBs, TEs continue to be silenced by the cytoplasmic PIWI proteins, and it is only at the onset of differentiation, when DjPIWIA and DJPIWIC expression is lost, that TE deleterious activity increases.

The question as to whether DjPIWIB/SMEDWI-2 is an epigenetic TE silencer remains outstanding, and proving so may potentially help in the understanding of NB maintenance and differentiation (Grohme et al. 2018). Since the planarian genome is replete with TEs, it is likely that the epigenetic silencing of these parasitic elements has shaped both the genome architecture and gene regulatory networks. For example, the deposition of the heterochromatic H3K9me3 mark at TEs in Drosophila germline stem cells has been shown to bleed to nearby gene promoters, causing their repression, or at least dampening of expression (**Figure 4B**) (Sienski et al. 2012). For planarian NBs, this would lead to a trade-off whereby TEs neighbouring highly expressed NB genes escape epigenetic silencing so that the NB gene expression program is not compromised. Thus, these TEs would be transcribed, but may be cleaved by the cytoplasmic PIWI proteins thereby preventing genomic stress. Conversely, it is possible that genes with high expression in the differentiated compartment (Xins) are able to establish heterochromatic marks at neighbouring TEs as they have no transcriptional activity in NBs. If this hypothesis is true, genes with high Xins expression will be aberrantly expressed in NBs following DjPIWIB/SMEDWI-2 knockdown, concomitant with a loss of the H3K9me3 mark at promoters. In order for Xins genes to be expressed correctly in the differentiated compartment, the effect of DjPIWIB/SMEDWI-2 on H3K9me3 must be counteracted, or at least dampened if DjPIWIB persists in all differentiated cells (Shibata et al. 2016). Alternatively, purifying selection may remove deleterious TEs that have suppressive effects on neighbouring NB genes at least in the asexual biotype, owing to lower heterozygosity that unmasks deleterious recessive alleles (Hollister and Gaut 2009; Barrett and Charlesworth 1991). These evolutionary

scenarios can be explicitly tested in the planarian NB system and can reveal how the genome architecture of animals can be shaped by the co-evolution with parasitic TEs.

Planarian NB piRNAs themselves may mediate small RNA transgenerational inheritance between successive NB divisions. As the asexual species must persist as an adult population the somatic NBs must be collectively immortal and underpin the homeostatic turnover of adult tissue. The maintenance of genomic integrity is therefore vital and cannot be compromised. In Drosophila, maternal deposition of cytoplasmic piRNAs to the developing egg prior zygotic transcription is important in kick-starting the piRNA generation system of the embryo in two distinct ways. In the nucleus, inherited piRNAs add the H3K9me3 to activate piRNA clusters in the embryo. In the cytoplasm, inherited piRNAs initiate the ping-pong cycle by providing an initial substrate for the cytoplasmic PIWI proteins (Le Thomas et al. 2014). A similar system may exist in planarian NBs to ensure genomic integrity between successive NB divisions. Moreover, if planarian piRNAs exist that are complementary to coding elements and are necessary for the suppression of differentiation genes, this raises an exciting possibility that NB identity can be preserved by piRNA mediated transgenerational inheritance.

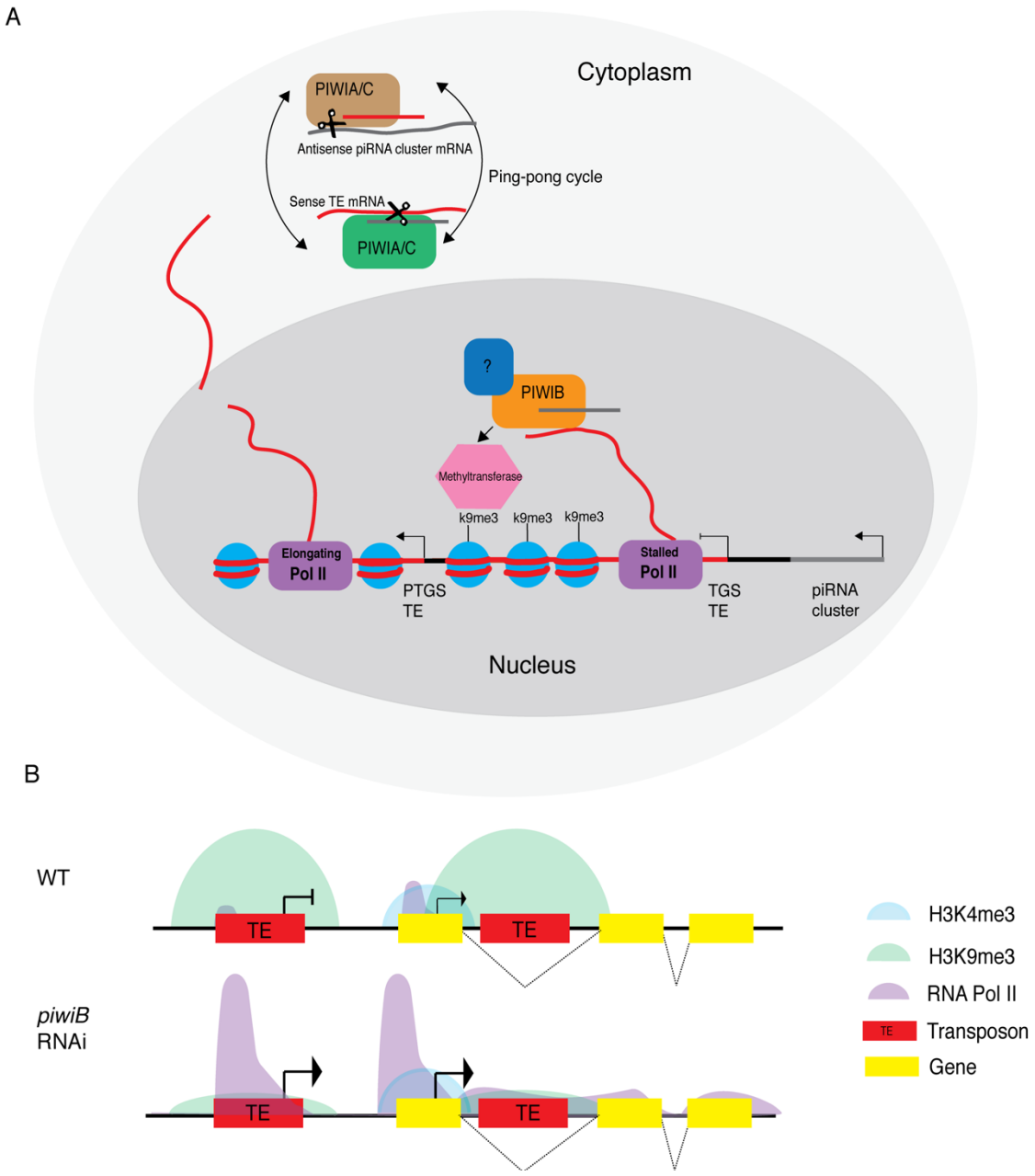


Figure 5. A. Hypothesis as to the function of the three planarian PIWI genes. PIWIB (SMEDWI-2 in *Schmidtea mediterranea*) functions in epigenetic silencing (transcriptional gene silencing; TGS) of transposons in the nucleus by recruiting an H3K9me3 methyltransferase. TEs that escape epigenetic silencing are cleaved in the cytoplasm and are post-transcriptionally silenced (PTGS) by binding of PIWIA (SMEDWI-1) and PIWIC (SMEDWI-3). These two cytoplasmic PIWI proteins participate in a ping-pong pathway to cleave TEs and produce piRNAs. **B.** TEs in wild-type (WT) planarian NBs are transcriptionally silent owing to H3K9me3 deposition preventing the recruitment and/or elongation of RNA Pol II. Following knockdown of *DjPiwib* in *Dugesia japonica* or *smedwi-2* in *Schmidtea mediterranea*, TE activity increases. In the case of a TE within an intron of a protein-coding gene, TE transcription and gene transcription both increase following knockdown.

2.11 Discussion: establishing a planarian program for studying the epigenetics of stem cell regulation

In this review, we have documented studies revealing the involvement of the NuRD, PRC complexes, SET1/MLL family of proteins, and PIWI in the epigenetic regulation of the planarian stem cell pluripotency and differentiation. In all four examples, elucidations of the function of these complexes have unravelled interesting biology that is relevant to mammalian stem cells and disease. Utilisation of epigenomic techniques such as ChIP-seq and ATAC-seq will potentially reveal gene targets of these chromatin complexes that may have been previously overlooked in traditional stem cell culture based systems. Moreover, comparative analyses between the epigenomes of planarian NBs and mammalian stem cells will also uncover conserved and divergent areas of the epigenome that potentially relate to the biology, life-history, and environment of the animal in question.

Thus far, the community has only considered the roles of canonical epigenetic complexes by RNAi, but it would be worth taking a top-down approach. For example, during planarian de-growth and growth, the animal remodels itself proportionally in response to nutritional status (González-Estévez et al. 2012; Mangel et al. 2016). How NBs are able to co-ordinate this feat at a molecular level remains an outstanding question, but the underlying mechanism is likely to involve feedback signalling from the differentiated cells mediating epigenetic changes to the entire stem cell compartment. Additionally, the manner in which stem cells read their position in the animal to activate region-specific differentiation programmes, most likely depends on the translation of signal gradients and positional control genes by epigenetic complexes to initiate both general and cell-type specific differentiation programs (Wurtzel et al. 2017). The recent advent of single-cell ChIP-seq and ATAC-seq/DNase-seq provides a promising avenue into these planarian investigations, and will help to characterize regulatory differences between individual NBs (Rotem et al. 2015; Clark et al. 2016; Buenrostro et al. 2015b; Cao et al. 2017; Clark et al. 2018; Cao et al. 2018) .

Chapter III

Asexual genome annotation and FACS categorization of annotated loci

Chapter III and IV have been published in part as: "*Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells*" in *Genome Research (2018)*. The reproduction of text and figures is in line with Copyright terms on the journal website. I share first authorship with Damian Kao, and with us Aziz Aboobaker wrote and revised the manuscript. Other authors assisted with the optimization of ChIP-seq experiments. The asexual genome annotation was generated by Damian Kao and figures are referenced as such. All other figures in Chapter III and Chapter IV and corresponding analyses have been generated by myself.

Abstract

Analysing data from ChIP-seq experiments requires a well-annotated genome. Moreover, in order to correlate the prevalence of particular histone modifications to gene expression, it is also necessary to have prior information regarding the expression and/or function of individual annotated genes. In this chapter, we present our methodology for annotating transcriptionally active loci in the asexual *Schmidtea mediterranea* genome, which contrasts with homology based methodologies such as MAKER. For each annotated locus, we allocate proportional values for a gene's expression in either S/G2/M NBs (X1), G1 NBs + post-mitotic progeny (X2), or differentiated cells (Xins) – the three broadly-defined cellular compartments that are isolated using Fluorescence Activated Cell Sorting (FACS). Our proportional categorization of loci agrees with the known expression patterns of specific genes, and also correlates with available single-cell RNA-seq data. In doing so, we have laid the groundwork with which to correlate both genome-wide and individual gene transcription with epigenetic data arising from planarian ChIP-seq experiments.

3.1 Introduction

Planarian NBs are the only proliferative cell type in the asexual biotype of planarians, and are responsible for both maintaining and regenerating the plethora of tissue types within the animal using a well-coordinated system of stem cell differentiation. In order to understand the unique biology of these animals, as well as the novel and conserved properties of their stem cells, an essential experimental tool is the isolation of NBs from the remaining post-mitotic cells of the animal.

Fluorescence Activated Cell Sorting (FACS) is a widely used technique for the compartmentalization of cell types from enzymatically dissociated tissues using fluorescent labelling and flow cytometry. For instance, combinatorial use of monoclonal antibodies to specific cell surface antigens has been used to isolate and fractionate rare subpopulations of the mouse hematopoietic system (Spangrude et al. 1988; Baum et al. 1992). However, the applicability of FACS to novel model systems and poorly characterized tissue types is generally hampered owing to lengthy and costly developmental procedures involved in the production of monoclonal antibodies to cell surface markers. Consequently, in more recent times, transgenic technologies has enabled the FACS isolation of individual cells harboring a florescent protein reporter. In *Hydra*, a burgeoning model system for regeneration studies, this has allowed for the convenient isolation and transcriptomic characterization of the three mitotic lineages (endodermal, ectodermal, and interstitial) (Juliano et al. 2014; Wittlieb et al. 2006; Hemmrich et al. 2012) .

Planarians are a stark anomaly in that they represent a relatively popular model organism that is not able to be transformed with exogenous DNA or edited using CRISPR/Cas9. Indeed, a published successful attempt in transforming NBs involved the electroporation of whole animals with DNA transposon vectors containing a P3-EGFP promoter target of the eye-specific TF Pax6 (Gonzalez-Estevez et al. 2003). However, this methodology has not been replicated by the same or other planarian groups. The reason as to why planarians are resistant to transformation is unknown, but we can speculate that since the NB itself has to be transformed, electroporating whole animals is an

inefficient way to go about this. Instead, electroporating sorted NBs with transgenic vectors, and reinjecting into NB-depleted animals following irradiation would be an obvious way in which to increase the chance of selection for the DNA construct, but is a technically challenging prospect. Given that the planarian genome has a -60% transposable element (TE) content (Grohme et al. 2018), with many TEs retaining transcriptional activity, endogenous elements may present an alternative way in which to integrate foreign DNA into the planarian genome.

In the absence of transgenic approaches and antibodies for confirmed cell lineage markers in planarians, FACS gating cell populations stained with Hoechst (nuclear dye) and calcein (cytoplasmic dye), or alternatively by Hoechst Blue/Red excitation, is the only available tool for isolating NBs, progeny, and differentiated cells (Hayashi et al. 2006; Romero et al. 2012). FACS allows for two irradiation sensitive compartments to be discerned: the 'X1' gate representative of S/G2/M-phase NBs with >2C DNA content; and the 'X2' gate representative of G1 phase NBs and post-mitotic progeny with 2C DNA content. The third FACS population, 'Xins', represents an irradiation-insensitive population with a higher cytoplasmic to nuclear ratio (**Figure 1A** and **1B**). These cell compartments are heterogeneous, with subpopulations of NBs expressing distinct lineage-specific markers present in the X1 population (Scimone et al. 2014; Van Wolfswinkel et al. 2014; Wurtzel et al. 2015), the X2 compartment consisting of an amalgam of G1 NBs and lineage-committed post-mitotic progeny (Baguñá and Romero 1981; Zhu et al. 2015; Hayashi et al. 2006; Molinaro and Pearson 2016), and Xins representing a broad collection of differentiated cell types.

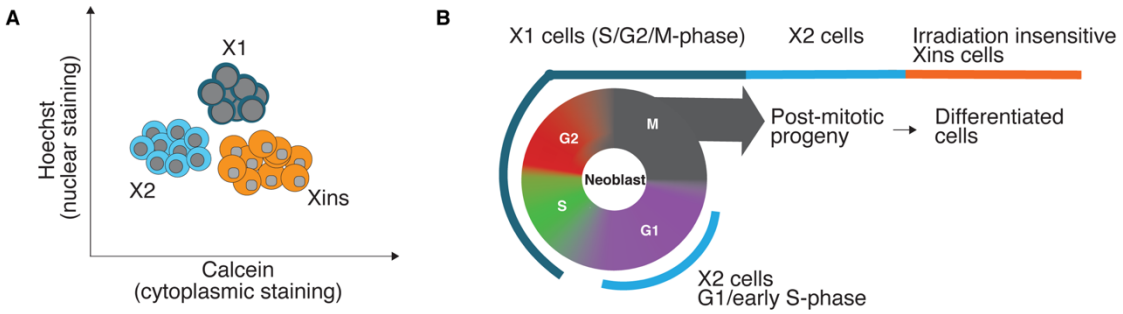


Figure 1: (A and B) Separation of FACS populations based on Hoechst and Calcein uptake. Cells with a high Hoechst content, indicative of a high DNA content, as well as a low cytoplasmic content, will be X-ray irradiation-sensitive X1 stem cells. G1 stem cells and post-mitotic progeny (X2 compartment) have a lower DNA content and low cytoplasmic content. Differentiated cells which are X-ray irradiation-insensitive (Xins) will have a low DNA content and high cytoplasmic content.

A number of labs have produced RNA-seq data from cells collected from these three distinct compartments. In this chapter, we introduce a new expression-based annotation of the asexual *S. mediterranea* genome, and use FACS RNA-seq datasets to produce consensus proportional expression profiles for annotated loci in X1 (S/G2/M stem cell), X2 (stem cell progeny + G1 stem cells) and Xins (differentiated cell) compartments. As a result, we are able to make predictions about a given gene's function at least with respect to a role in stem cell maintenance, lineage commitment or the maintenance of the differentiated state. Our FACS proportional analysis was corroborated by the existing expression profiles of individual planarian genes as well as tissue-specific single-cell data. Moreover, by analyzing gene expression in the context of the genome, we are able to now easily correlate the presence/level of histone modifications at genic promoters with their level of expression in the three FACS compartments.

3.2 Pipeline for establishing an expression-based annotation of the asexual S. mediterranea genome

We sought to produce an expression based annotation of all transcribed loci on the asexual *S. mediterranea* genome (SmedAsxl v1.1) utilising both *de novo* assembled transcriptomes and 164 independent RNA-seq datasets covering RNAi knockdown-, regenerating-, whole worm-, and cell

compartment-specific datasets. The inclusion of these diverse datasets was to improve the overall representation of the genome, and may be useful for discovering potential non-coding RNAs and protein-coding genes expressed at low levels, both of which may not have been fully covered by individual studies limited by read number, or reliant on homology based annotation processes such as MAKER (Robb et al. 2008, 2015). Consequently, our aim was to annotate all transcribed loci present in the genome utilizing this broad set of transcriptomic information.

To carry out this expression based annotation, we obtained *de novo* transcriptome assemblies and known gene sets from Planmine (Brandl et al. 2016), SmedGD (Robb et al. 2007) and NCBI. These sequences were mapped to the SmedAsxl 1.1 genome and were consolidated to produce an assembly of 67,037 transcripts. We also produced an independent reference assembly that took into account data from 164 RNA-seq datasets that covered a variety of experimental and biological conditions – this yielded a set of 91,464 transcripts from 47,427 potential genes (**Figure 2A**).

We next merged and consolidated both types of assemblies, and cleaned the resultant assembly to resolve cases where two transcriptional isoforms are assembled as separate loci owing to differences in splice junctions. To this end, we first clustered transcripts on the basis of their genomic coordinates overlapping. We then calculated a set of pairwise distances based on intron coordinates for each cluster. For each pair of transcripts (A and B) in a cluster, we found the length of the intersection in introns between both transcripts (intronLenA_B). We then divided intronLenA_B by the sum of all intronic lengths of transcript A (allintronLenA) and all intronic lengths of transcript B (allintronLenB). intronLenA_B was divided by allintronLenA to give a jaccard index value j_A, and also divide intronLenA_B by allintronLenB to give us j_B. We filtered all pairwise comparisons if both j_A and j_B are greater than 0.9 (**Figure 3)**. These filtered pairwise comparisons were used as edges on graphs, from which we then sought maximal cliques. These cliques are essentially groups of transcripts that are highly similar to each other, and which are most likely not true isoforms owing to this high degree of similarity. In the instance where no cliques are found, transcripts remain as isolated nodes, unconnected by edges, and as such are treated as isoforms (**Figure 4**).

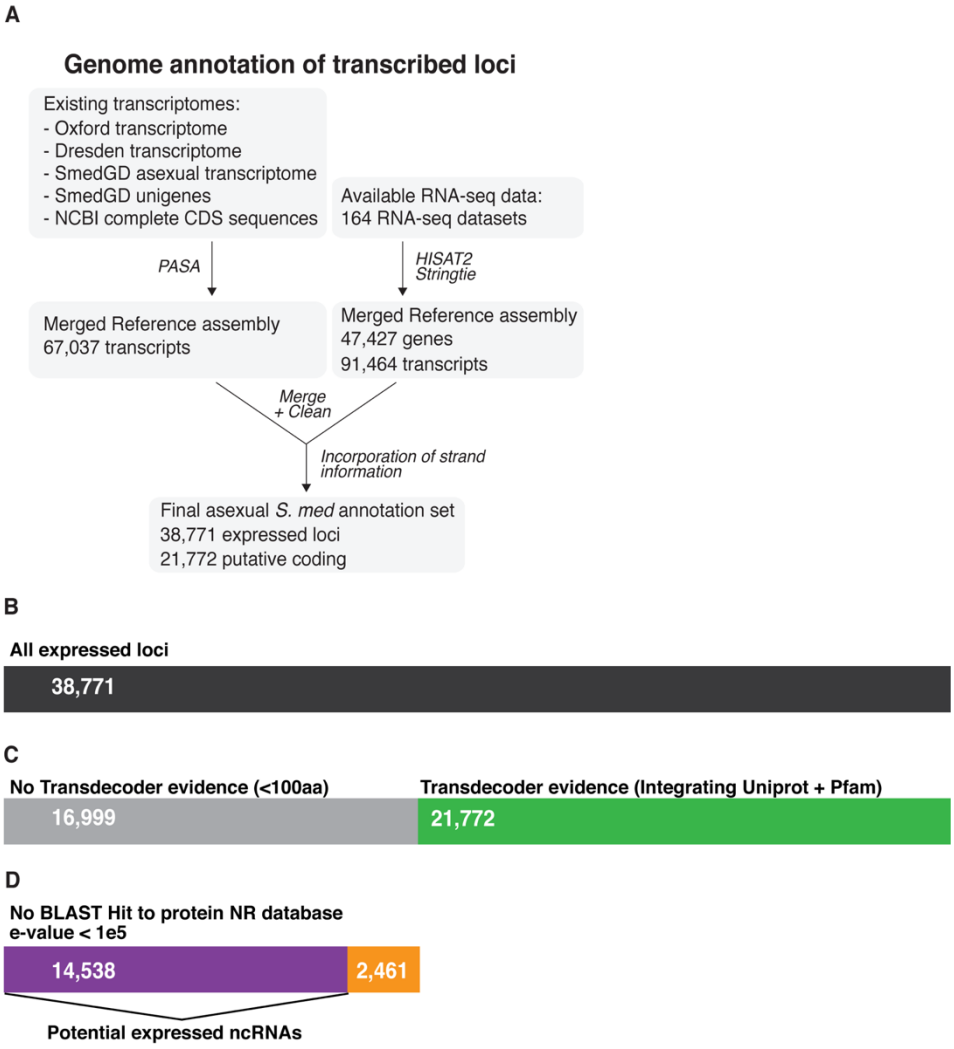


Figure 2: Overview of methodology for annotating the Schmidtea mediterranea asexual genome based on expression. One hundred sixty-four RNA-seq data sets, three de novo transcriptome assemblies, NCBI complete CDS sequences, and SmedGD Unigenes were mapped to the SmedGD Asxl v1.1 genome. Reference assemblies were merged, cleaned to remove potential splice variant redundancies, and the best representative transcript for each genomic locus was chosen. Strand information was obtained by BLAST to Uniprot, prediction of longest ORF, and data from strand-specific libraries. This process yielded a total set of 38,771 loci. (B) These expressed loci were filtered for those with TransDecoder evidence. (C) The remaining loci with <100aa length were then searched for in the NCBI NR database, and those without a hit were assigned as putative expressed ncRNAs.

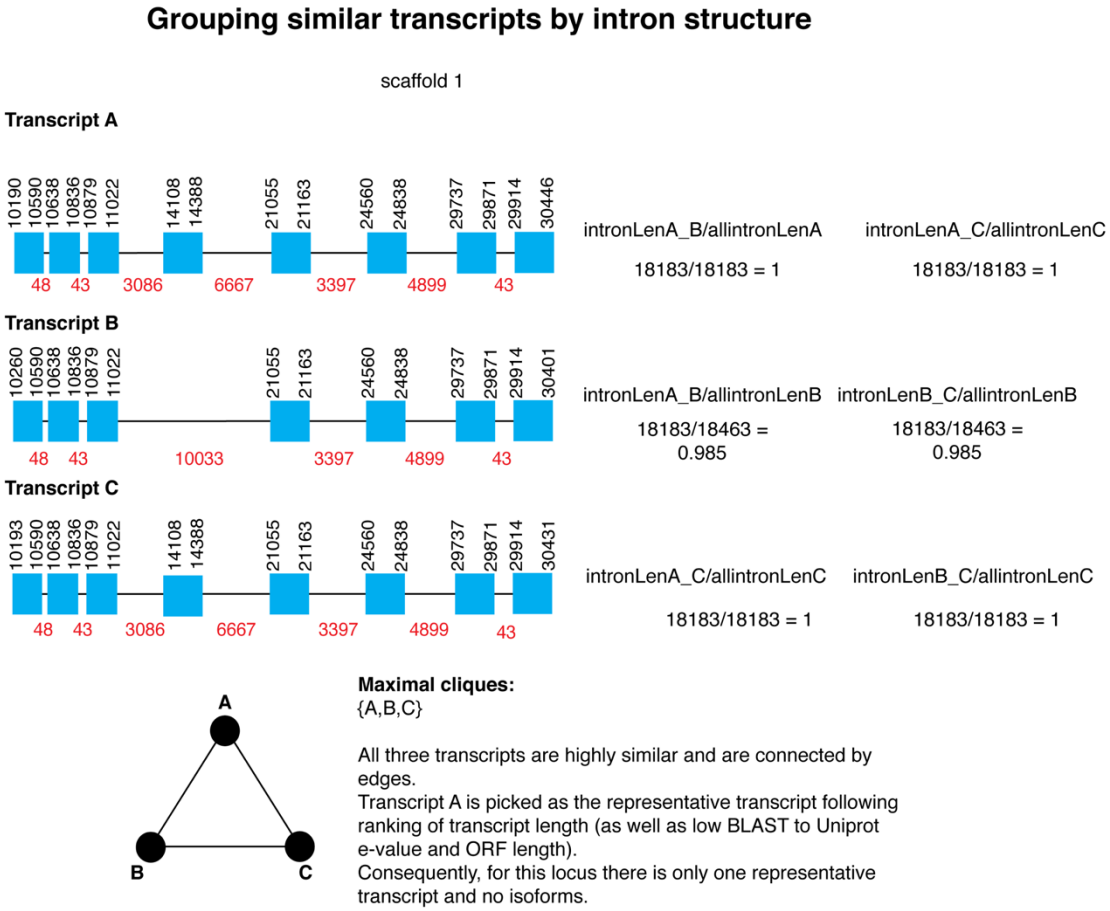


Figure 3: Example of choosing representative transcript for a particular genomic locus in our *S. mediterranea* (asexual) genome annotation. For each pair of transcripts (e.g. A and B, B and C, and A and C) in a cluster, we found the length of the intersection in introns between transcript pairs (e.g. intronLenA_B, intronLenA_C, intronLenB_C). We then divided the value of the intron intersection (e.g. intronLenA_B) by the sum of all intronic lengths for each transcript in the pair (e.g. all intronic lengths of transcript A (allintronLenA)) and transcript B (allintronLenB)). In the example, all pairs have a resulting jaccard index value above 0.9. The best representative transcript is picked on the basis of ORF length and BLAST to Uniprot e-value.

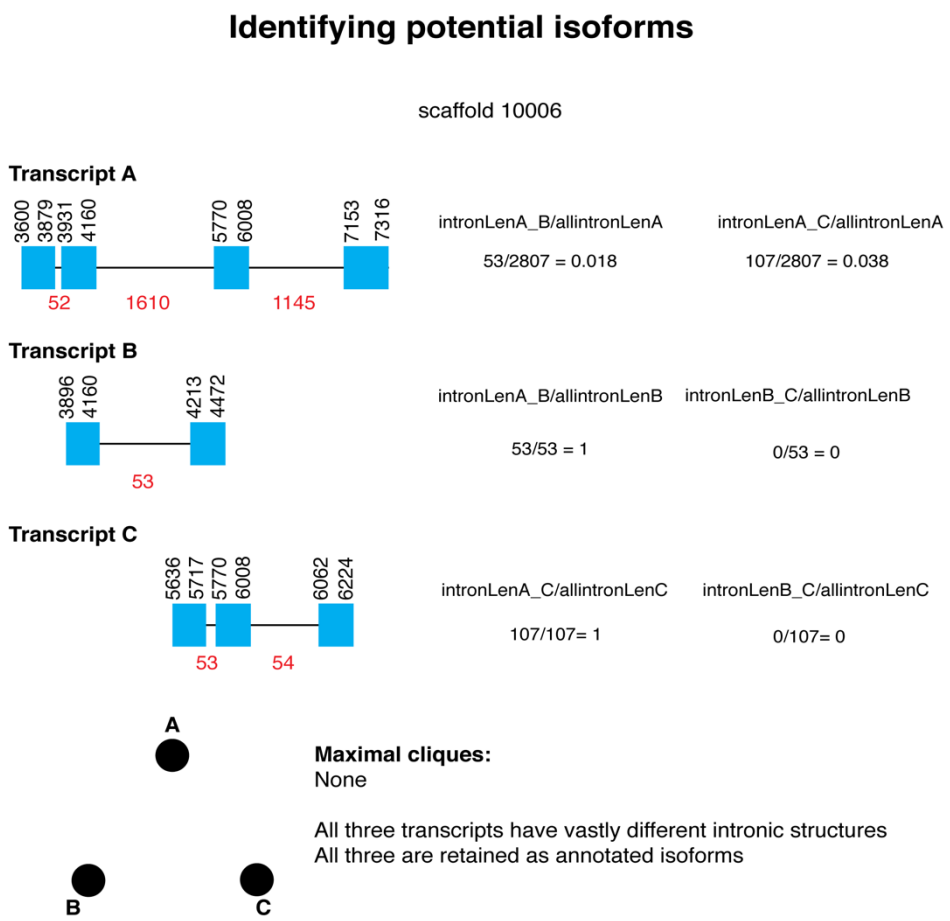


Figure 4: Example of isoform annotation in our *S. mediterranea* (asexual) genome annotation. In the above example of three transcripts, no pair of transcripts have a jaccard index (e.g. intronLenA_B/allintronLenA and intronLenA_B/allintronLenB) that are both above 0.9. Therefore, all three transcripts have vastly different intronic structures.

Our new expression-based annotation identified 38,711 expressed loci, 21,772 of which are predicted to be coding (>100aa). We utilize a broad definition of loci as genes that are both protein-coding as well as RNA genes that likely are non-coding (as per a definition of ORF length <100aa and lack of Pfam/Uniprot hits) (**Figure 2B**).

We additionally incorporated information on strand orientation by both BLASTx to Uniprot (metazoa). When information regarding strand information was lacking, we then utilized the strand information following mapping of strand-specific RNA-seq libraries to the SmedAsxl v1.1 genome. If, even in this instance, strand information was absent, we took the longest ORF and used this as

the default strand information. If both strands gave ORFs of the same length, we summed up the number of BLASTx hits to Uniprot metazoa in both forward and reverse strands, as well as the number of reads in the mapped BAM files for strand-specific libraries, and simply assigned strandedness based on the highest value. Despite this, we found 64 loci where strand information was ambiguous; it was likely that these few instances were ncRNAs that are transcribed from both directions.

Compared to the current available annotation of the *S. mediterranea* asexual genome (Smed GD 2.0) (Robb et al. 2007), our annotation discovered 10,210 new potential protein coding loci that are expressed at similar overall levels to previously annotated protein coding genes. A total of 6,300 genes from the existing MAKER homology-based annotation were not present in our expression driven annotation. Further analysis of these MAKER-specific genes shows that they generally have no or very little potential expression within the 164 RNA-seq libraries utilised for our annotation (**Figure 5**).

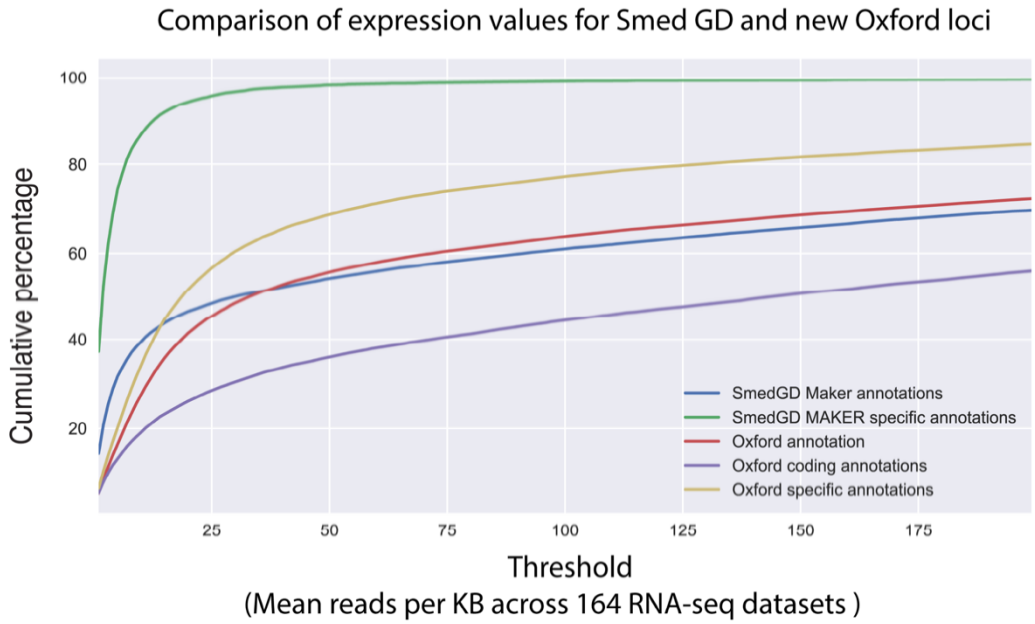


Figure 5: A comparison of mean expression levels for loci in the Smed GD 2.0 MAKER annotation with our new expression based annotation of the *S. mediterranea* asexual genome (Oxford). Expression value is mean estimated counts (from Kallisto output) per KB across 164 RNA-seq datasets. The graph shows cumulative percentages of annotations at a range of expression value thresholds for SmedGD MAKER annotations, Oxford annotations, SmedGD MAKER exclusive annotations, Oxford exclusive annotations and Oxford coding annotations predicted by TransDecoder. Loci in the Smed GD Maker annotation that are not included in the Oxford annotation have, on average, low expression values. Oxford-specific annotations incorporate a wider range of expression levels. **Figure and analysis by Dr. Damian Kao.**

3.2 Categorization of annotated loci by proportional expression of in FACS populations

We utilized publicly available RNA-seq datasets for the three different FACS populations in order to compare the expression of our annotated loci in these three distinct compartments (Önal et al. 2012; Labbé et al. 2012; Van Wolfswinkel et al. 2014; Zhu et al. 2015; Duncan et al. 2015). We first looked at the normalized TPM expression levels for annotated loci in our newly-annotated genome in the FACS population datasets originating from four different planarian labs. This revealed a rough congruence between different FACS populations from different labs (**Figure 6A**).

We transformed absolute TPM expression values into proportional values for each FACS compartment in each of the datasets (**Figure 6B**). These proportional values were then averaged

across datasets, to produce a final set of X1:X2:Xins proportions for 27,206 loci (18,010 of which are predicted to be protein-coding) that had at least 10 reads mapped in at least one FACS RNA-seq library. Consequently, we were able to sort all annotated genes by whether their predominant expression (i.e. >= 50% expression is in X1 (S/G2/M-phase NBs), X2 (NBs and stem cell progeny) or Xins (differentiated cells) (**Figure 7**).

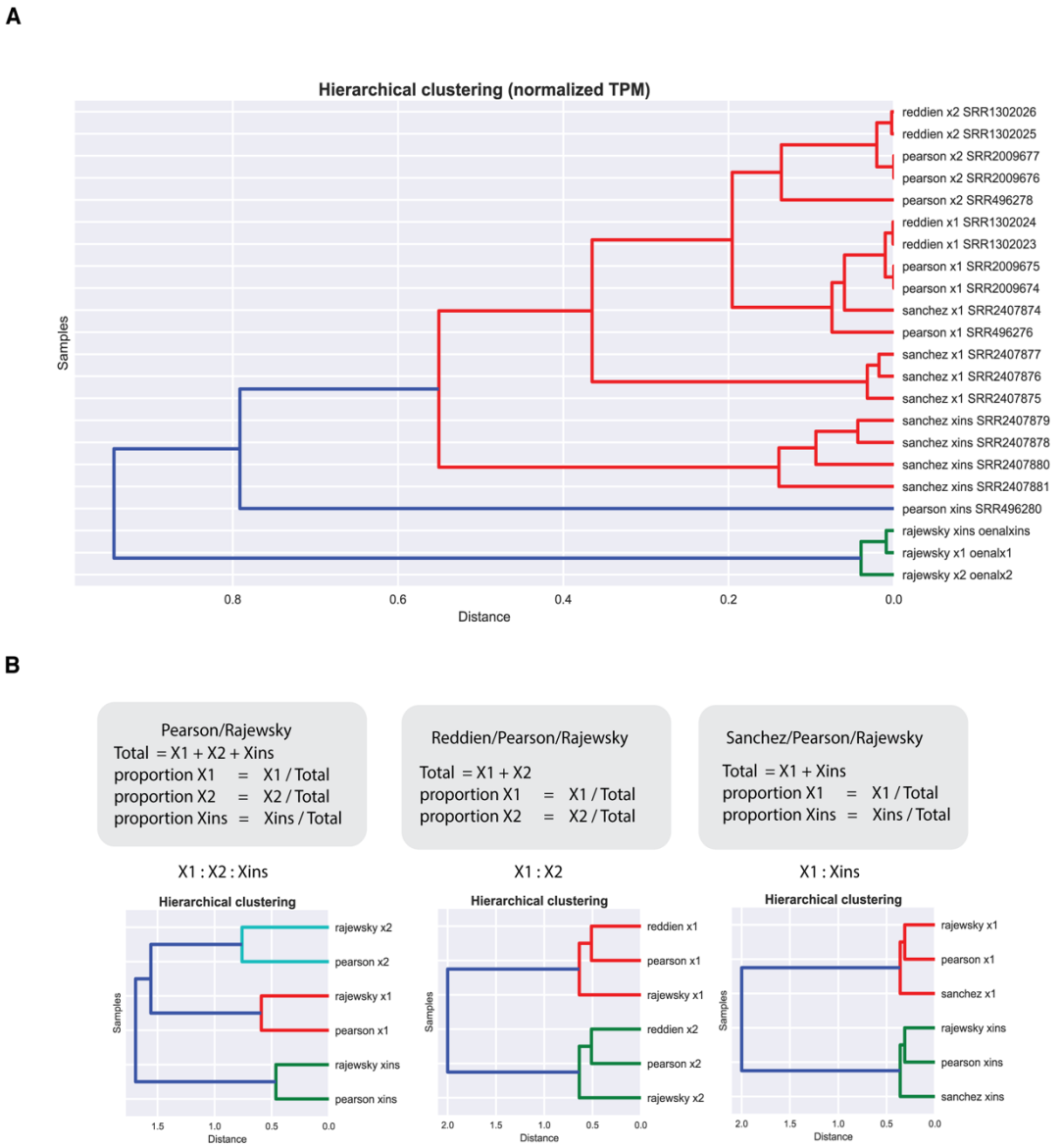


Figure 6: (A) Hierarchical clustering based on distance correlation between FACS datasets using normalized transcripts per million (TPM). Note how some X1 RNA-seq datasets (i.e. Sanchez SRR2407875- SRR2407877) do not cluster with other X1 RNA-seq datasets. This is most likely owing to a difference in FACS gating between experimental replicates and labs. (**B**) TPMs for each lab were averaged between labs in three groups. Pearson and Rajewsky labs have all three FACS datasets. Reddien lab has only X1 and X2 datasets. Sanchez lab has X1 and Xins. As such X1, X2, Xins average proportional values were calculated for Pearson and Rajewsky labs. Average X1 and X2 proportional values were calculated for Pearson, Rajewsky and Reddien datasets. Average X1 and Xins proportional values were calculated for Sanchez, Rajewsky, and Pearson datasets. Hierarchical clustering based on core distance between proportional values shows a consistent congruence between X1, X2, and Xins. **Figure and analysis by Dr. Damian Kao.**

A

Categorization of 38,771 total loci

Category	Criteria	Loci	Coding loci (% of category)
X1 enriched	X1 proportional expression =>50%	2,253	1,544 (68%)
X2 enriched	X2 proportional expression =>50%	8,444	4,781 (57%)
Xins enriched	Xins proportional expression =>50%	5,119	3,887 (76%)
X1/X2 enriched	X1 + X2 proportional expression =>75% Neither enriched in X1 nor X2	4,538	3,107 (68%)
X2/Xins enriched	X2 + Xins proportional expression =>75% Neither enriched in X2 nor Xins	3,652	2,688 (74%)
X1/Xins enriched	X1 + Xins proportional expression =>75% Neither enriched in X2 nor Xins	303	0 (0%)
Ubiquitous	Loci with roughly equal proportion in X1, X2, and Xins	2,897	2,003 (69%)
Unclassified	Loci with <10 reads in all FACS RNA-seq libraries	11,565	3,762 (33%)

Figure 7: 27,206 total annotated loci were categorized on the bases of FACS cell enrichment, with the remaining 11,565 not being categorized owing to have less than 10 read counts in all FACS RNA-seq libraries.

3.3 *Verification of FACS proportional categorization by Gene Ontology and individual known gene profiles*

We confirmed our FACS categorization analysis by checking for the enrichment of gene classes that correlate with known biological processes, as well as the expression of previously characterized individual genes. For the X1 NB category, Gene Ontology (GO) analysis revealed an enrichment for terms involved in cell-cycle/division related processes, as well mRNA processing and RNA-binding (**Figure 8A**). Indeed, conserved cell cycle related genes (e.g., *mcm2* (Treisman et al. 1995; Salvetti et al. 2000), *cdc6* (Bueno and Russell 1992), *cdk1* (Hartwell et al. 1973), *NCAPH* (Hirano 2012; Lai et al. 2018)) and RNA binding proteins that play a role in planarian NB maintenance (e.g., *bruli* (Guo et al. 2006), *piwi-1*, *piwi-2* (Reddien et al. 2005b; Palakodeti et al. 2008), and *tud-1* (Solana et al. 2009)) are all enriched within the X1 compartment (**Figure 8B**). For example, the Tudor gene family have a known role in mammalian germline stem cells for acting as molecular scaffold between PIWI

proteins and their piRNA targets (Siomi et al. 2010) . Knockdown of Tudor domain containing genes (TDRDs), such as TDRD9 (Shoji et al. 2009) and TDRD12 (Pandey et al. 2013), leads to embryonic lethality owing to de-silencing of TEs. In situ hybridisation of a planarian-specific TDRD, *tud-1*, as well as a homolog of TDRD9 revealed an mRNA expression pattern specific to NBs, and protein expression of TUD-1 was shown to be localized to the perinuclear chromatoid bodies (CBs) (**Figure 7C**). These organelles are analogous to the cytoplasmic granules known to be responsible for RNA processing in germ cells, and it is therefore not unexpected that genes involved in RNA-binding be enriched in NBs. Our analyses revealed the enrichment of genes in NBs that have been unappreciated in previous planarian stem cell studies, most likely owing to low absolute expression levels (Labbé et al. 2012; Önal et al. 2012; Solana et al. 2012) . For instance, we found that that *tert*, telomerase reverse transcriptase, is highly enriched in the X1 stem cell compartment. As with ESCs, this gene is involved in the maintenance of telomere length, and ensuring the maintenance of cellular immortality in planarians (Tan et al. 2012; Armstrong et al. 2005).

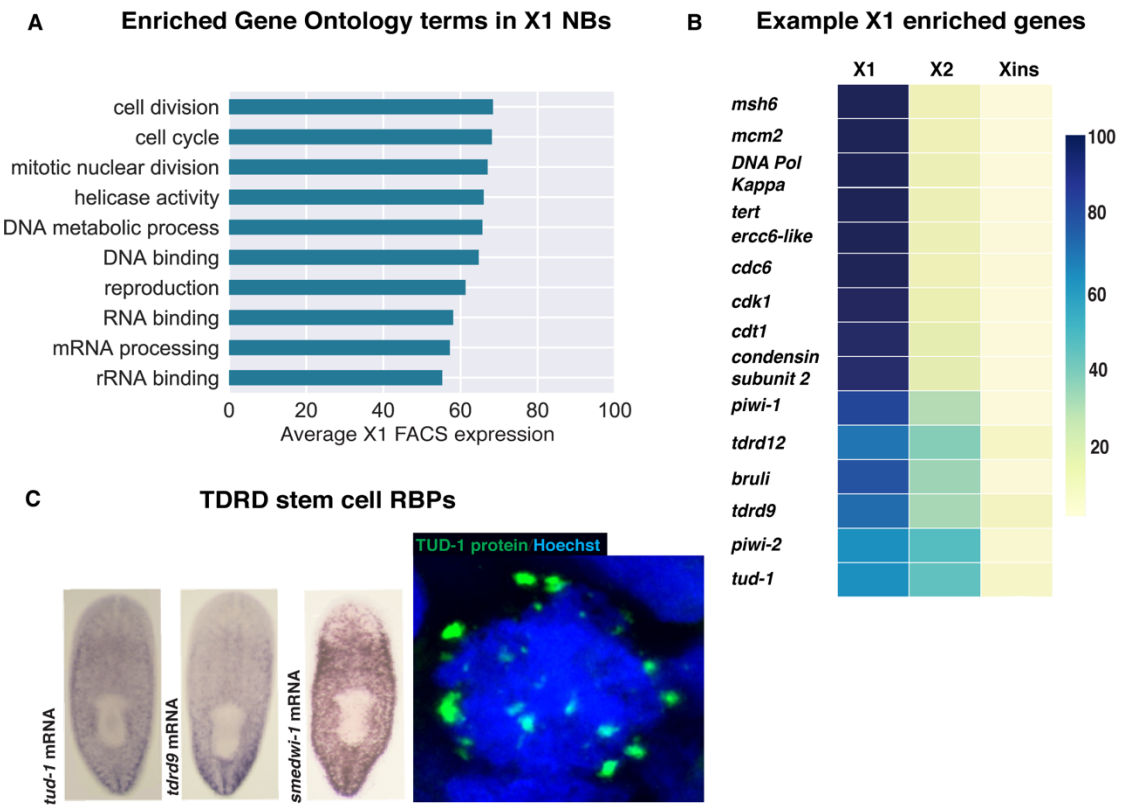


Figure 8. (A) Gene Ontology (GO) terms enriched in X1 enriched genes. Cell cycle, nuclear division, and DNA binding terms are indicative of genes involved in the maintenance of replication and division of NBs. (B) Example genes that are enriched in NBs. Genes involved in replication and cell-cycle processes such as *mcm2*, *cdc6*, *DNA Pol Kappa*, *condensin subunit2*, have a high overall expression in stem cells, even though they may be active at only one cell-cycle stage. RNA-binding proteins such as *tdrd12*, *tdrd9*, *tud-1* and *piw-1* and *piwi-2* are involved in the tethering of small-RNAs and likely play a role in the maintenance of genomic stability in NBs. (C) *In situ* hybridisation for two of these RNA-binding proteins, *tud-1* and *tdrd9*, shows that they have a typical stem cell pattern, as indicated by an absence of staining in the pharynx and inside gut. Protein staining against TUD-1 shows that the protein is located in perinuclear, granular structures called chromatid bodies, which are homologous to the RNA-processing centres of germ cells.

No significant enrichment for GO terms were found for X2 enriched genes, however, we verified that genes known to be involved in planarian lineage commitment had an expression profile of =>50% in this FACs compartments (**Figure 9A**). For instance, *soxP-3*, *egr-1* and *p53* are transcription factors (TFs) with a known role in activating genes for the development of the epidermal lineage, and as such have maximal expression in the X2 compartment, with smaller proportional representation in the X1 compartment. However, planarian studies have shown that the transcript levels of these three genes are detectable with *in situ* hybridisation in a fraction of *piwi-1*+ NBs, and prime stem cells to the epidermal lineage (Pearson and Alvarado 2010; Van Wolfswinkel et al. 2014). Other genes such as *prog-1* have comparatively little proportional representation in the

X1 compartment, consistent with previous results only 5-9% of *prog-1*+ cells being *piwi-1*+ as assayed by *in situ* hybridisation (Zhu et al. 2015) (**Figure 9B**). We reasoned that RNAi phenotypes which reduce the ability of stem cells to differentiate would lead to the mis-regulation of genes that are necessary for this process. Consequently, we looked at genes downregulated following RNAi of the planarian homolog of the RNA-binding protein MEX3, which results in a loss of multiple lineages but most obviously the epidermis. Indeed, an average FACS proportional expression of *mex3-1* 275 downregulated genes showed that most of these loci were expressed in the X2 FACS class, and loci known to be associated with epidermal progenitor specification such as *pmp-3*, *pmp-8, pmp-9*, *pmp-11* were all defined by a high X2 expression (all => 78% X2 gene expression) (**Figure 9C**). As an aside, defining the function of these *pmp* genes with no conserved domains and BLAST homology would help to define loci that are necessary for lineage commitment.

Genes with =>50% expression in the Xins FACS compartment were enriched in GO terms associated with ageing, formation of extracellular matrix, cell-cell signalling and peptidase activity – processes which are in line with the Xins compartment being comprised of short-lived cells in tissues such as the intestine, muscle, and neurons (**Figure 10A)**. Recently, genes involved in defining various cellular lineages of the worm have been identified by single-cell sequencing and verified by *in situ* hybridisation, and these same genes are Xins-enriched in our FACS proportional dataset (Fincher et al. 2018) (**Figure 10B and 10C)**.

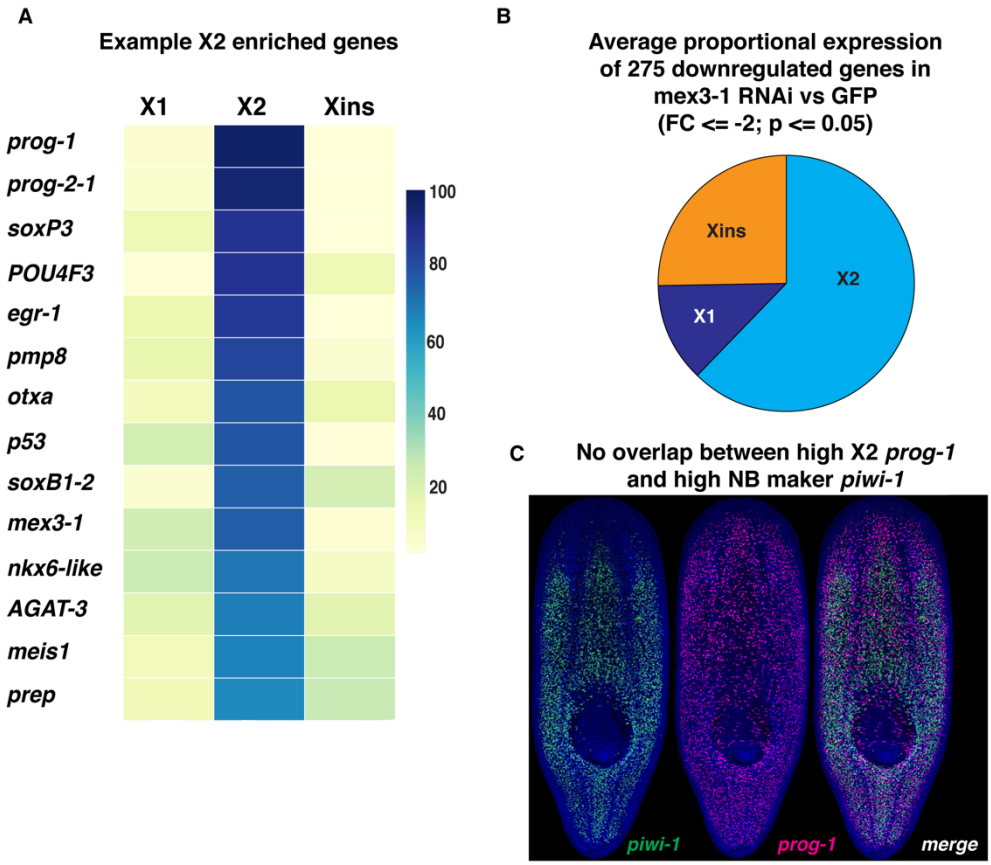


Figure 9. **A**. Example genes highly enriched for expression in the X2 compartment **B**. Pie chart showing the average wild-type FACS proportional expression for downregulated genes following *mex3-1* RNAi. **C**. *In situ* hybridisation pattern for X2-enriched gene *prog-1* which is a marker used for post-mitotic early epidermal progenitors. Note prog-1+ and piwi-1+ are distinct cell type, with very few cells expressing both genes, reflective of our proportional analyses. X2 genes are mostly downregulated consistent with a role for *mex3-1* in lineage commitment. Figure 9C are images from **Dr**. **Prasad Abnave**.

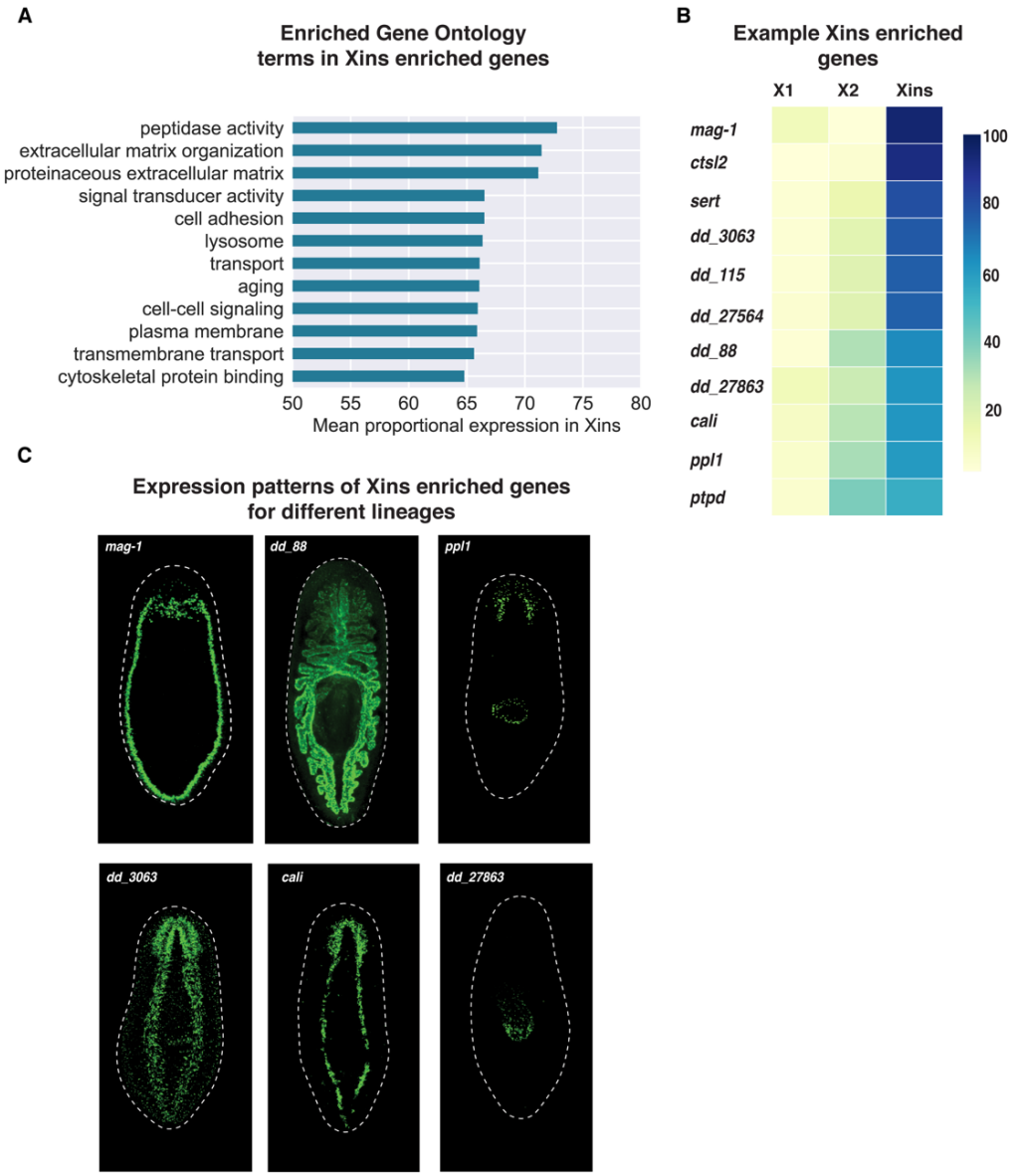


Figure 10. **A.** Gene Ontology (GO) terms enriched in genes with => 50% expression in the Xins FACS compartment. **B.** Example genes that are enriched in Xins and which have been documented to be specific to terminally differentiated lineages documented in Fincher et al. 2018. (C) *In situ* hybridisation patterns taken for these Xins enriched genes from the **https://digiworm.wi.mit.edu** database (Fincher et al. 2018).

3.4 Verification of FACS proportional categorization by analysis of single-cell datasets

Single-cell RNA-seq (scRNA-seq) technologies have recently been used to identify genes involved in lineage-commitment and teasing out heterogeneity within cell populations. In planarians, these have revealed heterogeneity in the expression profiles of NBs and have provided persuasive evidence

for the existence of lineage-committed NBs, such as progenitors of the epidermis and gut (i.e. ζ, and γ NBs) (Van Wolfswinkel et al. 2014). Moreover, we reasoned that genes with functions in lineage commitment and the maintenance of the differentiated cell types, would be representative of the X2 and Xins categories in our FACS proportional dataset. We consequently re-mapped reads emanating from two single-cell RNA-seq studies to our genome annotations, and extracted the top one thousand transcripts ranked by TPM gene expression for each cell type as defined by these two studies. Firstly, Wurtzel et al. (2015) have determined the transcriptomes of 619 individual planarian cells sorted by X1 and Xins FACS gates to identify 13 distinct cell types based on the portioning of cells into groups using the Seurat algorithm and identifying the expression of known planarian genes within these groups. Molinaro and Pearson (2014) produced transcriptomes for the 72 X1 and 96 X2 FACS isolated cells from the head region, and utilised this dataset to provide evidence for the existence of ν NBs – a population of *piwi-1* low, but *piwi-2* high stem cell neural precursors. We consequently looked at the position of these top one thousand genes in our FACS expression spectra sorted by X1 expression, and noted that the single-cell analysis fits expected patterns of expression thus independently validating our proportional categorizations (**Figure 11A**). For instance, cells defined as presumptive NBs (σ, ζ, and γ, and head X1) have top ranking genes with a higher mean expression in the X1 compartment by our analysis compared with the genes expressed in head X2 cells as well as Xins differentiated cell types (**Figure 11B**). We also note that all Xins cell types, with the exception of the epidermis II class, have an average enrichment for genes in the X2 class, with many more genes being expressed that have Xins enrichment compared to NB cells . These analyses meet our expectations given known planarian biology, as NB classes (σ, ζ, and γ) have genes that are enriched in the X1 class according to our proportional analyses, and are presumably involved in the maintenance of pluripotency, whereas differentiated cell classes express genes that are typically enriched in the post-mitotic X2 and Xins cell compartments.

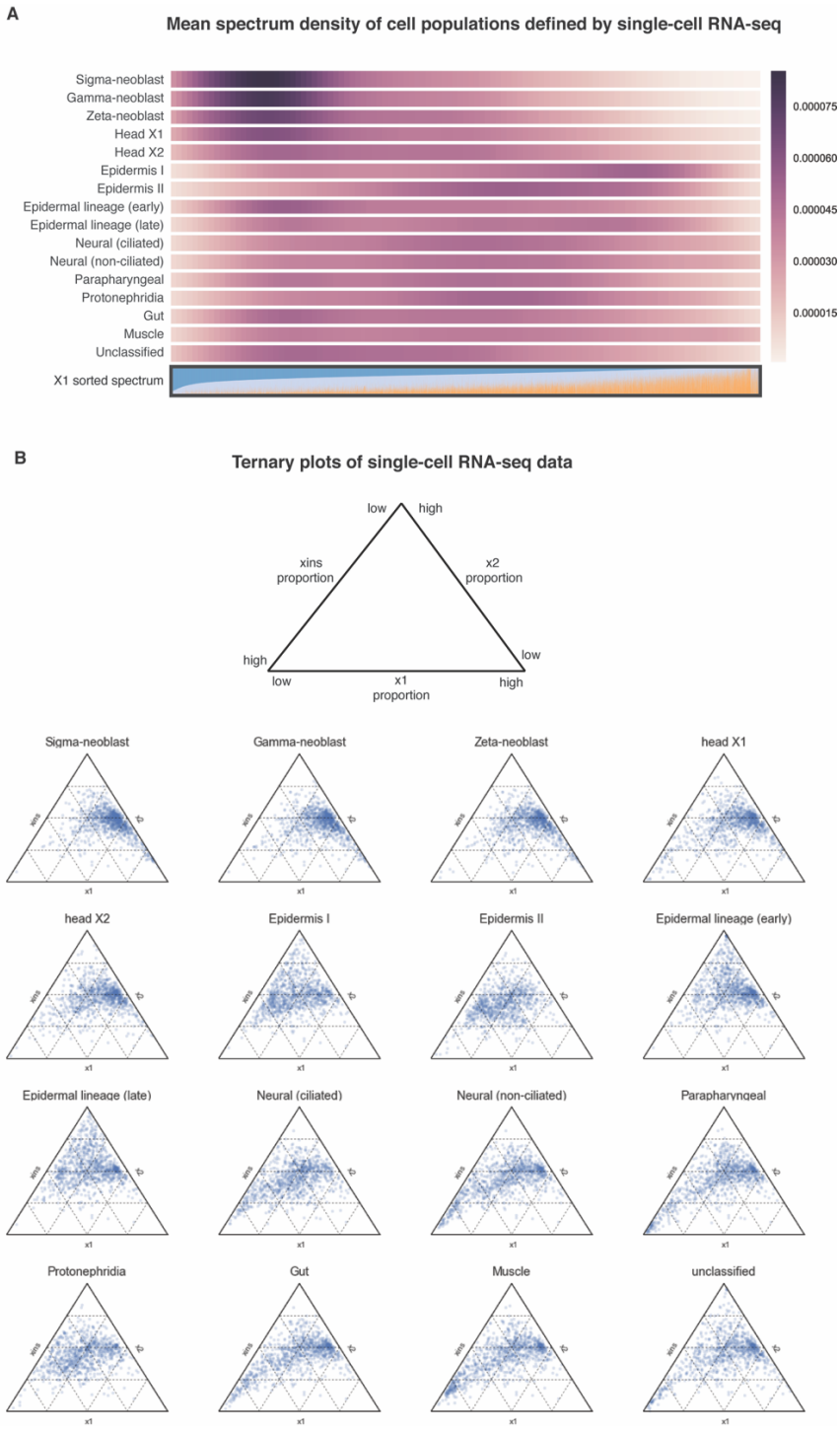


Figure 11: TPM values for single-cell RNA-seq libraries, mapped to our new asexual *S. mediterranea* annotation, for specific cell types were averaged and top 1000 genes were extracted. **A**. Cell types were classified previously by (Wurtzel et al. 2015) and are labelled accordingly. X1 and X2 single cells from (Molinaro and Pearson 2016) are from the head region and labelled 'head X1' and head X2'. Each row represents a cell type and the intensity colour represents the density of genes at the position on the proportional expression spectra. **B.** Ternary plots for each cell type were made and dots represent individual loci and their wild-type X1/X2/Xins proportional expression.

3.5 Discussion

In this chapter, we have presented our strategy for generating an expression-based annotation of the asexual *S. mediterranea* genome. Moreover, for each annotated locus we calculated proportional values pertaining to the level of expression in the three FACS compartments X1, X2 and Xins. This has allowed us to identify genes with predominant expression in the stem cell, post-mitotic, and differentiated cell compartments. Our analyses are consistent with reported expression profiles of individual genes and also correlates with data from single-cell sequencing experiments.

Given recent advancements in scRNA-seq technologies, we can now either use our FACS proportional data or single-cell expression maps to elucidate a given gene's function and lineage expression. In reality, both methods are equally important and complementary. For instance, we can imagine that a TF with low, but non-absent, expression in the NB compartment will enable the lineage-commitment of a few stem cells, and this TF's expression will increase during differentiation down a specific lineage. Given the low depth of sequencing and sensitivity in library making procedures, scRNA-seq may not effectively identify lowly-transcribed transcripts. Consequently, if one were to rely solely on transcriptional profiles resulting from scRNA-seq an impression would be gained that the TF is specifically expressed in more differentiated cells, whereas bulk FACS RNA-seq would identify low expression in NBs as well. However, FACS RNA-seq would not be able to identify which lineages specifically genes are expressed in, owing to the considerable heterogeneity within the three FACS categories. Thus, interpreting both single-cell and FACS RNA-seq datasets will enable for a truer representation of a gene's expression profile.

These considerations aside, in Chapter IV, we utilise our genome annotation as a mapping template for ChIP-seq data, and utilize our FACS proportional analyses to correlate the presence of specific histone marks in NBs with the transcriptional output of given annotated locus.

Chapter IV

Epigenetic analyses of planarian neoblasts demonstrates conservation of bivalent promoters in animal stem cells

Abstract

Currently, little is known about the importance of the epigenetic status of NBs and how histone modifications regulate homeostasis and cellular differentiation. In this chapter, we describe an improved and optimized ChIP-seq protocol for NBs and describe genome-wide profiles, with respect to the *S. mediterranea* asexual genome assembly, for four definitive epigenetic marks indicative of transcriptional status: the active marks H3K4me3 and H3K36me3, and suppressive marks H3K4me1 and H3K27me3. The genome-wide profiles of these marks were found to correlate well with NB gene expression profiles described in the previous chapter. We found that genes with little transcriptional activity in the NB compartment but which switch on in post-mitotic progeny during differentiation are bivalent, being marked by both H3K4me3 and H3K27me3 at promoter-proximal regions. In further support of this hypothesis bivalent genes also have a high level of paused RNA Polymerase II (Ser5P) at the promoter-proximal region.

Overall, this chapter confirms that epigenetic control is important for the maintenance of a NB transcriptional program and makes a case for bivalent promoters as a conserved feature of animal stem cells and not a vertebrate specific innovation. By establishing a robust ChIP-seq protocol and analysis methodology, we further promote planarians as a promising model system to investigate histone modification mediated regulation of stem cell function and differentiation.

4.1 Introduction

The promoters of developmental genes in mammalian embryonic stem cells (ESCs) are frequently marked with both the silencing H3K27me3 mark and active H3K4me3 marks. These marks are typically found in an asymmetric configuration, on sister histone tails of single nucleosomes (Voigt et al. 2012) (**Figure 1**). It has been proposed that this 'bivalent' state precedes resolution into full transcriptional activation or repression depending on ultimate cell type commitment (Voigt et al. 2013; Harikumar and Meshorer 2015; Bernstein et al. 2006). The advantage is that bivalency represents a poised or transcription-ready state, whereby a developmental gene is silenced in ESCs, but can be readily rendered active during differentiation to a defined lineage (**Figure 2**). Evidence for this comes from the finding that 51% of bivalent promoters in ESCs are bound by paused polymerase (RNAPII-Ser5P), compared with 8% of non-bivalent promoters (Brookes et al. 2012; Lesch and Page 2014); demonstrating a strong but not complete association. Bivalency may also protect promoters against less reversible suppressive mechanisms, such as DNA methylation (Lesch and Page 2014). Bivalent chromatin has also been discovered in male and female germ cells at many of the gene promoters that regulate somatic development, and may underpin the gametes' ability to generate a zygote capable of producing all cellular lineages (Lesch et al. 2013; Sachs et al. 2013; Lesch and Page 2014).

It remains unclear whether the poised bivalent promoters of developmental genes are an epigenetic signature of vertebrates or arose earlier in the ancestor of all animals. Recently, the orthologues of bivalent genes that sit at the top of transcriptional hierarchies in mammalian development, were also found to be poised in chicken male germ cells (Lesch et al. 2016). Sequential ChIP has also established H3K4me3/H3K27me3 co-occupancy of promoters in zebrafish blastomeres (Vastenhouw et al. 2010). Conversely, comparatively few bivalent domains were identified in *Xenopus* embryos undergoing the midblastula transition (Akkers et al. 2009). *Xenopus* genes which appear to have signals for both H3K4me3 and H3K27me3 originate from cells in distinct areas of

the embryo, and as such the observed bivalency can be explained by cellular heterogeneity (Akkers et al. 2009).

As ESC pluripotency requires bivalent chromatin, planarian NBs represent one possible scenario where poised promoters could have an important role in invertebrates, if this regulatory feature is conserved. Planarian NBs are a population of adult dividing cells that collectively produce all differentiated cells during homeostatic turnover and regeneration (Aboobaker 2011; Rink 2013). Several RNA-binding proteins, such as *piwi* and *vasa*, typically associated with nuage of germ cells are also expressed in planarian NBs where they function in the maintenance of pluripotency (Reddien et al. 2005b; Palakodeti et al. 2008; Solana 2013; Shibata et al. 2016; Lai and Aboobaker 2018). Moreover, the ability of NBs to differentiate upon demand must also require well-regulated transcriptional and epigenetic processes, and poised, bivalent promoters may constitute an effective way of coordinating the differentiation of these stem cells.

In this chapter, we describe an optimized ChIP-seq methodology for planarian NBs and combine this with informatics approaches to establish robust approaches for studying histone modifications at transcriptional starts sites (TSSs). By combining transcriptional and epigenetic analyses, we were able to identify genes with both inactive/low expression and bivalent promoters in the NB population, and which increase in transcription in post-mitotic NB progeny that are actively differentiating. Our findings indicate that bivalent promoters in pluripotent stem cells are not just a facet of vertebrates, but may have a role in regulating pluripotency in embryonic and adult stem cells across animals.

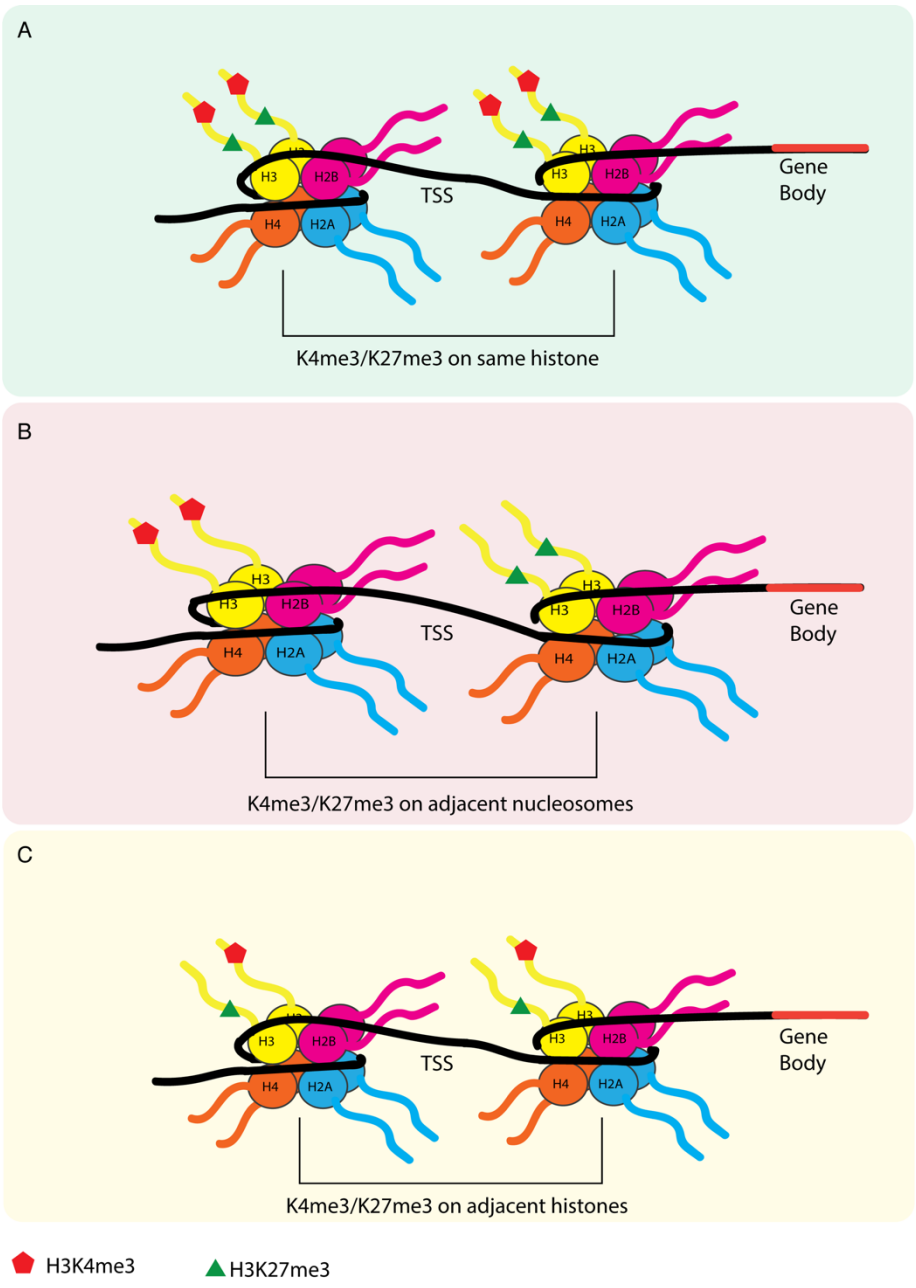


Figure 1. Potential conformations of bivalent promoters. Figure adapted from Voigt et al. 2013. **A.** H3K4me3 and H3K27me3 modifications may co-occupy the same histone H3 molecule in a nucleosome (i.e. symmetrically distributed bivalent histone modifications). **B.** Alternatively, H3K4me3 and H3K27me3 may occupy neighboring nucleosomes in the vicinity of the TSS. **C.** Single nucleosome Mass Spectrometry suggests that H3K4me3 and H3K27me3 co-occupy the same nucleosome in an asymmetric fashion featuring differentially modified copies of H3 (Voigt et al. 2012).

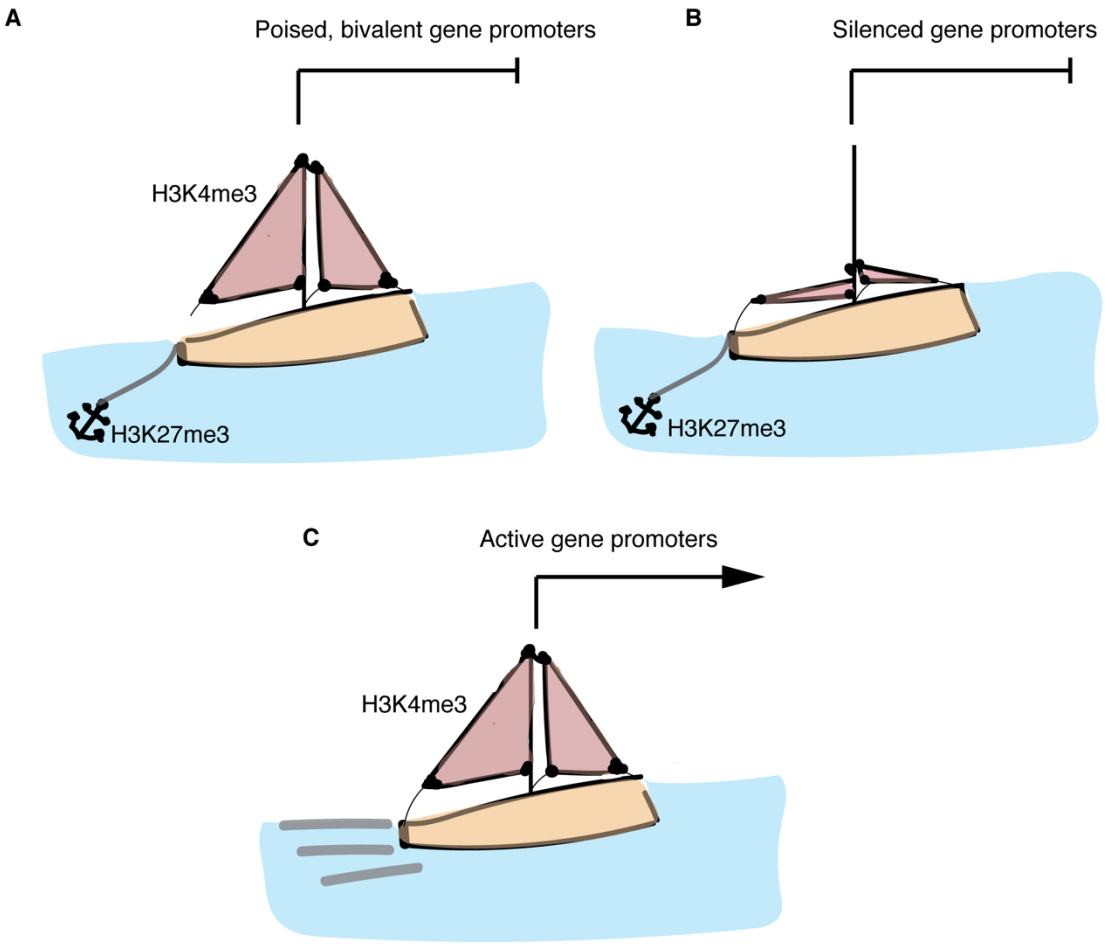


Figure 2. A. Bivalent gene, depicted as a boat, has its sail up (H3K4me3) but is held in static by its anchor (H3K27me3). **B.** A silent gene has its sail down and anchor down. It is easier for the 'poised' boat in **A** to transition to the active state in **C**, than for the stably silent boat in **B** to transition to **C**, as it simply requires the anchor to be lifted. Moreover, in order for a bivalent gene to become silent in a particular lineage (the transition from **A** to **B)** H3K4me3 simply has to be removed, analogous to the sail being pulled down. Figure adapted from Harikumar and Meshorer (2015).

4.2 Overview of an optimized ChIP-seq protocol for use with FACS-isolated planarian NBs

Research into the epigenetic mechanisms governing stem cell pluripotency in planarian NBs is still in its infancy (Dattani et al. 2019). Previous work has uncovered a lack of endogenous DNA methylation in the *Schmidtea mediterranea* genome, and characterized loss of function phenotypes for members of the NURD complex (Jaber-Hijazi et al. 2013; Vásquez-Doorman and Petersen 2016; Scimone et al. 2010), COMPASS and COMPASS-like families (Hubert et al. 2013; Duncan et al. 2015; Mihaylova et al. 2018). The first study to utilize ChIP-seq in planarians documented the effects

of *mll1/2* and *set1* RNAi with respect to the transcriptional activation mark H3K4me3 (Duncan et al. 2015). However, we revisited this data and noted that the total number of ChIP-seq reads from ~1 million X1 sorted NBs was relatively low in comparison to those from *Drosophila melanogaster* S2 'carrier' cells. The authors of this paper mixed Drosophila S2 cells (around 10^7) with FACS-isolated planarian NBs (around 10^6) to provide the molecular mass needed for ChIP-seq. Despite the Drosophila genome being an estimated 7-8 times smaller than the planarian genome, the total planarian content of their ChIP-seq reads was considerably smaller than expected.

Consequently, we developed an optimized ChIP-seq protocol for FACS sorted X1 NBs without the addition of excess 'carrier' cells. We were able to generate high quality uniquely mapped reads to our annotated *Schmidtea mediterranea* genome using only 150-200,000 X1 cells per immunoprecipitation (IP) – 5 to 7 times less material than the previously established planarian protocol (Duncan et al. 2015) . Although we have not tested this previous methodology for planarian NB ChIP-seq, possible reasons as to why metrics for our protocol indicate that it is more sensitive include: (1) a shorter duration between FACS isolation and commencement of ChIP protocol; (2) DNA/protein crosslinking on nuclei as opposed to whole cells; (3) antibody affinity to *S. mediterranea* (the previous protocol used H3K4me3 Millipore 08-473 whereas we used Abcam ab8580; but the Abcam ab09050 antibody against H3K36me3 remained the same); (4) utilisation of the NEBNext Ultra II kit that can generate DNA-seq libraries from starting material of 500pg, thus avoiding the need to bulk up chromatin input with Drosophila material. A low genomic coverage of ChIP and input reads can reduce the likelihood of capturing weakly enriched areas of the genome, especially when using peak calling methodologies such as MACS2 (Jung et al. 2014). Therefore, our protocol allows for a truer representation of the histone modification landscape of the planarian genome.

Prior to all IPs, we also added an estimated 3% *Drosophila* S2 spike-in simply as a method to normalize any technical differences across replicate libraries of an IP (Orlando et al. 2014). The addition of an exogenous reference genome meant that replicates of the same IP could be scaled

according to the Drosophila amount added (as measured by read total in input) and the amount retrieved following the IP. In this way, replicate IPs could be rendered comparable in read coverage. This normalization method was also used by our lab in a separate study to effectively compare IPs performed on two biologically different samples (e.g. wild-type and LPT RNAi samples). By normalizing to a constant Drosophila S2 spike in, local differences in read tag density can be identified that are as a result of the experimental condition and which would not otherwise be uncovered by traditional reads per million (RPM) ChIP-seq normalization methods (Orlando et al. 2014; Mihaylova et al. 2018). Given that our wild-type IPs documented in this chapter already contain this spike-in, we can perform ChIP-seq on RNAi NBs at a later date and compare to wild-type samples provided the RNAi samples also contain a spike-in.

For the results described in this chapter, *Drosophila* spike-in reads accounted for ~17 % of X1 H3K4me3 libraries compared to an average of ~87% in the previous study's X1 H3K4me3 libraries. Moreover, *Drosophila* spike-in reads accounted for ~9% of our X1 H3K36me3 libraries compared with ~99% of the single X1 H3K36me3 replicate included in a previous study (Duncan et al. 2015) (**Figure 3**).

Comparison of Aboobaker and Sanchez (Duncan et al. 2015) X1 ChIP-seq libraires

		S. med fragments	*D.mel* fragments	
Aboobaker	H3K4me3	14,724,474	3,572,880	SRR4089722
	H3K4me3	13,252,896	993,932	SRR4089758
	H3K4me3	26,551,125	8,812,331	SRR4089769
Sanchez	H3K4me3	3,550,931	15,244,101	SRR2726623
	H3K4me3	3,473,370	17,332,482	SRR2726624
	H3K4me3	5,392,596	30,787,848	SRR2726649
	H3K4me3	3,578,055	29,897,763	SRR2726650
	H3K4me3	765,305	20,051,292	SRR27226607
	H3K4me3	2,370,282	18,811,157	SRR27226608
Aboobaker	H3K36me3	22,724,137	2,635,922	SRR7187811
	H3K36me3	22,856,257	1,264,684	SRR7187812
Sanchez	H3K36me3	571,728	39,104,178	SRR2842081

Figure 3: Comparison of number of mapped fragments to *S. mediterranea* asexual genome. Sanchez single-end ChIP-seq H3K4me3 and H3K36me3 libraries made from X1 FACS isolated cells are from Duncan et al. 2015. Aboobaker paired-end ChIP-seq libraries are from Dattani et al. 2018a and Mihaylova et al. 2018.

4.3 An optimized ChIP-seq protocol reveals H3K4me3 and H3K36me3 levels correlate with active gene expression in planarian NBs

We tested the robustness of our ChIP-seq protocol with reference to both H3K4me3 and H3K36me3 – epigenetic marks that are known to positively correlate with gene expression in other model systems. H3K4me3 is laid down at active and bivalent promoters by the Trithorax group (TrxG) complexes containing SET or MLL enzymes (Hu et al. 2013b; Bledau et al. 2014; Denissov et al. 2014) . H3K36me3 is a mark of transcriptional elongation, and is deposited on histones as they are displaced by RNA polymerase II and as such this modification is enriched towards the 3' end of genes (Li et al. 2002; Wagner and Carpenter 2012). H3K36me3 is hypothesized to prevent spurious transcriptional initiation at cryptic promoter-like sequences within exons and, in yeast, this is achieved by the recruitment of histone deacteylase complexes (HDAC) that erases elongation-associated acetylation (Carrozza et al. 2005; Joshi and Struhl 2005).

As predicted, ChIP-seq of H3K4me3 in X1 NBs revealed a high average peak around the TSSs of genes characterized as being X1 enriched (**Figure 4A**). Conversely, we observed comparatively lower H3K4me3 deposition at the TSSs of Xins enriched genes not expressed or expressed only at very low levels in X1 cells. Intermediate levels of H3K4me3 in the X2 compartment are consistent with this FACS population being a mixture of G1 NBs and post-mitotic progeny. Indeed, genes with the highest proportion of X2 expression (i.e. 'high ranking X2 genes') indicative of expression in post-mitotic progeny but not NBs had lower levels of H3K4me3 in X1 cells compared with low ranking X2 genes that retain expression in cycling G1 NBs **(Figure 4B).** We next calculated a base by base Spearman's Rank correlation coefficient between ChIP-seq signal to FACS proportional expression values of annotated loci across a 2.5 kb region either side of the TSS. This was done by producing two vectors for all 50bp windows around the TSS's of annotated loci – one vector for proportional expression (either X1, X2, or Xins), and another for coverage – and then calculating a Spearman's rank correlation coefficient for this assayed window. The resultant graph shows a positive correlation between the level of X1 proportional expression and the level of H3K4me3 deposition close to the TSS (**Figure 8A**). On the other hand, there is a negative correlation between H3K4me3 deposition and Xins proportional expression across the same region. Thus, a high H3K4me3 ChIP-seq signal reflects higher expression of a locus in X1 NBs, whereas lower H3K4me3 signal reflects lower X1 NB gene expression but higher expression in the differentiated Xins compartment.

ChIP-seq plots of H3K36me3 split by FACS gene expression revealed, as predicted, a higher average peak around X1 enriched genes when compared with the X2 and Xins FACS enrichment categories (**Figure 4C**). Importantly, the average peak for X1 genes is located towards the 3' end of genes, whereas the smaller Xins peak is promoter-proximal by comparison. This can be explained by a higher level of transcriptional elongation of X1 transcripts in NBs compared with Xins genes that have a predominant expression in the differentiated compartment. When splitting X2 enriched genes by rank order, we observe that genes with highest expression in the X2 compartment and, as a

consequence lowest transcript abundance in NBs, have an enrichment for H3K36me3 at the promoter-proximal end of the gene (**Figure 4D**). Conversely, with decreasing X2 proportional expression and a concomitant increase in transcriptional activity in the NB compartment, the average peak of H3K36me3 is shifted downstream of the TSS towards the 3' ends of genes.

We also looked at the individual H3K4me3 and H3K36me3 profiles of genes known to be highly expressed in NBs, and compared this to the signal for the suppressive marks H3K4me1 and H3K27me3 (see later). We confirmed that known metazoan genes associated with stem cell maintenance, such as cell-cycle and replication related genes (i.e. *mcm2, cyclin-B1, wee1, cdt1*), RNA-binding proteins (*piwi-1, ddx52*), DNA-damage response (DDR) genes (*errc6-like, exonuclease 1*) and epigenetic-related genes (*setd8-1)*, all have high levels of H3K4me3 at the promoter-proximal end and H3K36me3 in the gene body, but a comparatively low signal for the suppressive marks H3K4me1 and H3K27me3 (**Figure 5**).

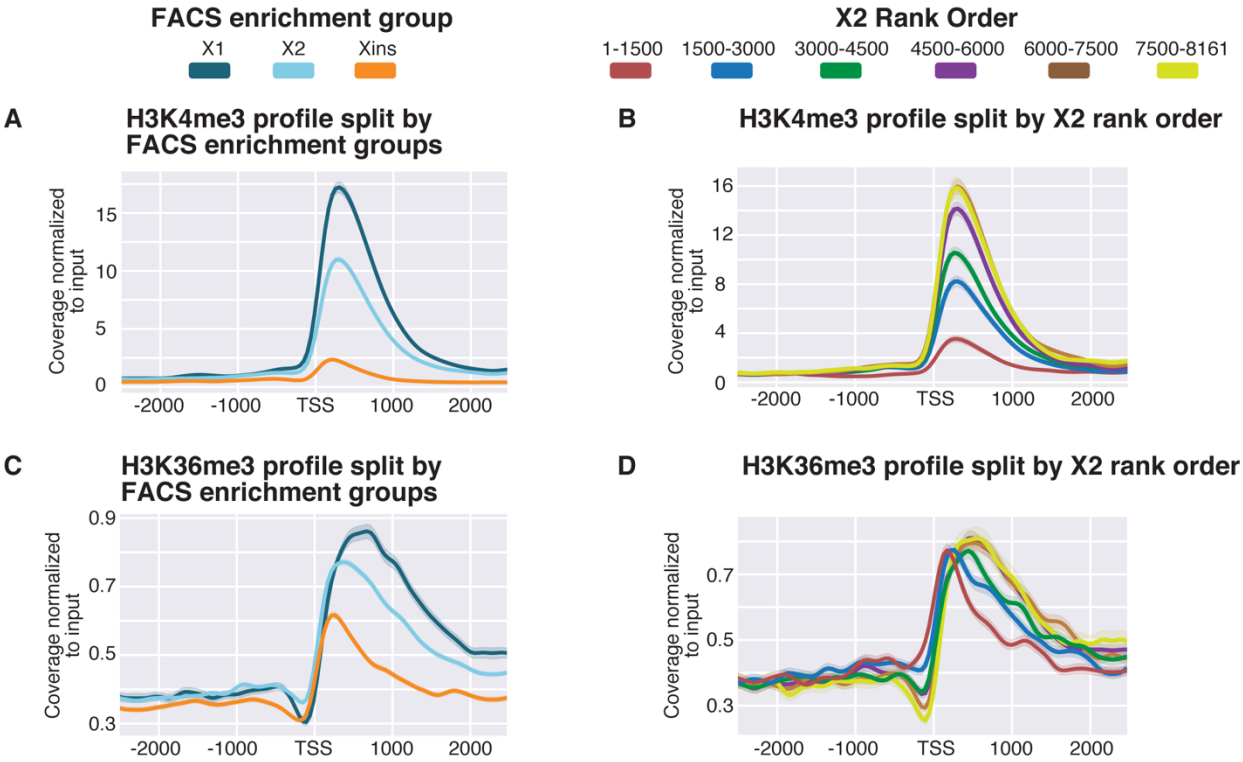


Figure 4. Histone marks for actively transcribed genes in X1 NBs. **A**. Average H3K4me3 ChIP-seq coverage profiles across X1-, X2-, and Xins-enriched loci in X1 NBs across biological replicates following outlier removal. The y-axis represents the difference in coverage between sample and input, and the x-axis represents 2.5 kb upstream of and downstream from the TSS. Shaded area around line is representative of the confidence interval for mean ChIP-seq signal. H3K4me3 signal is highest around the promoter-proximal region close to the TSS for X1-enriched loci in NBs consistent with the role of H3K4me3 in active transcription. **B.** H3K4me3 ChIP-seq profiles following outlier removal for X2 genes ranked from high to low X2 proportional expression. H3K4me3 signal in NBs decreases with an increase in proportion of X2 expression, indicative of high-ranking X2 genes having a predominant role in post-mitotic progeny as opposed to NBs. **C**. Average H3K36me3 ChIP-seq profile across X1-, X2-, and Xins-enriched loci in X1 NBs across biological replicates following outlier removal. The y-axis represents the difference in coverage between sample and input, and the x-axis represents 2.5 kb upstream of and downstream from the TSS. Shaded area around line is representative of the confidence interval for mean ChIP-seq signal. H3K36me3 signal is promoter-proximal for Xins genes, whereas the magnitude of signal is greater and shifted 3′ for X1 genes. **D**. H3K36me3 ChIP-seq profiles following outlier removal for X2 genes from high to low X2 proportional ranking. H3K36me3 signal in NBs shifts to the 3′ end with a decrease in X2 proportion, consistent with these lowly ranked genes having transcriptional activity in NBs.

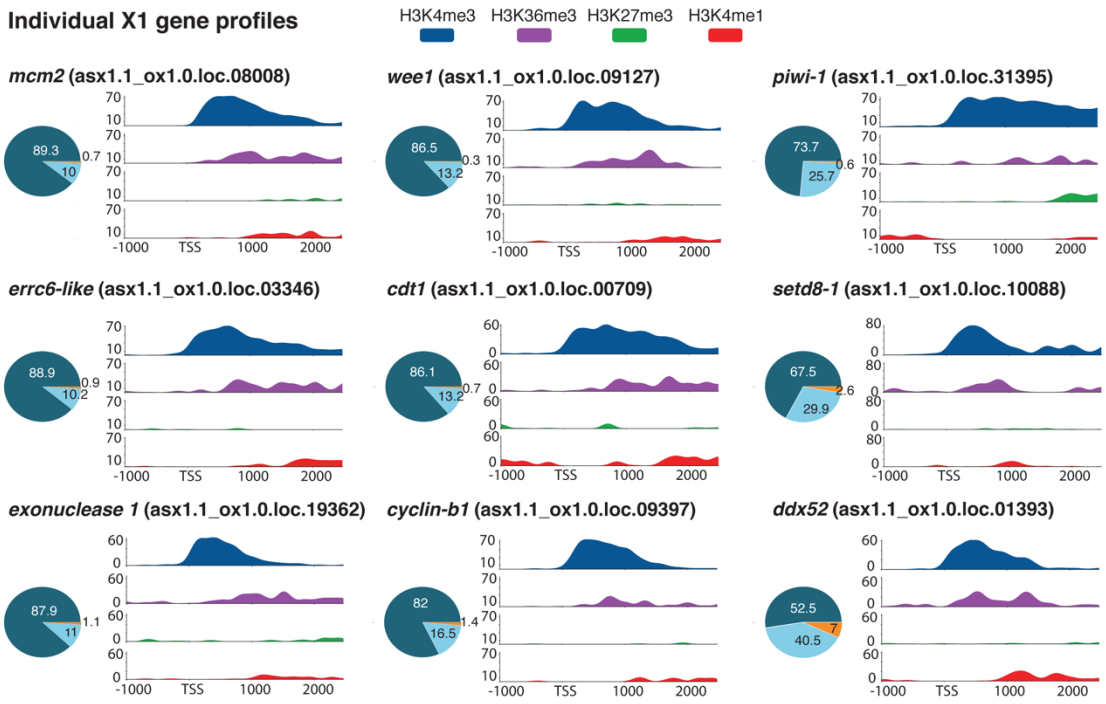


Figure 5. H3K4me3 and H3K36me3 (active marks) and H3K4me1 and H3K27me3 (suppressive marks) ChIP-seq profiles for highly expressed X1 genes in NBs. The y-axis represents percentage coverage for each mark and allows for the four epigenetic marks to be directly compared. The x-axis represents 1.0 kb upstream of and 2.5 kb downstream from the TSS. Pie charts represent proportional expression for each gene in X1 (dark blue), X2 (light blue), and Xins (orange).

4.4 H3K27me3 and H3K4me1 levels at the TSS anti-correlate with gene expression

Utilising our optimized ChIP-seq protocol, we investigated the occurrence of two additional histone modifications: H3K27me3, a repressive promoter mark catalysed by the PRC2 complex, and H3K4me1, a mark mediated by the MLL3/4 family of histone methyltransferases that correlates both with active enhancers and inactive promoter regions (Calo and Wysocka 2013; Cheng et al. 2014).

Genes that are categorized as being X1 enriched have low levels of H3K27me3 deposition at the TSS, compared with Xins enriched genes which are silenced in NBs (**Figure 6A**). A positive correlation is observed between the level of H3K27me3 and expression in the Xins compartment in a window from the TSS to 1kb downstream. This fairly broad domain of H3K27me3 deposition is consistent with previous studies in mammals (Hawkins et al. 2011; Pauler et al. 2009). Conversely,

a negative correlation at the TSS is observed between H3K27me3 signal and genes with high X1 expression (**Figure 8B**). Consequently, the genome wide pattern for H327me3 is the opposite to that observed for H3K4me3. When splitting X2 genes by rank we note that genes with higher transcriptional enrichment in the post-mitotic compartment have a higher overall level of H3K27me3 at the promoter proximal region compared to genes that have G1 NB expression (**Figure 6B)**.

The distribution of the H3K4me1 mark is noticeably different compared to that observed for either H3K27me3 or H3K4me3. Specifically, Xins loci have high levels of H3K4me1 at the TSS in X1 NBs, consistent with these genes being transcriptionally silent or expressed at low levels in NBs, whereas X1 loci have H3K4me1 peaks that are on average -1kb downstream of the TSS **(Figure 6C)**. This data suggests that the H3K4me1 signal shifts away from the TSS for genes that are actively expressed in NBs, in agreement with previous observations in mammals (Cheng et al, 2014). Further evidence of this peak shifting comes from analysis of X2 enriched genes sorted by rank order of expression (**Figure 6D**). Highly ranked X2 genes are marked with H3K4me1 at the promoter-proximal region. As the proportion of X2 enrichment decreases, indicative of increasing expression in the G1 NB compartment, the average H3K4me1 profile becomes bimodal, eventually shifting downstream of the TSS.

We plotted the epigenetic profiles of individual genes known to have high Xins proportional expressions and that have validated expression patterns both by single-cell RNA sequencing data and *in situ* hybridisations (**Figure 7**) (Fincher et al. 2018; Plass et al. 2018). These genes are expressed almost exclusively in the muscle (*COL21A1, slit1*), parenchyma (*glipr1, tolloid-like 1*), cathepsin+ cells (*dd961, aquaporin 1*), non-ciliated neurons (*tph*, *dd8060*), and protonephridia (*Na/Ca exchanger-like*), and all have high H3K27me3 signal at the TSS consistent with these genes being silenced in NBs. Moreover, these Xins enriched genes all have a high H3K4me1 signal at the TSS that anti-correlates with H3K4me3 deposition, in support of an earlier hypothesis that H3K4me1 limits the role of H3K4me3 interacting proteins (Cheng et al. 2014). We also observe an atypical

placement of H3K36me3 at the TSS of individual Xins genes which supports the previous suggestion that that H3K36me3 may silence loci when placed at a promoter-proximal region (Wu et al. 2011).

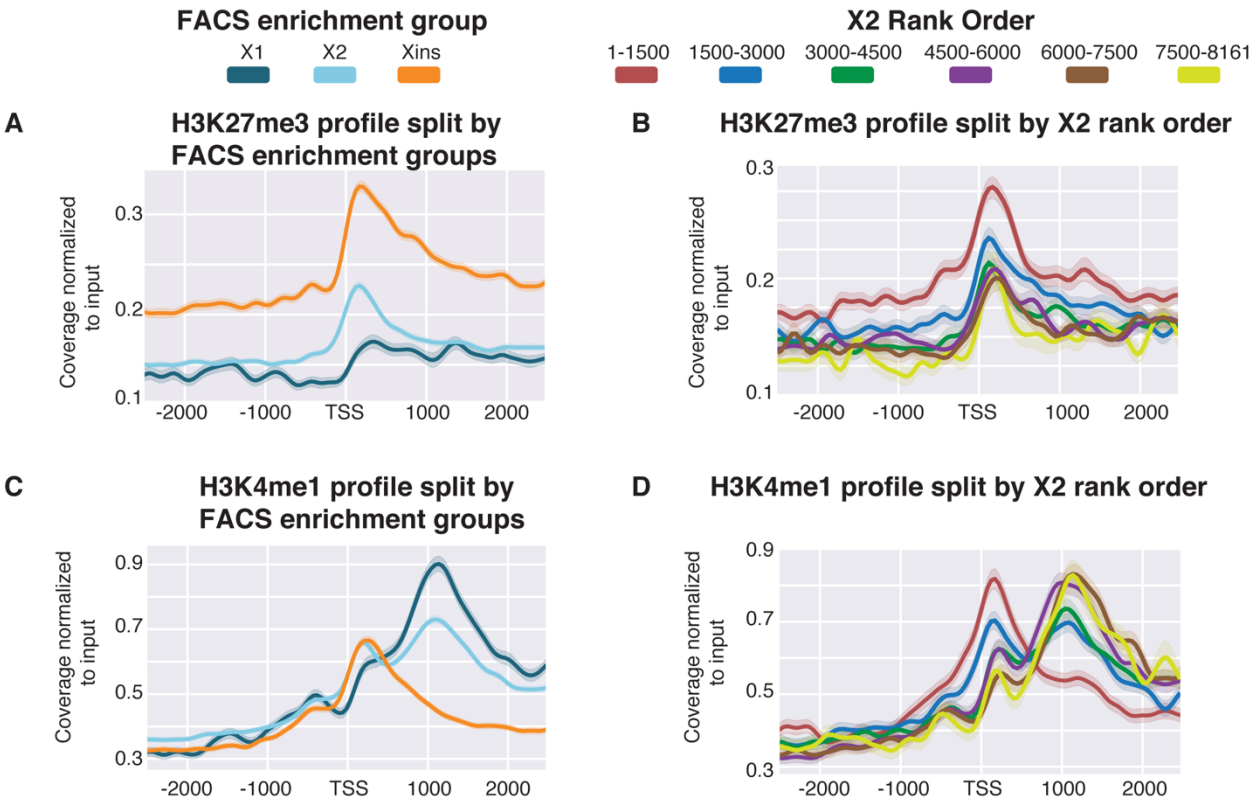


Figure 6. Histone marks for inactive genes in X1 NBs. **A.** Average H3K27me3 ChIP-seq profile across X1-, X2-, and Xins-enriched loci in X1 NBs across three biological replicates following outlier removal. The y-axis represents the difference in coverage between sample and input, and the x-axis represents signal 2.5 kb upstream of and downstream from the TSS. Shaded area around line is representative of the confidence interval for mean ChIP-seq signal. **B.** H3K27me3 ChIP-seq profiles following outlier removal for X2 genes from high to low X2 proportional ranking. H3K27me3 signal increases with an increase in proportion of X2 gene expression, indicative of these high-ranking X2 genes being transcriptionally silenced or lowly expressed in NBs. **C.** Average H3K4me1 ChIP-seq profiles following outlier removal across X1-, X2-, and Xins-enriched loci in X1 NBs. The y-axis represents the difference in coverage between sample and input, and the x-axis represents signal 2.5 kb upstream of and downstream from the TSS. Shaded area around line is representative of the confidence interval for mean ChIP-seq signal. **D.** H3K4me1 ChIP-seq profiles following outlier removal for X2 genes from high to low X2 proportional ranking. Highly ranked X2 genes have a H3K4me1 signal at the promoter-proximal region, and a decrease in X2 ranking coincides with a peak shift −1 kb downstream from the TSS.

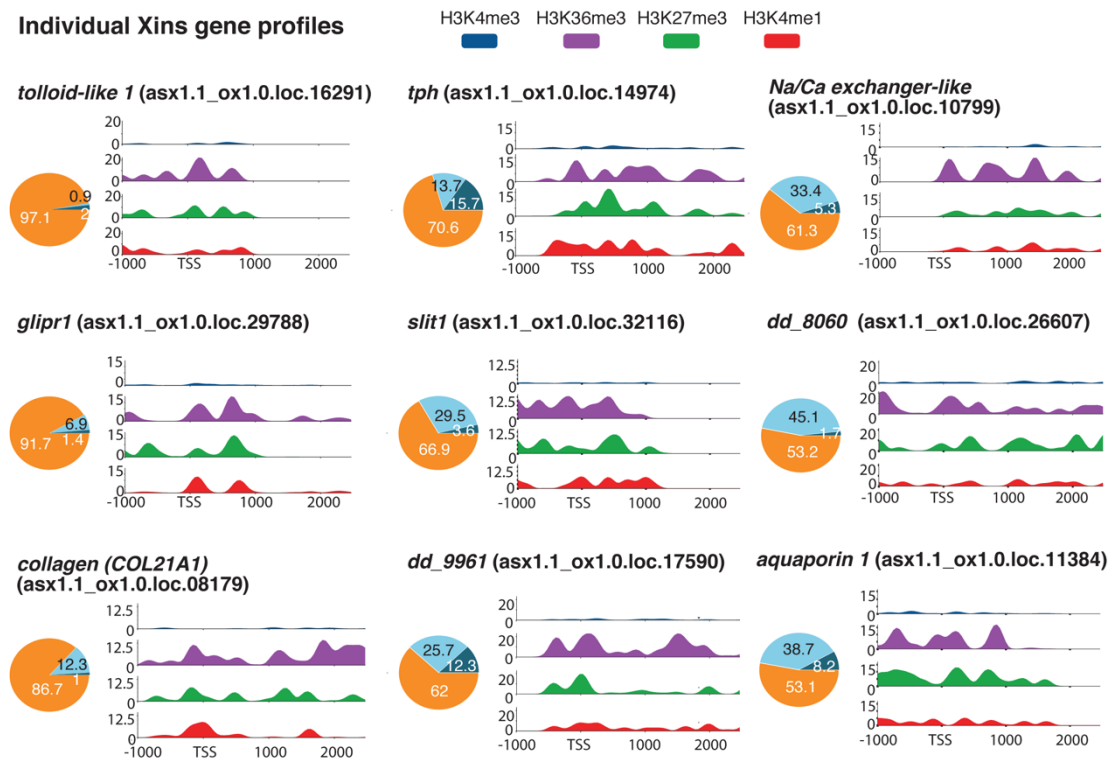


Figure 7. H3K4me3, H3K36me3, H3K4me1, and H3K27me3 NB ChIP-seq profiles for highly expressed Xins genes. The y-axis scale represents percentage coverage for each mark, and the x-axis represents 1.0 kb upstream of and 2.5 kb downstream from the TSS. Pie charts represent proportional expression for each gene in X1 (dark blue), X2 (light blue), and Xins (orange).

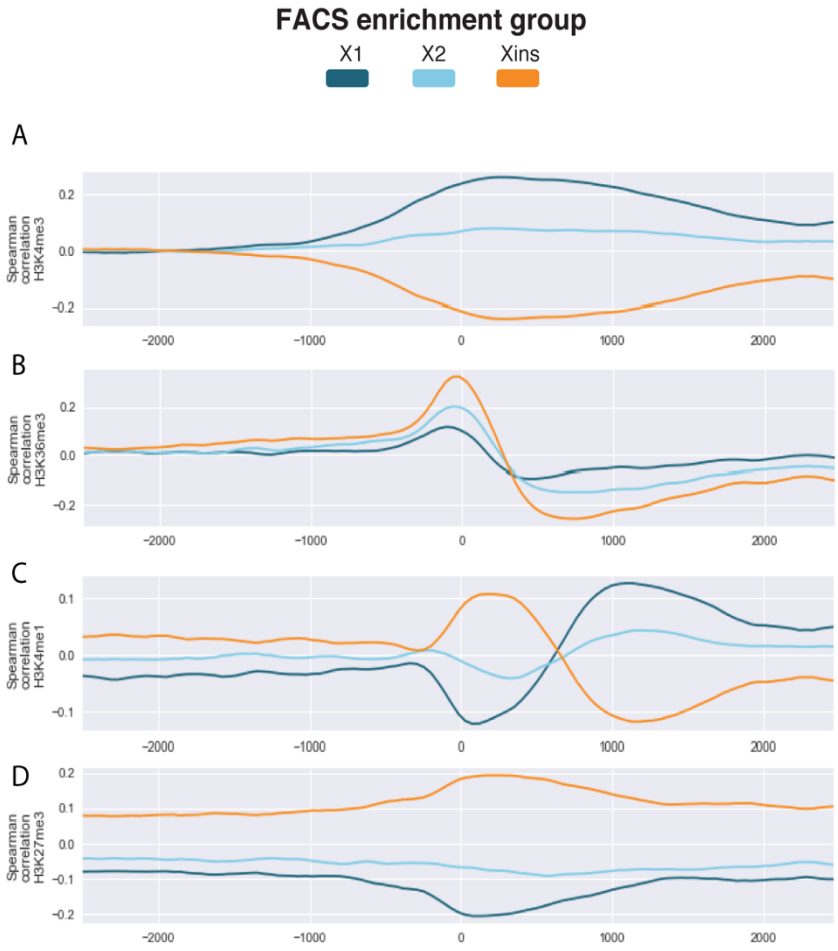


Figure 8: Spearman's rank correlation coefficient plots between ChIP-seq signal at a particular 50bp window around the TSS and the proportional expression value in X1, X2, and Xins**.** A positive correlation value means that the higher the ChIP-seq signal, the higher the proportional expression value. A negative correlation means that the lower the ChIP-seq signal, the higher the proportional expression value. **A.** A high ChIP-seq signal for H3K4me3 at the TSS correlates with higher X1 proportional expression, whilst genes with high Xins proportional expression have a low ChIP-seq signal for H3K4me3 at the TSS. Genes with high X2 proportional expression also, on average, have H3K4me3 signal at the TSS, but the correlation is weaker than that for high X1 genes. **B.** The profile for H3K36me3 shows that the highest signal for this mark correlated with high Xins proportional expression. As you move downstream of the TSS there is little, or negative correlation, between proportional expression and this mark (but we note that X1 and H3K6me3 correlation coefficient is higher compared with Xins and H3K36me3 correlation downstream of the TSS). This may be due to variations in exon length between genes resulting in a non-uniformly distributed H3K36me3 downstream of the TSS, thus leading to a weak correlation between X1 proportion and H3K36me3 signal. **C**. The profile for H3K27me3 indicates that genes with a high Xins proportional expression have, on average, a high H3K27me3 signal at the TSS. Genes with a high X1 or X2 proportional expression have a lower H3K27me3 signal at the TSS. **D.** The profile for H3K4me1 indicates that higher Xins proportional expression correlates with a high H3K4me1 signal at the TSS. Genes with higher X1 proportional expression are, on average, enriched for H3K4me1 downstream of the TSS. We observe a weak negative correlation for X2 expression and H3K4me1 at the TSS, and a weak positive signal between X2 expression and H3K4me1 downstream of the TSS. This is consistent for the X2 compartment being an admixture of genes that are almost exclusively expressed in the X2 compartment and genes that retain NB expression.

4.5 Correlations of H3K27me3 and H3K4me3 profiles against FACS proportions provide evidence for promoter bivalency in NBs

Having demonstrated that known active and suppressive marks correlate with gene expression in planarian NBs, we investigated whether promoter bivalency could act to keep genes in a poised state prior to the onset of differentiation. Bivalent promoters are characterized by the presence of both the activating mark H3K4me3 and repressive mark H3K27me3. The simultaneous presence of both these marks keeps the gene in a poised transcriptional state, with low or no expression, and upon differentiation resolves such that only one of the two marks is dominant. We reasoned that loci that are off or have relatively low proportional expression in X1 NBs, but which are upregulated during the differentiation process in post-mitotic progeny (high X2 expression), would be good candidates for potential regulation by bivalent promoters in NBs. Additionally, in the absence of sequential or co-ChIP-seq technologies for planarians, using genes with no or very low expression in NBs greatly reduces the likelihood that any bivalent signals are due to cell heterogeneity. This is because these genes would not be expected to have high levels of H3K4me3 in any (or at least very few cells) in the X1 NB compartment.

We plotted the percentage of maximum coverage for both H3K4me3 and H3K27me3 for the top 1000 genes for each of the three FACS enrichment categories (**Figure 9A-C**). A plot for the top 1000 X1 genes shows that these genes have a higher level of H3K4me3 compared to H3K27me3 (**Figure 9A**), whereas the top 1000 Xins genes have on average a much higher H3K27me3 signal compared to H3K4me3 (**Figure 9C)**. Consistent with our hypothesis, the top 1000 X2 genes, have, on average, peaks that are of similar magnitude for both of these functionally opposing epigenetic marks (**Figure 9B)**.

We also plotted the epigenetic profiles of genes that are downregulated following RNAi of the planarian homolog of the RNA-binding protein MEX3 (Zhu et al. 2015). Previously, *mex3-1* has been shown to be necessary for generating the differentiated cells of multiple lineages, and consistent

with a role in the differentiation process we found that the downregulated genes (downregulated 2-fold; p-value <= 0.05) had a higher average X2 proportional expression value (62.4%) compared with that of X1 (12.5%). As expected, we found a paired H3K4me3 and H3K27me3 ChIP-seq signal for these *mex3-1* downregulated genes (**Figure 9D)**.

One possibility is that our observations are as a result of some highly ranked X2 genes having only the H3K4me3 mark, whereas other genes exist in a H3K27me3-only state in NBs. This would produce an average profile that appears bivalent when many genes are looked at simultaneously. To account for this possibility, we plotted the distribution of Pearson correlation coefficients between H3K4me3 and H3K27me3 for the top 500 ranked X1, X2 and 285 *mex3-1(RNAi)* downregulated loci. This showed a strong positive correlation between H3K4me3 and H3K27me3 for top 500 X2 loci and *mex3-1* downregulated loci, compared to a weak or no average correlation for X1 loci (**Figure 9E**). This is consistent with the interpretation that bivalency is present at promoters of genes that are highly enriched for expression in the X2 compartment.

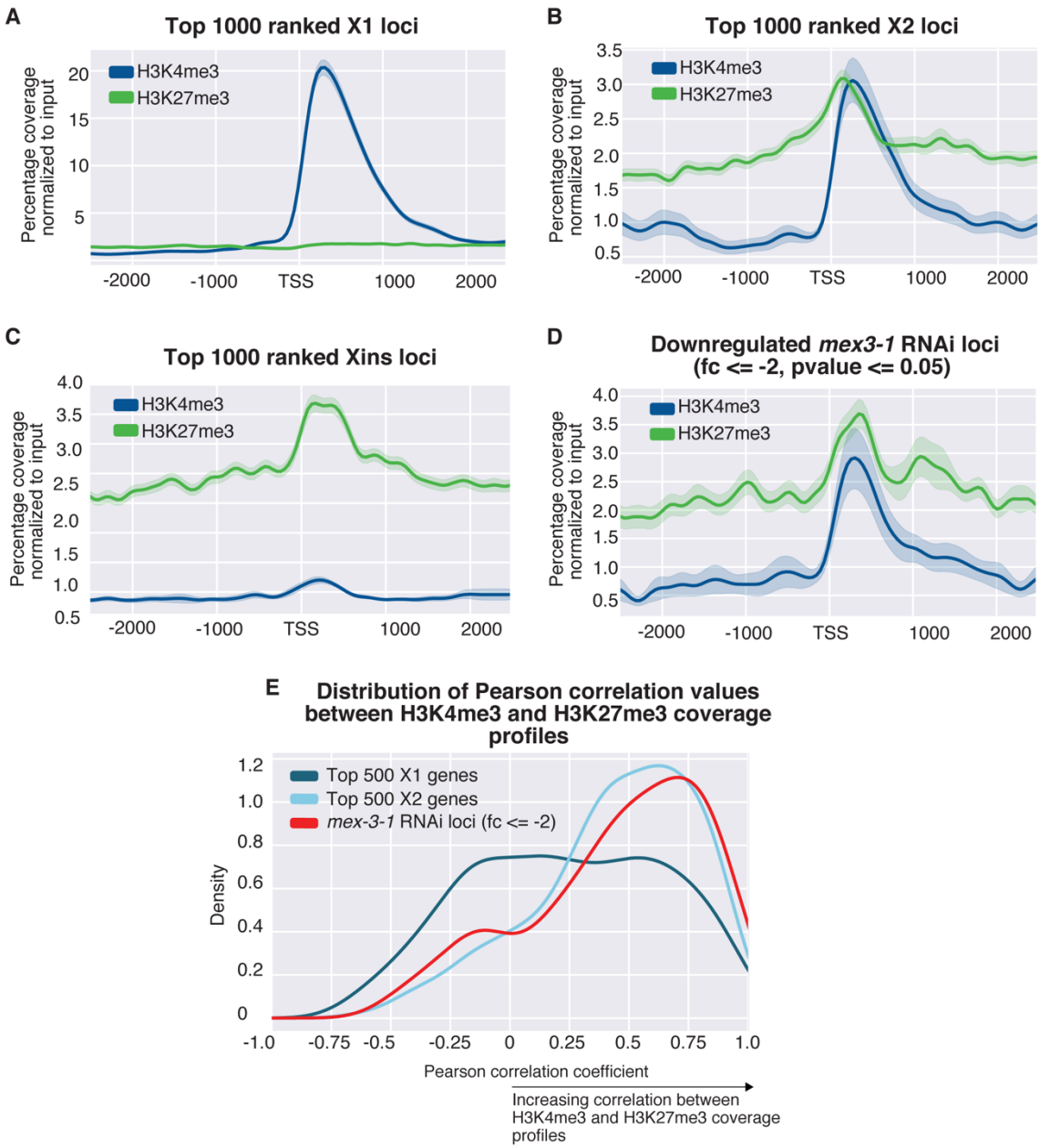


Figure 9. A–E. Average H3K4me3 and H3K27me3 ChIP-seq profiles in X1 NBs across three biological replicates. The y-axis is percentage coverage after normalization to input to allow both ChIP-seq profiles to be directly compared. Shaded area around line is representative of the confidence interval for mean ChIP-seq signal as a percentage coverage.Plots are shown for the following: **A.** top 1000 ranked X1 genes by expression; **B.** top 1000 ranked X2 genes; **C.** top 1000 ranked Xins genes; and **D.** 285 *mex3-1* down-regulated loci with greater than twofold change ($P < 0.05$). (E) A distribution of Pearson correlation values for the top 500 X1 expressed loci, the top 500 X2 expressed loci, and 285 loci ≥ twofold down-regulated after mex3-1(RNAi). The Pearson correlation coefficient was calculated between the H3K4me3 and H3K27me3 values at each 50-bp window −1000 bp and +1500 bp around the TSS.

4.6 Planarian orthologues to mammalian bivalent genes are marked by H3K4me3, H3K27me3 and paused RNA Pol II at the promoter-proximal region

RNA Polymerase II (RNAPII) pausing at genes that are highly inducible has been hypothesized to play a pivotal role in preparing genes for rapid induction in response to environmental or developmental stimuli. In a number of mammalian cellular contexts, bivalent genes have been shown to have a high density of paused RNA Pol II at the promoter-proximal region compared to genes which are actively transcribed, therefore allowing genes to be maintained in a transcriptionally poised state (Stock et al. 2007; Ferrai et al. 2017; Liu et al. 2017). Paused RNA Pol II can be distinguished from other forms by a phosphorylation at Ser5 (Ser5P) of the YSPTSPS heptad repeat at the C-terminus of the largest subunit of the Pol II complex. This heptad repeat is conserved across metazoans, and is found in *S. mediterranea* (Corden 2013; Yang and Stiller 2014).

ChIP-seq for RNAPII-Ser5P in NBs revealed that X2 enriched genes have a higher level of paused RNA Pol II at the promoter proximal region compared to X1 genes (**Figure 10A**). More significantly, highly ranked X2 genes with high expression in post-mitotic progeny and little expression in NBs have the highest amount of paused RNA Poll II close to the TSS, and with increasing expression in NBs the enrichment for this mark decreases (**Figure 10B**).

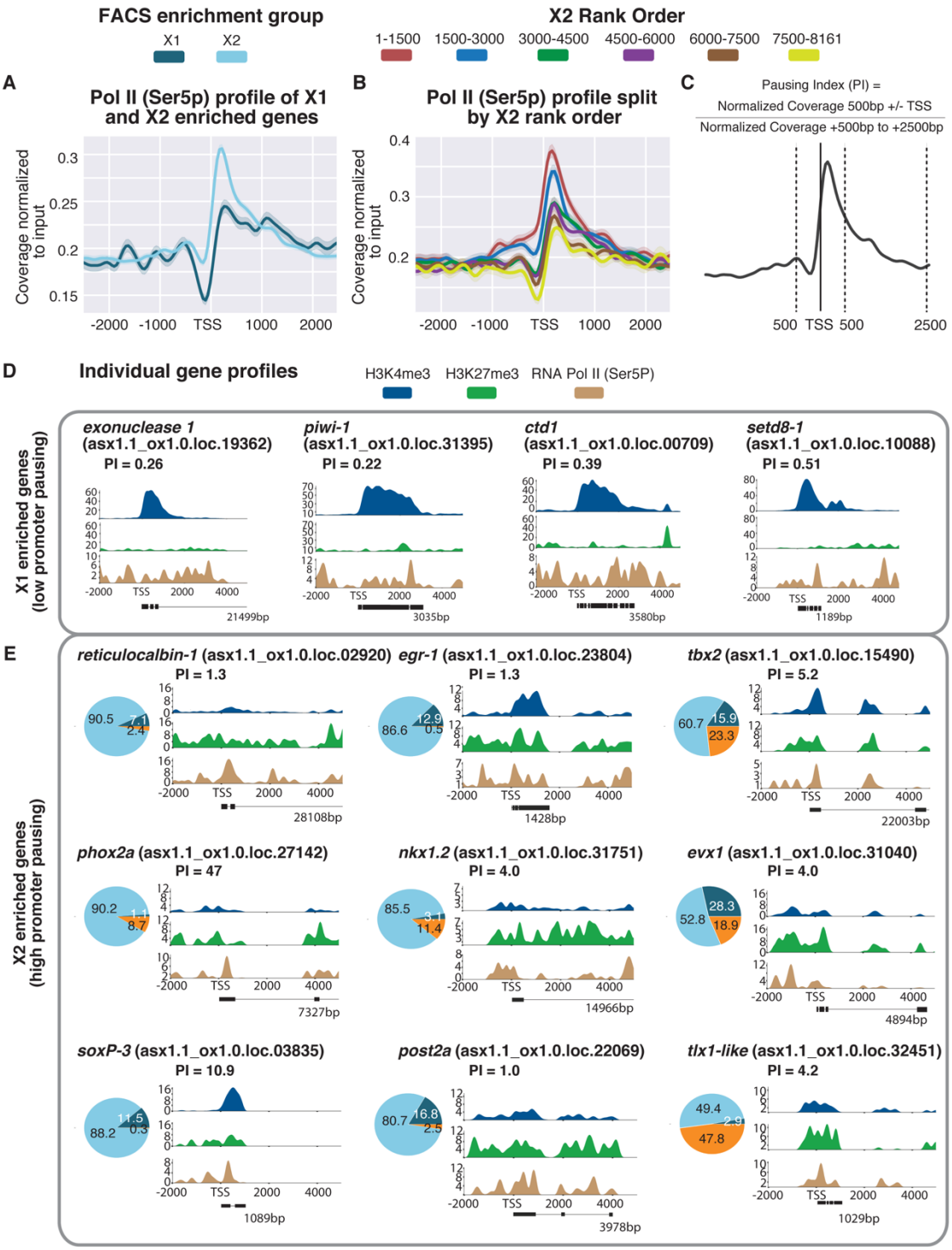
We calculated the pausing index (PI) for all annotated genes in our genome that have a total annotated length of ≥ 1kb. For our particular genome annotation, we calculated the PI as the read coverage (normalized to input) +/- 500bp either side of the annotated TSS divided by the normalized read coverage from +500bp to +2500bp from the TSS (**Figure 10C**). We defined a gene as being significantly stalled for transcription if the PI ≥ 1. As expected, individual genes highly expressed in the NB compartment had both a low PI score and were not enriched for RNAPII-Ser5P at the promoter-proximal region, thereby confirming our methodology was accurate at the gene level (**Figure 10D**). We also found that X2 genes with high PI scores had, on average, higher Pearson correlation coefficients between H3K4me3 and H3K27me3 (indicative of a bivalent state) compared

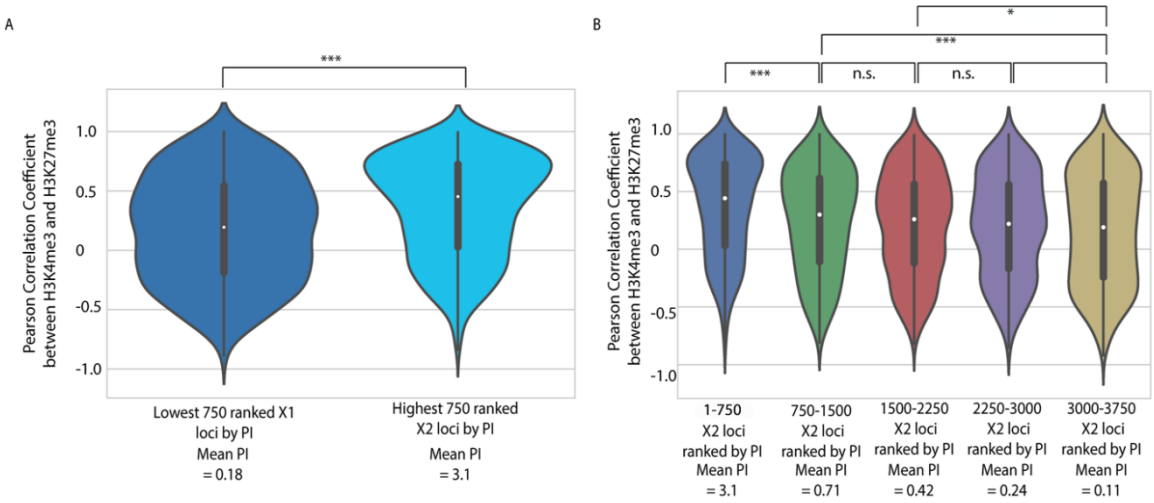
with both X1 and X2 genes that have lower PI scores (**Figure 11**). Given this correlation, we chose individual X2 enriched genes with high PI values and plotted the ChIP-seq profiles for H3K4me3, H3K27me3 and RNAPII-Ser5P as a percentage of maximum coverage for each mark.

Amongst genes enriched for these three signatures of bivalent promoters were those that have orthology to transcription factor (TF) families and include the Hox (*post2a*), Nkx (*nkx1-2*), Even-skipped (*evx-1*), Paired-like (*phox2A)*, T-box (*tbx2*) and Tlx (*tlx1-like*) gene classes (**Figure 10E**). Indeed, previous studies in both mouse ESCs (Bernstein et al. 2006) and quiescent muscle stem cells (Liu et al. 2013) have shown that members of these gene families are typically marked by both H3K4me3 and H3K27me3. A paired level of these marks at the TSS for these individual genes suggests the existence of bivalent chromatin states at these conserved developmental genes and confirms our correlational analysis of X2 loci (**Figure 9E**). Moreover, lineage trees made from psuedotemporally ordered cells in single cell RNA-seq sequencing data show that these genes are predominantly expressed in the post-mitotic cells of specific lineages, and is also consistent with few NBs having detectable levels of expression of these genes (**Figure 12**) (Plass et al. 2018).

One caveat of our analyses is that the bivalent profiles of X2 enriched differentiation related genes may, for some individual genes that appear bivalent, reflect admixture of transcriptionally active and repressed states within the X1 NB compartment. For example, previous work has shown that the X1 compartment is highly heterogeneous with subsets of *piwi-1*+ NBs expressing lineage specific TFs (Van Wolfswinkel et al. 2014). These genes, such as *soxP-3* and *egr-1,* which are in fact X2 enriched according to our dataset and others (Labbé et al. 2012), appear to have a paired H3K4me3 and H3K27me3 signal (**Figure 10E**). Given that they are known to be expressed in a subset of cells in the X1 compartment and are definitive markers of lineage-primed NB subsets that will go through one more cell division (as validated by *in situ* hybridisation, condensin knockdown studies (Lai et al. 2018; Van Wolfswinkel et al. 2014) and single-cell RNA-seq data (Wurtzel et al. 2015; Plass et al. 2018; Fincher et al. 2018) no definitive conclusions concerning bivalency of these particular genes can be reached.

Figure 10: **A**. Average paused RNAPII-Ser5P ChIP-seq profile across X1- and X2-enriched loci in X1 NBs across biological replicates following outlier removal. The y-axis represents the difference in coverage between sample and input, and the x-axis represents signal 2.5 kb upstream of and downstream from the TSS. Shaded area around line is representative of the confidence interval for mean ChIP-seq signal. **B**. RNAPII-Ser5P ChIP-seq profiles following outlier removal for X2 genes from high to low X2 proportional ranking. RNAPII-Ser5P signal increases with an increase in proportion of X2 gene expression, indicative of these high-ranking X2 genes being transcriptionally silenced but maintained in a permissive state for rapid induction. **C**. Calculation for pausing index (PI) of genes ≥1 kb. We divided normalized coverage between ±500 bp TSS by normalized coverage +500 bp to +2.5 kb. For genes <2.5 kb, we inspected RNAPolI-Ser5P profiles visually to confirm whether Pol II pausing was enriched at the promoter-proximal region. **D**. Individual profiles for H3K4me3 and H3K27me3 of highly enriched NB X1 genes. X1 genes have a high level of H3K4me3, and levels of H3K27me3 correspond to intron regions and are not enriched at the promoter-proximal region. RNAPII-Ser5P signal is not enriched at the promoter-proximal region compared with the gene body; as a result, PI <1. **E**. We selected highly enriched X2 genes with a PI ≥1 that have both H3K4me3 and H3K27me3 enriched at the promoter-proximal region, together with an enrichment of RNAPII-Ser5P close to the TSS. Pie charts represent proportional expression for each gene in X1 (dark blue), X2 (light blue), and Xins (orange).


FACS enrichment group
X1
X2
X2 Rank Order
1-1500
1500-3000
3000-4500
4500-6000
6000-7500
7500-8161
A
Pol II (Ser5p) profile of X1 and X2 enriched genes
Coverage normalized to input
0.3
0.25
0.2
0.15
-2000 -1000 TSS 1000 2000
B
Pol II (Ser5p) profile split by X2 rank order
Coverage normalized to input
0.4
0.3
0.2
-2000 -1000 TSS 1000 2000
C
Pausing Index (PI) =
Normalized Coverage 500bp +/- TSS
Normalized Coverage +500bp to +2500bp
500 TSS 500 2500
D
Individual gene profiles
H3K4me3
H3K27me3
RNA Pol II (Ser5P)
X1 enriched genes (low promoter pausing)
exonuclease 1 (asx1.1_ox1.0.loc.19362)
PI = 0.26
21499bp
piwi-1 (asx1.1_ox1.0.loc.31395)
PI = 0.22
3035bp
ctd1 (asx1.1_ox1.0.loc.00709)
PI = 0.39
3580bp
setd8-1 (asx1.1_ox1.0.loc.10088)
PI = 0.51
1189bp
E
X2 enriched genes (high promoter pausing)
reticulocalbin-1 (asx1.1_ox1.0.loc.02920)
PI = 1.3
90.5
7.1
2.4
28108bp
egr-1 (asx1.1_ox1.0.loc.23804)
PI = 1.3
86.6
12.9
0.5
1428bp
tbx2 (asx1.1_ox1.0.loc.15490)
PI = 5.2
60.7
15.9
23.3
22003bp
phox2a (asx1.1_ox1.0.loc.27142)
PI = 47
90.2
1.1
8.7
7327bp
nkx1.2 (asx1.1_ox1.0.loc.31751)
PI = 4.0
85.5
3.1
11.4
14966bp
evx1 (asx1.1_ox1.0.loc.31040)
PI = 4.0
52.8
28.3
18.9
4894bp
soxP-3 (asx1.1_ox1.0.loc.03835)
PI = 10.9
88.2
11.5
0.3
1089bp
post2a (asx1.1_ox1.0.loc.22069)
PI = 1.0
80.7
16.8
2.5
3978bp
tlx1-like (asx1.1_ox1.0.loc.32451)
PI = 4.2
49.4
2.9
47.8
1029bp



C

	Correlation	PI		Correlation	PI
ddx52	-0.45	0.16	evx-1	0.95	4.0
cyclinb-1	-0.3	0.07	phox2a	0.91	47
setd8-1	-0.15	0.51	soxP-3	0.81	10.9
mcm2	-0.11	0.25	tlx-like	0.76	4.2
exonuclease-1	-0.1	0.25	tbx-2	0.71	5.2
err6L	0.24	0.08	hoxb-9	0.68	1.0
wee1	0.40	0.90	nkx1.2	0.59	4.0
ctd1	0.40	0.39	egr-1	0.26	1.3
piwi-1	0.51	0.22	reticulocalbin-1	-0.3	1.3

Figure 11: A. Distribution of Pearson correlation coefficients between H3K4me3 and H3K27me3 for lowest 750 X1 enriched loci ranked by Pausing Index (PI) versus highest 750 X2 enriched loci ranked by PI. A Pearson correlation coefficient between H3K4me3 and HK27me3 percentage coverage at each 50bp window between TSS and 1.75kb upstream of TSS was calculated for each locus. A gene with high Pearson correlation coefficient would be, on average, more likely to be bivalent. X2 enriched loci with high PI scores, on average, have higher Pearson correlation coefficients, indicative of a bivalent state. Visual inspection of potential bivalent loci is further required to confirm a paired H3K4me3 and H3K27me3 signal at the TSS. **B**. X2 enriched loci (>1kb) split into rank order of PI. X2 enriched genes with high PI scores have, on average, higher Pearson correlation coefficients between H3K4me3 and H3K27me3. As PI decreases, so does the Pearson correlation coefficient. **C.** Individual Pearson correlation coefficients and Pausing Index scores for X1 enriched loci (as in Figure 1) and X2 bivalent loci (as in Figure 5). Again, on average, X2 loci have a higher Pearson correlation coefficient than X1 loci but individual gene ChIP-seq tracks must be inspected to verify paired H3K4me3 and H3K27me3 signal as this calculation does not take into account absolute magnitude of signal. For example, piwi-1 does not fit the expected trend owing to H3K27me3 signal in introns (resulting in a higher than expected Pearson correlation coefficient between H3K4me3 and H3K27me3) and reticulocalbin-1 has a broad H3K27me3 signal compared to H3K4me3 (resulting in a lower than expected Pearson correlation coefficient).

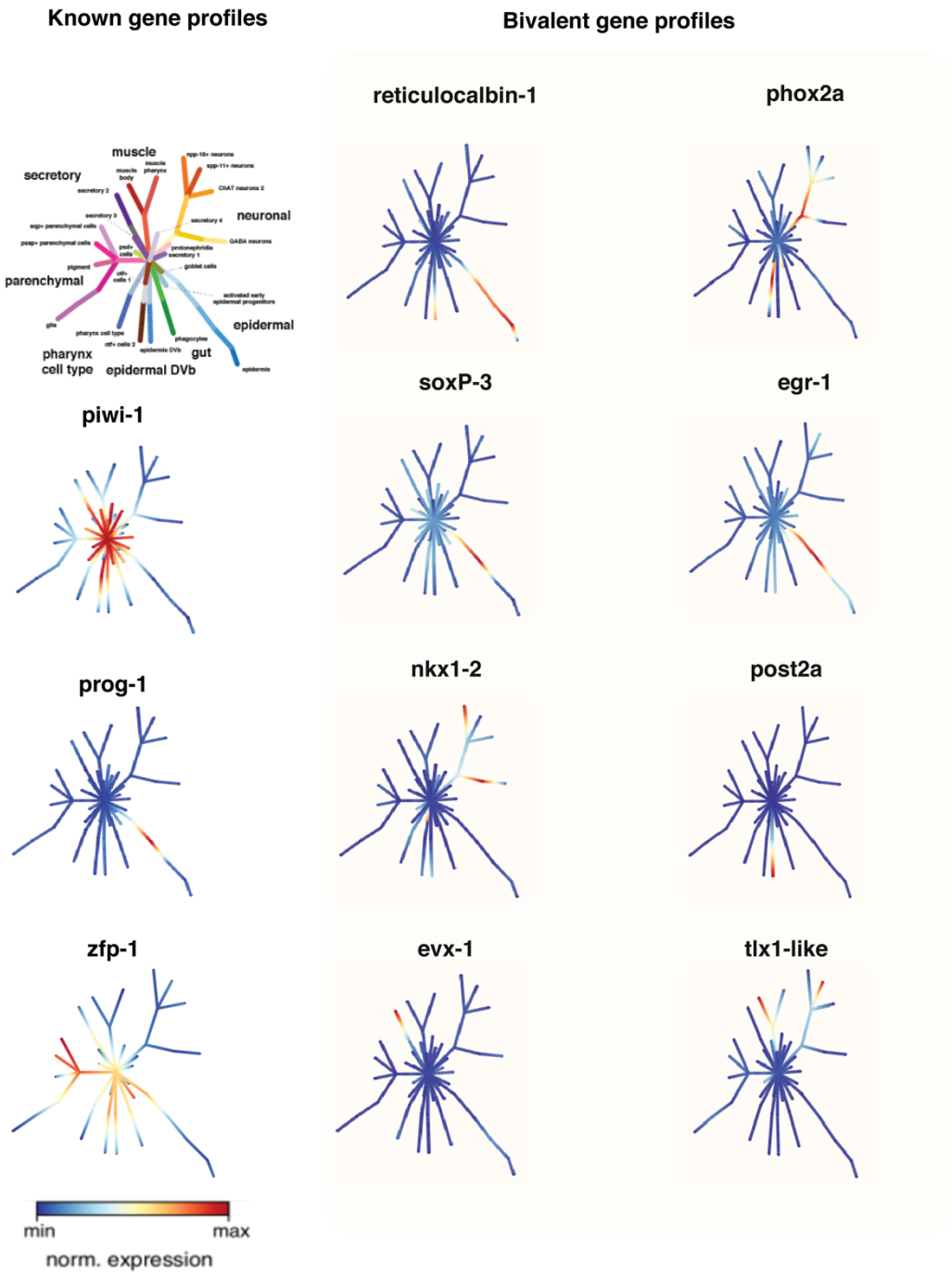


Figure 12: Single cell lineage trees (from https://shiny.mdc-berlin.de/psca/) of identified bivalent genes indicates that these genes are expressed in defined cellular lineages. These are similar profiles to other X2-enriched genes such as *prog-1* that were not found to have bivalent histone modifications in our analysis. In contrast, *piwi-1* is expressed predominantly at the core of the tree which is indicative of high NB expression. *zfp-1* (marker of ζ NBs) is also expressed at the core of the tree, indicative of high stem cell subset expression and a role in early NB lineage-commitment. *soxP-3* and *egr-1* are known to be expressed in epidermal-committed NBs, and as bivalent may reflect heterogeneity. In these lineage plots a small amount of normalized expression is detected at the core of the tree, but highest expression is in epidermal branch arms indicative of post-mitotic cells.

4.8 Discussion

In this chapter, we have outlined an optimized ChIP-seq protocol, and employed this to generate robust genome-wide profiles of the active H3K4me3 and H3K36me3 marks and repressive H3K4me1 and H3K27me3 marks in planarian NBs. We find that the active marks H3K4me3 and H3K36me3, and suppressive H3K4me1 and H3K27me3 marks in X1 NBs correlate with the proportion of total transcript expression of these loci in X1 cells, validating our NB ChIP-seq methodology.

Moreover, these analyses showed that genes associated with stem cell differentiation, and which are expressed at low levels in X1 population but activated at high levels in the X2 population, are marked with both H3K4me3 and H3K27me3 marks at comparable levels at the TSS. Moreover, these genes were also highly marked with paused RNA polymerase (RNAPII-Ser5P) at the promoter region consistent with the definition of transcriptionally poised bivalent genes. Although we cannot entirely rule out cell heterogeneity within the X1 NB population as a factor contributing to our observation of promoter bivalency, our focus on genes with high X2 expression (post-mitotic NB progeny) and orthology to vertebrate transcription factors known to have bivalent profiles provide strong evidence that bivalent histone marks are involved in poising of genes for activation upon NB commitment and differentiation in planarians.

The existence of promoter bivalency in invertebrates, prior to our work here, has been contentious. For example, the mammalian orthologues of bivalent genes in *Drosophila* germ cells were found to have only repressive H3K27me3 deposited at their promoters (Schuettengruber et al. 2009; Gan et al. 2010; Lesch et al. 2016). However, in a more recent study using fly embryos, the Pc-repressive complex 1 (PRC1) that binds to H3K27me3 was shown to co-purify with both Fsh1 (orthologue of mammalian BRD4) that binds to acetylated histone marks and to Enok/Br140 (orthologues to subunits of mammalian MOZ/MORF histone acetyltransferase complex). ChIP-seq identified two groups of PRC1/Br140 genomic binding sites that were either defined by strong H3K27me3 signal

or strong H3K27ac signal (i.e. actively transcribed genes). Both groups were also marked with narrow peaks of H3K4me3 at the TSS (Kang et al. 2017). These recent findings also argue for the existence of bivalent-like promoters outside of vertebrates, at least with respect to the binding of chromatin regulatory complexes, and extends the model to suggest that acetylation may be important in the resolution of bivalent protein complexes during development. Moreover, ChIP-seq of cercariae of the parasitic flatworm *Schistosoma mansoni* indicate that loci are marked by the bivalent H3K4me3/H3K27me3 signature. It was also shown that at this developmental stage there is very little active transcription, and that the animal possesses very few stem cells. Consequently, this study makes the case that bivalency does not have to be associated with pluripotency, and may be a general methodology to keep genes in a transcriptionally poised state before transitioning into the next developmental state (Roquis et al. 2015).

One key role of bivalent chromatin is to allow the maintenance of pluripotency in ESCs, by having genes involved in differentiation and commitment both silent but competent to switch on if the right signals are received. Our data suggest that this mechanism is likely to be important for pluripotency in planarian NBs, as genes that can switch on rapidly upon differentiation appear to be the bivalent. Indeed, these genes also included planarian orthologues to mammalian TFs that have been documented to be bivalent in ESCs. Consequently, we are able to present a case for promoter bivalency in planarian NBs and in doing so demonstrate that this process is not necessarily vertebrate-specific. This novel finding adds to the growing body of evidence which suggests a deep conservation of regulatory mechanisms involved in stem cell function (Solana et al. 2016; Juliano et al. 2010; Solana 2013; Alié et al. 2015; Lai and Aboobaker 2018) as well as combinatorial patterns of post-translational modifications (Schwaiger et al. 2014; Sebé-Pedrós et al. 2016; Gaiti et al. 2017). Epigenetic studies in the unicellular relative of metazoans, *Capsaspora owczarzaki* (Sebé-Pedrós et al. 2016) could not find any evidence of bivalency given the absence of H3K27me3, and epigenetic studies in the sponge *Amphimedon queenslandica* (Gaiti et al. 2017) and the cnidarian *Nematostella vectensis* (Schwaiger et al. 2014) have also not revealed any evidence for this approach to gene regulation. Further work will be required to establish when bivalent chromatin evolved in animals.

Overall, our development of a robust ChIP-seq protocol for use with planarian sorted NBs, together with good coverage for four definitive and essential epigenetic marks, establishes a resource for both future planarian studies investigating the epigenetic regulation of stem cell function as well as comparative epigenetic studies across metazoan phyla.

Chapter V

Deciphering the planarian stem cell regulome

Chapter V has not been reproduced elsewhere, and is a work in progress. Author contributions: Aziz Aboobaker and Anish Dattani conceived the project, which is still evolving. Anish Dattani and Divya Sridhar together prepared preliminary ATAC-seq libraries. Analyses were conducted by Anish Dattani.

Abstract

Transcriptomic approaches have allowed for the elucidation of genes involved in planarian NB pluripotency maintenance and subsequent differentiation into the various post-mitotic lineages. However, almost nothing is known about how transcription factor (TF) mediated cis-regulatory mechanisms control lineage commitment both under homeostatic and regenerative conditions. In the chapter, we document the repertoire of potential TFs in the *S. mediterranea* genome, and utilise FACS proportional categorization to find those enriched in NBs. By this methodology we identify TFs that have not been functionally characterized in planarians such as various Zinc Finger (ZNF-C2H2) TFs and potentially novel TFs derived from DNA transposable elements (TEs). We utilise ATAC-seq to identify regions of the genome in a state of open chromatin in the three FACS compartments. We find that the promoters of Xins-enriched genes become more accessible upon differentiation from the X1 NBs, consistent with their transcriptional activity in differentiated cells. Conversely, the promoters of X1-enriched genes remain accessible throughout differentiation. We also identify regions of the genome that are potential enhancers, and which correlate with the expression of proximally located genes. We summarize the experiments needed to be carried out to validate these findings, and strategies that can be employed to identify both the promoter targets and TF mediators of these enhancers.

5.1 Introduction

The binding of transcription factors (TFs) at cognate DNA motifs within enhancers can either positively or negatively influence the expression of a gene (Lambert et al. 2018; Lee and Young 2013). TF-bound enhancers are able to regulate the transcription of nearby or distant genes through physical contacts mediated by cohesin – a macromolecular complex that allows for the looping of DNA between enhancers and their target promoters (Kagey et al. 2010). TFs that enable transcriptional activation recruit conserved co-activators, such as p300 and the Mediator complex, that together form a preinitiation complex at enhancer/promoter junctions and recruit RNA Polymerase II (Juven-Gershon and Kadonaga 2010; Malik and Roeder 2010; Sikorski and Buratowski 2009). Once RNA Pol II initiates transcription, it typically pauses 20-50bp downstream of the TSS, and will transition to elongation following the recruitment of elongation factors and pause-release TFs, such as *c-Myc* (Adelman and Lis 2012; Rahl et al. 2010). Moreover, TFs and their associated complexes can also recruit histone modification and nucleosome remodelling complexes that either positively or negatively influence a gene's transcriptional environment (Nagaich et al. 2004; Boeger et al. 2008; Voss et al. 2011; Zaret and Carroll 2011)

The importance in precisely regulating gene expression for cellular and developmental processes is evident throughout all domains of life, but the exact TF repertoires from diverse organisms remains unchartered. Moreover, the genomic locations of specific TF enhancer targets are difficult to predict *in silico* for organisms where TF-DNA interactions have not been studied by ChIP-seq (Andersson et al. 2014). Recent alternative efforts to identify regulatory elements have utilised nucleases such as DNase I and Tn5 transposases to cleave nucleosome depleted regions (Giresi et al. 2007; Thurman et al. 2012; Stergachis et al. 2013; Buenrostro et al. 2015b). Moreover, these regulatory elements often have transcriptional expression as a result of the production of eRNAs (de Santa et al. 2010; Djebali et al. 2012; Murakawa et al. 2016; Andersson et al. 2014). Consequently, in different organisms, large repertoires of regulatory elements have been identified by determining the DNA

accessibility of different cell types and developmental stages. Association of enhancers with target promoters can be further inferred by their proximity to first exons, or alternatively can be assayed with Hi-C (Lieberman-Aiden et al. 2009; Schoenfelder et al. 2015a; Hughes et al. 2014; Dryden et al. 2014).

In systems where transgenesis is possible, validation of enhancer sequences can be carried out by enhancer-reporter constructs. For instance, ATAC-seq has been used in *C. elegans* to identify highly accessible chromatin regions enriched in TF motifs that modulate their accessibility between developmental stages and tissue types, and reporter constructs validated their tissue/developmental stage specific expression patterns (Daugherty et al. 2017). Similarly, enhancers specific to endothelial cells were identified in zebrafish using ATAC-seq on GFP+ nuclei of cells sorted from dissociated whole animals expressing a transgenic reporter specific to the endothelial compartment (Quillien et al. 2017). Evidence for putative endothelial enhancers was also provided on the basis of ATAC-seq peaks correlating with conserved enhancer histone epigenetic marks – such as H3K4me1 and H3K27ac – from whole zebrafish ChIP-seq data. Overall, both these studies suggest that multiple lines of independent evidence are necessary when validating the occurrence of enhancers in animal genomes.

In this chapter, we utilize the recent *S. mediterranea* genome assembly for the sexual biotype that has a N50 length of 3,854,845 bp compared with the asexual assembly that has an N50 of 77,506 bp (Robb et al. 2007; Grohme et al. 2018). Consequently, the higher contiguity of this assembly allows us to visualize genes and their regulatory elements in tandem on the same scaffold. We annotated this genome using both *de novo* transcriptomes and an assembly made in-house using publicly available RNA-seq datasets. Using this annotation, we identified proteins with DNA binding domains and characterized FACS proportional expression of genes belonging to major metazoan TF families. This enabled us to identify TFs that are enriched in the NB, post-mitotic, and differentiated cell compartments.

We next carried out ATAC-seq on the three FACS isolated populations of planarians – S/G2/M NBs (X1), stem cell progeny + G1 NBs (X2), and differentiated cells (Xins) - in order to identify regions of the genome that are in a state of open chromatin, and which are specific to the particular cell-types. We observed that between cell-types, genes that are active in the Xins population become more accessible at the promoter region following differentiation, consistent with their transcriptional upregulation between stem cell and differentiated state. However, stem cell (X1) genes did not ostensibly lose accessibility at the promoter region following differentiation, potentially indicative of these genes being transcriptionally permissive, albeit less active, in differentiated cells. We observed that putative enhancers are flanked by H3K4me1 – a mark known to be enriched at active and primed enhancers in other model systems (Calo and Wysocka 2013) - and make attempts to identify elements whose accessibility positively correlates with nearby gene expression.

5.2 Identification of TFs in the planarian genome

We sought to identify the repertoire of TFs in our expression-driven annotation of the sexual *S. mediterranea* genome. To generate this set of genomic annotations, we utilized a publicly-available platform Mikado, which analyses an ensemble of transcriptomic assemblies mapped to the genomic template and clusters overlapping transcripts based on genomic position (Venturini et al. 2018). In our case, we mapped *de novo* transcriptomes, known *S. mediterranea* gene sets (SmedGD UniGenes), and a transcriptome assembly made from 183 publicly available RNA-seq datasets to the sexual genome. Representative transcripts within these Mikado processed gene clusters were then picked using a scoring methodology based on ORF length (by TransDecoder (Haas et al. 2013)), BLAST homology with Uniprot metazoa, and splice-junctions (using Portcullis (Mapleson et al. 2017)). Valid alternatively spliced transcripts that are non-redundant with picked transcripts were also brought back to be included as isoforms. This process identified 29,038 genomic loci and 25,280 that potentially protein coding owing to an ORF length of =>100aa. This annotation process was carried out prior to the release of gene models for the sexual *S. mediterranea*, and as such comparisons between the two annotations should be made in the future (Rozanski et al. 2019).

We next conducted a systematic domain annotation of our 25,280 protein-coding loci using the InterProScan resource - which integrates a number of protein signature databases, including Pfam and SUPERFAMILY. We generated a list of potential TFs by searching our annotation list for Pfam terms identified by the DNA Binding Domain (DBD) v2.0 database (Wilson et al. 2008) . Moreover, we also identified SUPERFAMILY terms that were indicative of potential TF activity. By using both of these distinct Hidden Markov Model lists we identified 816 potential TFs in our genome annotation (**Figure 1A**). This non-redundant classification included 502 genes preferentially categorized by PFAM annotation, and a further 314 that had SUPERFAMILY annotation only. The distribution of proteins across both Pfam and SUPERFAMILY classes revealed that most belong to the 'homeobox-like domain (SUPERFAMILY)', 'Zinc Finger, C2H2 (Pfam)', 'homeobox domain (Pfam)', 'winged-helix (SUPERFAMILY)', and 'Myc-type, basic Helix Loop Helix (bHLH)'.

We next mapped RNA-seq data from publicly available FACS RNA-seq databases, and noted that only 682 out of 816 putative TFs had 10 reads or more mapping reads, with the remaining genes being likely pseudogenes or lowly expressed homeodomain-containing DNA transposable elements (TEs) (**Figure 1B**). Using our previously described methodology, we allocated a FACS proportional value to each potential TF. We noted that TFs were distributed in enrichment across all three cellular compartments, indicative of a role for TFs in maintaining NB pluripotency, driving lineage commitment, and maintaining the differentiated tissue-specific state.

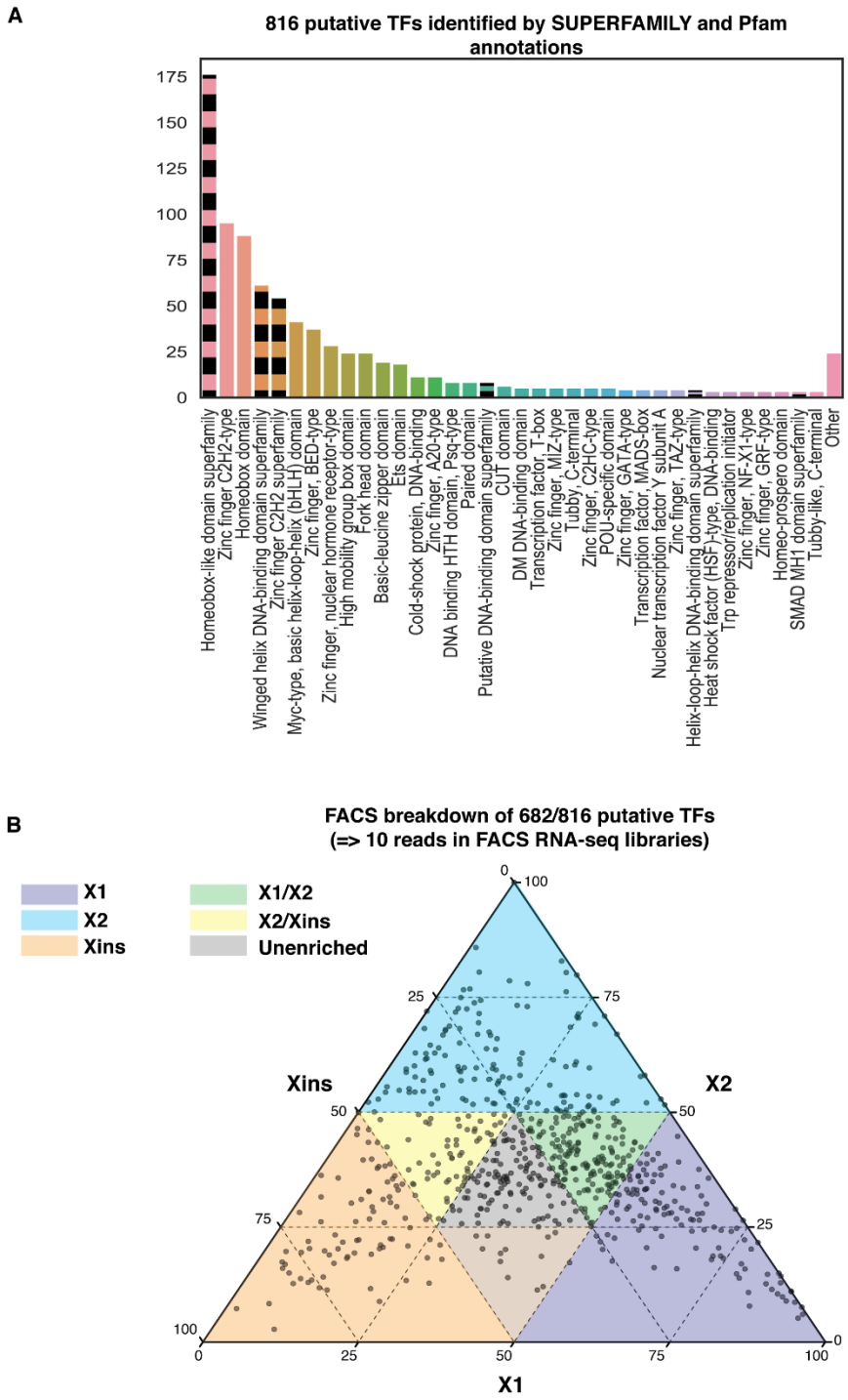


Figure 1: **A.** Number of genes in TF families identified in annotations of sexual *S. mediterranea* genome by InterPro utilising Pfam and selected SUPERFAMILY (Black stripes) IDs curated by the DBD database. Counts are non-redundant, such that PFAM annotations are given preference over SUPERFAMILY annotations in cases where genes are annotated by both. **B.** Ternary plot depicting FACS expression profiles for 682/816 TFs with => 10 reads in FACS RNA-seq libraries. Colours represent enrichment categories.

5.3 Identification of TFs enriched in planarian NBs

The contribution of TFs such as the Yamanaka factors (*Oct3/4, Sox2, Klf4, and c-Myc*) to the maintenance of pluripotency and self-renewal is well documented in mammalian ESCs. However, the role of TFs in the regulation of planarian NBs has received only minor attention. For instance, single-cell qPCR of X1 cells revealed that the σ NBs express two genes of the Sox/HMG box class of TFs (*soxP-1* and *soxP-2*) in addition to a SMAD family member *smad6/7-1* (Van Wolfswinkel et al. 2014). The *pbx* homeobox was also reported to be enriched in σ NBs compared to epidermal committed ζ NBs, but our proportional analyses from bulk FACS RNA-seq data indicates that this gene is enriched in the post-mitotic X2 class. Moreover, *pbx* has a well-established role in the establishment of planarian AP polarity (Blassberg et al. 2013; Chen et al. 2013). Likewise, the ζ NBs are characterized by a high expression of *zfp-1, p53, soxP-3* (Van Wolfswinkel et al. 2014). From these three TFs, only *zfp-1* is enriched in X1 NB compartment, with the other genes having maximal expression in the X2 compartment according to our proportional analyses. Similarly, γ NBs are distinguished by the expression of *nkx2.2-like*, *hnf4*, *gata4/5/6* and *prox-1*, from which only *prox-1* is enriched in the X1 NBs (Van Wolfswinkel et al. 2014) (**Figure 2**). Consequently, TFs that are expressed, but not enriched, in NBs likely have a predominant function in late lineage commitment, although we cannot rule out separate NB-specific functions for these genes.

The full repertoire of TFs enriched in NBs has never before been documented, and their roles in the maintenance of NB pluripotency and/or early lineage commitment are yet to be uncovered. In order to discover uncharacterized TFs with an enrichment of expression in planarian NBs, we manually curated a list of potential TFs with high X1 proportional expression. From an initial list of 119 TFs with => 50% X1 expression, we filtered down to 95 TFs by discarding genes whose orthologues are known to have DNA-binding functions that do not involve the regulation of transcription directly (**Figure 2**). For example, we removed *cdc6*, which has a predominant role in the maintenance of checkpoint mechanisms during cell cycle progression, and *topoisomerase III alpha,* known for reducing the number of DNA supercoils during transcription. Moreover, we noted numerous genes

with known functions in DNA replication and cell-cycle progression are also represented in a previous annotation of TFs in the planarian transcriptome as a result of their DNA-binding domains, and were not manually filtered to account for false-positives in this study (Swapna et al. 2018). It is also likely that our putative list of 816 TFs is an overestimation of the repertoire planarians, and further manual filtering of this list is needed for downstream functional analyses and phylogenetic comparisons of TF number with other organisms.

Within our X1 enriched TF list, we also included information, if available, as to the specific cell-types these TFs are expressed in according to the Drop-seq dataset by Fincher et al. 2018. This dataset includes transcriptomic information from 50,562 cells, which were Seurat clustered cells into 44 major groups that make up 9 distinct tissue types (NB, neural, intestine, muscle, epidermis, protonephridia, *cathepsin*+ cells, pharynx and muscle). From these initial major clusters, a further >150 sub-clusters were identified with distinct gene expression, that allowed for the identification of different specialized tissue types as well as lineage-committed *smedwi-1*+ NBs. Moreover, 9 potential NB sub-clusters that may be representative of the individually pluripotent cNeoblasts were identified owing to a high level of the *piwi-1* NB marker and exclusion of ζ and γ NBs (Fincher et al. 2018) (**Figure 3A**). Apart from 2 subclusters (subclusters 2 and 9), each subcluster can be distinguished by the specific expression of genes corresponding to the Dresden Transcriptome IDs: e.g. *dd_10988, dd_6998, dd_17796, SAMD15 (dd19710), dd_1122+, dd_13666*, and *PLOD1* (*dd3457*). Importantly, we cannot posit that all cells of these 9 subclusters will be cNeoblasts, and may in fact contain early precursors to lineage-committed NBs. It is likely, however, that these 9 NB sub-clusters will enrich for cNeoblasts compared to the NB population as a whole. Most of our identified X1-enriched TFs were not statistically enriched in any particular major planarian tissue type, likely owing to a low coverage for these TFs in individual single-cell libraries, and 14/95 of our TFs were in fact enriched in tissues other than NBs. For 26/95 TFs with enrichment in expression in NBs, we further looked for whether they were enriched in particular *piwi-1*+ sub-clusters. Interestingly, we found only 6/26 were enriched in one of the 9 potential cNeoblast clusters: *soxP-*

1, *soxP-2*, *runt-1, gcm-1,* an unknown ZNF gene (*Sxl_rink_2022*) and unknown bZIP-containing gene (*Sxl_rink_31340)* (**Figure 3**).

Our X1-enriched TF list uncovered a number of genes that are conserved with mammals, and which have not been functionally studied in planarians. For instance, one of two orthologues to the glial-cells missing (*gcm*) gene is highly enriched in the stem cells of planarians (X1). In Drosophila, *gcm* is known to function as a binary developmental switch between assuming glial or neuronal fate, whereas in mammals *gcm* genes are known to have a role in placental morphogenesis and development of the parathyroid gland (Hosoya et al. 1995; Jones et al. 1995; Basyuk et al. 1999; Wegner and Riethmacher 2001). In planarians, *gcm-1* is statistically enriched in the high *smedwi-1+* cluster *dd_10988* compared to other subclusters, but there are some cells outside of this compartment that also express this gene (**Figure 3C**). Given that this TF has subsumed a variety of functions through evolutionary time, RNAi of this gene in planarians may uncover a role in stem cell maintenance. Alternatively, the expression of this gene in the *dd_10988* compartment may indicate a role in early lineage commitment within high *piwi-1+* cells to glia (a recently discovered planarian cell type (Roberts-Galbraith et al. 2016; Wang et al. 2016)), similar to its role in flies.

Moreover, we identified a number of Zinc-finger C2H2 (C2H2-ZNF) containing genes, which have not been previously functionally characterized. Members of this gene family are known to function as transcriptional regulators, binding to DNA via their zinc-finger region; others, however are also known to bind to RNA, and their exact function is yet unknown (Theunissen et al. 1992; Grondin et al. 1996). In humans, C2H2-ZNF genes represent the largest group of TFs (747/1639) and alongside homeodomains (196/1639) represent more than half of known TFs (Lambert et al. 2018). Some of these C2H2-ZNF genes have well characterized functions in development and cell proliferation such as *Krox-20, snail, Gli1,* and *Krüppel-like* factors, but the vast majority have elusive functions that are only beginning to be uncovered. For instance, recently ZNF207 was shown to bind to the proximal enhancer of the pluripotency factor OCT4 human ESCs, and could also enhance reprogramming efficiency of neonatal fibroblasts to iPSCs (Fang et al. 2018). Importantly, the high

conservation of amino acids in C2H2-ZNF genes, as well as the repetitive nature of the C2H2-ZNF domain within a single protein, means that it is often difficult to work out orthologous relationships between genes from different animals, owing to a high baseline of sequence similarity (Knight and Shimeld 2001). Moreover, lineage-specific expansions as a result of tandem duplications are documented in mammals at least (Shannon 2003; Hamilton 2006; Tadepally et al. 2008) . For the C2H2-ZNF genes enriched in planarian NBs, we performed a BLAST to human, *Drosophila* and *C. elegans* in order to ascertain homology, if present. This provided identification for Sxl_rink_6833, which hits *bc11a* in mouse and known orthologues in *Drosophila* (CG9650) and *C. elegans* (*bc-11*). In the majority of cases, BLAST homology was uncertain, owing to low sequence similarity to proteins from all three organisms, and as such we assigned the gene as 'ZFN-unknown'. Molecular phylogenetic analyses is therefore necessary, using a large number of taxa, to ascertain where in metazoa these ZFN-unknown genes arose, and whether they are a unique expansion in flatworms.

Figure 2: List of TFs enriched (=> proportional 50% expression) in NBs. We also include predictions for lineage involvement using single-cell sequencing enrichment analyses from Fincher et al. (2018). We first checked for enrichment of genes in major clusters, representative of major planarian tissue types. Enriched genes were identified using both a receiver operating characteristic curve (ROCC) analysis and a likelihood ratio test (LRT) test based on zero-inflated data, thresholding for genes that show at least a 0.25 Log2fold average difference between all other clusters (Fincher et al. 2018). Where NB enrichment was observed for the major cluster, we ascertained enrichment if present in the subclusters of *smedwi-1*+ cells (given in red). For most genes, we do not observed an enrichment in any major cluster, which may be the result of very few cells expressing this gene and/or this gene is lowly expressed. Where orthology is not known for the Zinc Finger, C2H2 proteins, we have assigned this gene as "ZNF-unknown". DNA TEs are included in this list owing to 'homeodomain-like' domain hit. We highlight those TFs that have been tested for function with RNAi or have documented expression patters as assayed by *in situ* hybridisation.

Gene ID	BLAST hit	Domain	X1	X2	Xins	Major cell cluster, piwi+ subcluster
Sxl_rink_13565	seven-up	ZF,C4-type				N/A
Sxl_rink_26389	Cowles et al. (2013) hesl-1	bHLH				N/A
Sxl_rink_6123	Cowles et al. (2013) hesl-3	bHLH				N/A
Sxl_rink_13564	seven-up	ZF,C4-type				N/A
Sxl_rink_18380	glial-cells missing	GCM				piwi-1+, dd_10988+
Sxl_rink_31340	Unknown	bZIP				piwi-1+, dd_11796+
Sxl_rink_12027	Scimone et al. (2014) ZFP-2	ZF, C2H2-type				N/A
Sxl_rink_0590	EIF3C	winged-helix				N/A
Sxl_rink_16613	CCAAT-TF	CCAAT				N/A
Sxl_rink_33835	DNA TE	homeodomain-like				N/A
Sxl_rink_25031	Znf-Unknown	ZF, C2H2-type				N/A
Sxl_rink_1118	CREBBP	ZF, TAZ-type				N/A
Sxl_rink_31816	neuroD1	bHLH				N/A
Sxl_rink_31358	Cowles et al. (2013) hesl-2	bHLH				N/A
Sxl_rink_24148	Uknown	homeo,POU				N/A
Sxl_rink_8416	Znf-Unknown	ZF, C2H2-type				NB
Sxl_rink_1697	Znf-Unknown	ZF, C2H2-type				N/A
Sxl_rink_10870	RFX5	Winged-helix				N/A
Sxl_rink_7599	ZFP-CHHC-Uknown	ZFP,CHHC				NB
Sxl_rink_8777	Znf-Unknown	ZF, C2H2-type				piwi-1+, 15
Sxl_rink_16754	Cowles et al. (2013) sim-1	bHLH				Neural
Sxl_rink_16166	Znf-Unknown	ZF, C2H2-type				piwi-1+, zeta NBs
Sxl_rink_7275	Wenemoser et al. (2012) runt-1	RUNT				piwi-1+, dd_10988+
Sxl_rink_24482	rpa2	winged-helix				NB
Sxl_rink_6181	Wenemoser et al. (2012) six3-1	homeodomain				N/A
Sxl_rink_4293	Unknown	NR				N/A
Sxl_rink_1024	Scimone et al. (2014) twist	bHLH				Muscle
Sxl_rink_36711	YBOX4-like	Y-box				N/A
Sxl_rink_12086	TFIIE	winged-helix				N/A
Sxl_rink_37911	GATA3	GATA				N/A
Sxl_rink_8230	DNA TE	homedomain				N/A
Sxl_rink_32212	soxP-5	HMG box				piwi-1+, muscle
Sxl_rink_20259	YY1 TF	ZF, C2H2-type				N/A
Sxl_rink_14234	HMGB2	HMG-box				N/A
Sxl_rink_32069	van Wolfswinkel et al. (2014) ZFP-1	ZF,C2H2-type				piwi-1+, zeta/cathep
Sxl_rink_2022	Znf-Unknown	ZF,C2H2-type				piwi-1+, SAMD15
Sxl_rink_32835	Wagner et al. (2012) PROX-1	homeodomain				piwi-1+, gamma/intes
Sxl_rink_5158	HR46	ZF,C4-type				N/A
Sxl_rink_11751	nkx6-2	homeodomain				N/A
Sxl_rink_30626	DNA TE	homeodomain-like				NB
Sxl_rink_10428	musculin	bHLH				Muscle
Sxl_rink_30624	DNA TE	homeodomain-like				NB
Sxl_rink_6279	Wagner et al. (2012) soxP-1	HMG box				piwi-1+, 10988/neural1/2
Sxl_rink_7471	Unknown	bZIP				piwi-1+, N/A
Sxl_rink_35752	Wagner et al. (2012) soxP-2	HMG box				piwi-1+, dd_10988
Sxl_rink_14505	scratch-like	ZF,C2H2-type				Neural

Gene ID	BLAST hit	Domain	X1	X2	Xins	Major cell cluster, piwi+ subcluster
Sxl_rink_0878	IRX4	homeodomain				pharynx
Sxl_rink_11710	DMD3	DM				muscle
Sxl_rink_1484	XHOX-3-like	homeodomain				N/A
Sxl_rink_25579	YBOX2	CSD				All Lineages
Sxl_rink_23633	INSM1	ZF,C2H2-type				Neural
Sxl_rink_17944	Dorsal switch protein 1	HMG box				All Lineages
Sxl_rink_13833	DNA TE	homeodomain-like				N/A
Sxl_rink_1178	Gonzalez et al. (2012) SMAD6/7-1	SMAD				piwi-1+, 15
Sxl_rink_7014	Histone TF 4	ZF,C2H2-type				NB
Sxl_rink_35945	Znf-Unknown	ZF,C2H2-type				epidermal
Sxl_rink_25578	YBOX2	CSD				All Lineages
Sxl_rink_22167	Unknown	homeodomain-like				N/A
Sxl_rink_0654	Currie et al. (2016) zerknullt	homeodomain				N/A
Sxl_rink_24821	Lapan et al. (2011) sp6-9	ZF,C2H2-type				N/A
Sxl_rink_22361	DNA-TE	homeodomain-like				N/A
Sxl_rink_32142	Unknown	homeodomain-like				epidermal
Sxl_rink_24442	Unknown	homeodomain-like				N/A
Sxl_rink_12151	TFIIIB	homeodomain-like				NB
Sxl_rink_11380	Znf-Unknown	ZF,C2H2-type				piwi-1+, muscle/neural 2
Sxl_rink_15876	Unknown	homeodomain-like				piwi-1+, muscle
Sxl_rink_6833	Bcl11a	ZF,C2H2-type				Muscle/Neural
Sxl_rink_36813	TEAD1	TEA				N/A
Sxl_rink_6857	Monjo et al. (2015) soxB1-1	HMG box				N/A
Sxl_rink_9792	Tigger DNA TE	homeodomain-like				N/A
Sxl_rink_0371	nkx3-2	homeodomain				N/A
Sxl_rink_3076	Dmd-4	DM				N/A
Sxl_rink_12528	TFIIE	winged-helix				N/A
Sxl_rink_30445	GFI1	ZF,C2H2-type				N/A
Sxl_rink_4471	mitfl-3	bHLH				N/A
Sxl_rink_15355	Znf-Unknown	ZF,C2H2-type				NB
Sxl_rink_22195	Wagner et al. (2012) TCF15	bHLH				piwi-1+, zeta/musc/epid/cathep
Sxl_rink_38458	DNA TE	homeodomain-like				N/A
Sxl_rink_16235	POU5F1	homeodomain, POU				N/A
Sxl_rink_25587	YBOX1	CSD				All Lineages
Sxl_rink_12500	NK7	homeodomain				Neural/Epidermal
Sxl_rink_37376	hr39	ZF-C4-type				N/A
Sxl_rink_13385	DNA TE	homeodomain-like				N/A
Sxl_rink_11715	dorsal switch protein	HMG box				N/A
Sxl_rink_21310	dorsal switch protein	HMG box				piwi-1+, neural 1/2
Sxl_rink_36881	MTA-1	ZF,C2H2-type				N/A
Sxl_rink_9090	DNA TE	homeodomain-like				N/A
Sxl_rink_35950	Znf-Unknown	ZF,C2H2-type				N/A
Sxl_rink_32387	Unknown	homeodomain-like				N/A
Sxl_rink_34976	Znf 207	ZF C2H2-type				NB
Sxl_rink_8226	Znf-Unknown	ZF,C2H2-type				N/A
Sxl_rink_3267	ZMAT1	ZF,U1-like				N/A
Sxl_rink_36437	DNA TE	homeodomain-like				N/A
Sxl_rink_9684	nkx2-3	homeodomain				N/A

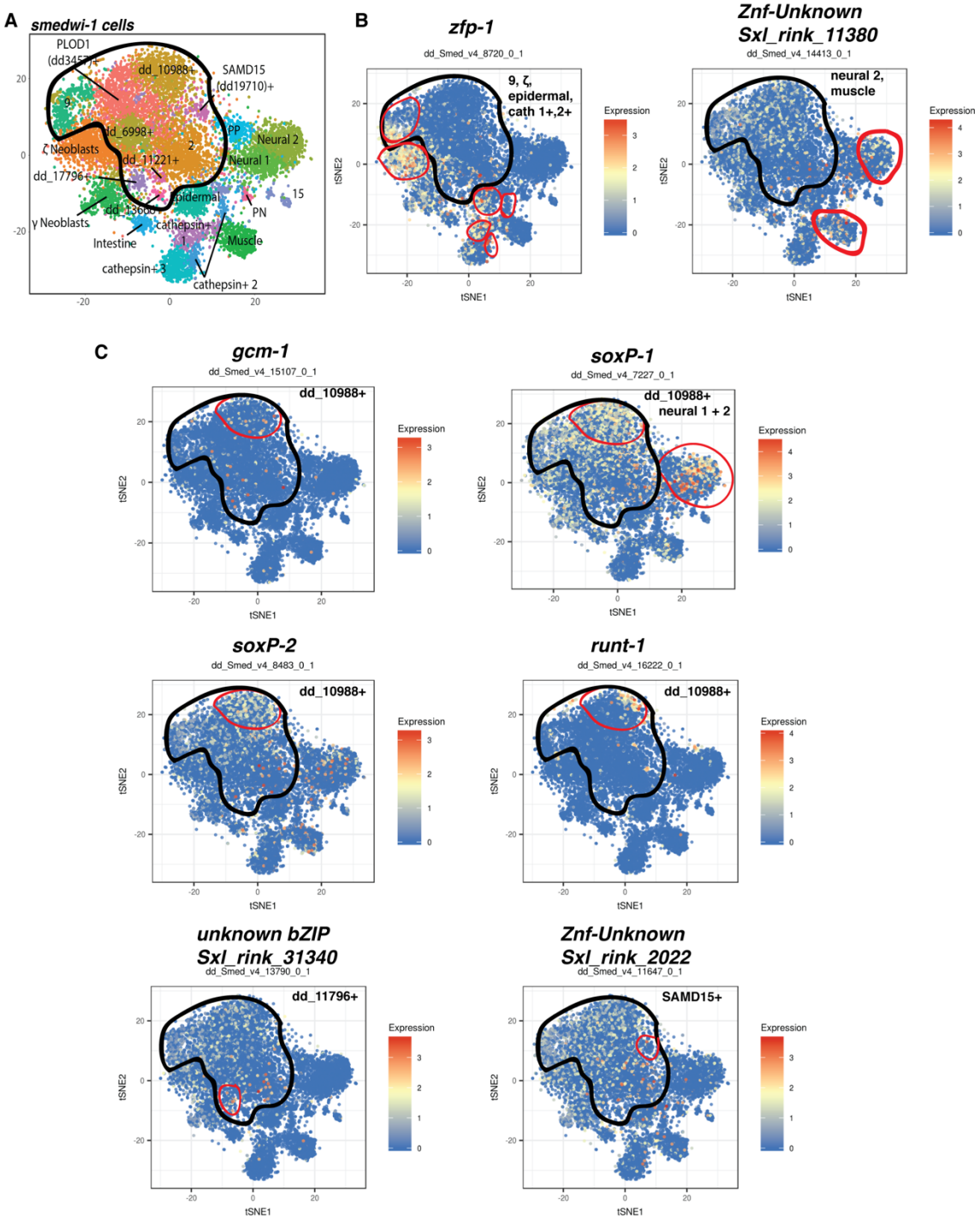


Figure 3: **A**. Data from https://digiworm.wi.mit.edu/. Single-cell tSNE plots of *piwi-1*+ cells predicts 22 sub-clusters. Circled in black are high *piwi-*+ NBs identified by Fincher et al. (2018) that exclude the lineage-committed ζ and γ NBs. **B.** For *zfp-1* and *Sxl_rink_11380*, expression of these genes are enriched in lineage-committed NBs (red circle). **C.** *gcm, soxP-1, soxP-2, runt-1, Sxl_rink_31340,* and *Sxl_rink_2022* are enriched for expression in specific clusters within the 9 high *piwi-1*+ population. Cluster of enrichment is shown in inset.

5.4 Homeodomain-containing DNA TEs are expressed in NBs

We found 128/682 putative TFs that had sufficient transcriptional coverage in our FACS RNA-seq datasets (=> 10 read counts in a single RNA-seq library) and which were annotated by RepeatMasker as DNA TEs. These selfish genetic elements encode a transposase within a single ORF that cuts at inverted flanking repeat sites at either end of the element. The DNA TE then re-inserts itself into the genome, and the previous excision site is repaired by either Homologous Recombination (generating a copy of itself) or by Non-Homologous End Joining (NHEJ). Consequently, DNA TEs are referred to as 'cut-and-paste' elements.

11 out of 19 DNA TE superfamilies (including Tc1/mariner, hAT (Hickman et al. 2005; Zhou et al. 2004) and PiggyBac (Mitra et al. 2008)) unambiguously contain an amino acid triad, consisting of two aspartic acid residues (D) and a third glutamic acid (E) or a third D, which are brought together in close proximity by an RNase-H-like fold to enable DNA cleavage (Haren et al. 1999; Yuan and Wessler 2011). Moreover, DNA transposons, at least of the Tc1/Mariner and Pogo classes, also have DNA-binding domains at their N-terminal regions that consists of a helix turn helix (HTH) (Vos and Plasterk 1994; Colloms et al. 1994). Indeed, many eukaryotic HTH domains likely originated from transposases, such as the paired domain of *PAX6* (Feschotte 2008; Chuong et al. 2017). As a result of their HTH motif, these DNA TEs appear in our putative TF list owing to Pfam HTH domain and 'homeodomain-like' SUPERFAMILY signatures. In contrast to DNA TEs, retroelements copy themselves via an RNA intermediate and as such are referred to as 'copy and paste' TEs. Long Terminal Repeat (LTR) and non-LTR (e.g. LINES and SINES) TEs represent the two major classes of retroelements and are distinguished by distinct replication mechanisms, (reviewed in (Levin and Moran 2011)), but both require the activity of a reverse transcriptase encoded by the retroelements itself. Although DDE motifs are present in at least the integrase enzyme encoded by the LTR TEs (Capy et al. 1997), neither LTRs nor non-LTR retroelements have an N-terminal 'homeodomain-like' DNA-binding domain. Consequently, retroelements are not found in our candidate TF lists.

Given that DNA TEs comprise of 15% of the planarian genome (**Figure 4A and B**), and 5% of expressed genomic loci **(Figure 4C and D**), we sought to provide *in silico* evidence for whether DNA TEs had functional activity within planarian NBs. Consequently, we looked at the normalized average TPM levels of RepeatMasker annotated DNA TEs in our X1 FACS datasets, as well as epigenetic marks that correlate with transcription. In particular, we generated genome-wide epigenetic profiles for H3K9me3 in X1 NBs – a mark known to correlate with the constitutive heterochromatin and TE silencing in animals (Martens et al. 2005; Rangan et al. 2011; Karimi et al. 2011; Bulut-Karslioglu et al. 2014). This analysis revealed low levels of H3K9me3 at DNA transposons, but higher levels of enrichment at LINEs (and to a lesser extent LTRs), indicating that DNA TEs avoid transcriptional suppression by host H3K9me3 modification machinery compared with retroelements (**Figure 4E**). There may be, however, mechanisms other than heterochromatin formation responsible for the silencing of active DNA TEs in the planarian genome, such as natural RNAi (Sijen and Plasterk 2003; Piast et al. 2005). Moreover, we also observed that the presence of H3K9me3 upstream or within an intron of genes is an indicator for the presence of retroelements (**Figure 4F**).

We found that 8/10 DNA TEs had low normalized TPM expression in NBs and were naked for all tested epigenetic marks (**Figure 5A**). Moreover, these DNA TEs were comparatively less evolved from their consensus sequence (i.e. low Kimura 2-paramter genetic distances), indicating that they may be evolutionarily younger. Consequently, these elements may represent TEs that have recently integrated within the genome and are enriched in expression in NBs. However, for two DNA TEs (Sxl_rink_30624 and Sxl_rink_30626) belonging to the Tc1/Mariner superfamily, we noted high levels of expression, together with high levels of the active H3K4me3 mark and low levels of heterochromatin associated H3K27me3 and H3K9me3. These TEs are also divergent from their Tc1/Mariner (ISrm11) family consensus sequence as indicated by a high Kimura 2-paramter genetic distance (**Figure 5B**). Consequently, these elements may represent evolutionarily old TEs that have accumulated mutations and, given their NB expression levels, have been co-opted to new functions beneficial to the host. An alignment of these two planarian Tc1/Mariner (ISrm11) sequences against

active Tc1/Mariner representatives from different animals shows that the two planarian elements both contain the conserved DDE/D motif in the transposase domain, as well as two helix-turn-helix motifs indicative of potential DNA binding activity (**Figure 5C**). We also found another NB expressed DNA TE (Sxl_rink_9792), but which lacked a RepeatMasker annotation, indicating that it may be highly derived. Moreover, this sequence had a high level of H3K4me3 deposition, and contained both a DDE domain as well as an N-terminus HTH motif annotated by Pfam.

Figure 4: **A.** Repeat content of the sexual *S. mediterranea* genome as identified by RepeatMasker. **B.** Kimura landscape showing Kimura 2-paramater distances to consensus sequence for all RepeatMasker annotated sites of the genome, and categorized as LINEs, LTR, DNA elements, RC Helitron elements, or Unclassified. **C.** Repeat content of sexual *S. mediterranea* genome annotations, which can be taken as transcriptionally active genomic loci owing to the annotation process. **D.** Kimura landscape for TEs present in genome annotations. Note all TEs in general that are expressed as evidenced by their annotation have shorter Kimura distances compared to genome-wide TEs, indicating that they are potentially younger. **E.** H3K9me3 profile in X1 NBs (1 replicate) across all RepeatMasker annotated TE sites shows that LINE and LTR elements have on average, higher levels of H3K9me3 compared to DNA elements. Annotated loci that have been filtered for expressed TEs (i.e. non-TE genes) have lower levels of H3K9me3 at the TSS, indicating that they are protected against heterochromatin formation. Both upstream and downstream of non-TE genes, there are higher levels of H3K9me3, indicative of the presence of TEs that need to be silenced (e.g. within introns and intergenic regions). **F.** Example RepeatMasker TE track (black) and genome annotation track (blue). Regions with high H3K9me3 (red) coverage correlate with the presence of TEs. TEs shown are likely transcriptionally inactive if there is no associated genome annotation. RNA-seq track is represented as normalized RPKM coverage using three whole worm RNA-seq libraries (SRR867386 - SRR867388). X1 ChIP-seq library tracks are shown H3K4me3, H3K27me3 and H3K9me3 as log2 ratio of normalization to input.

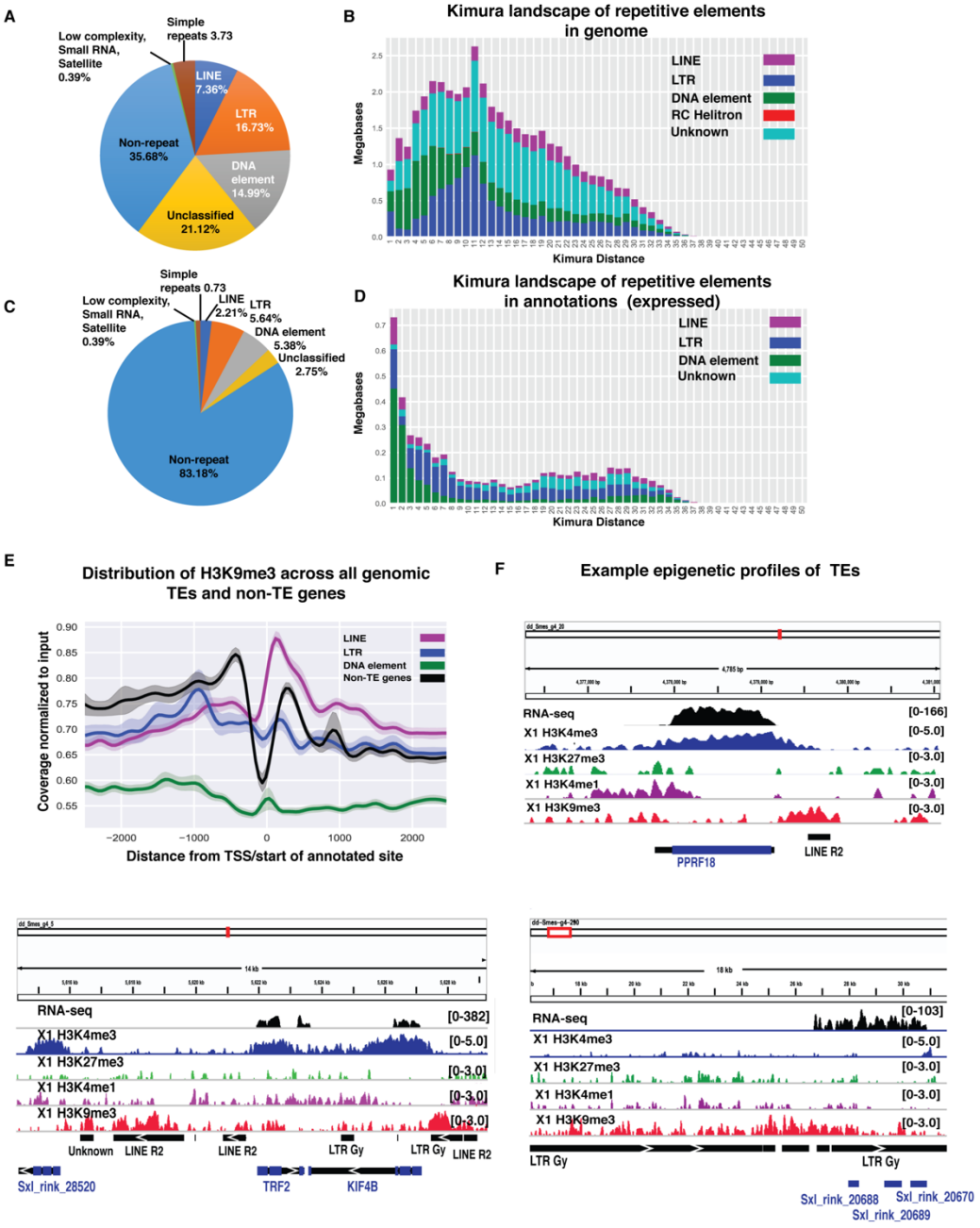

A
Low complexity, Small RNA, Satellite 0.39%
Simple repeats 3.73
LINE 7.36%
LTR 16.73%
DNA element 14.99%
Unclassified 21.12%
Non-repeat 35.68%
B
Kimura landscape of repetitive elements in genome
LINE
LTR
DNA element
RC Helitron
Unknown
Megabases
Kimura Distance
C
Low complexity, Small RNA, Satellite 0.39%
Simple repeats 0.73
LINE 2.21%
LTR 5.64%
DNA element 5.38%
Unclassified 2.75%
Non-repeat 83.18%
D
Kimura landscape of repetitive elements in annotations (expressed)
LINE
LTR
DNA element
Unknown
Megabases
Kimura Distance
E
Distribution of H3K9me3 across all genomic TEs and non-TE genes
LINE
LTR
DNA element
Non-TE genes
Coverage normalized to input
Distance from TSS/start of annotated site
F
Example epigenetic profiles of TEs
RNA-seq
X1 H3K4me3
X1 H3K27me3
X1 H3K4me1
X1 H3K9me3
PPRF18
LINE R2
Unknown
LTR Gy
Sxl_rink_28520
TRF2
KIF4B
Sxl_rink_20688
Sxl_rink_20670
Sxl_rink_20689

Figure 5. **A.** NB enriched DNA TEs with low X1 TPM levels. These TEs also have low Kimura distances from their consensus sequences, indicating that they are evolutionarily younger. These DNA TEs have representative genome annotations owing to expression in the transcriptome assemblies used to make the annotation. In this figure, RNA-seq track is represented as normalized RPKM coverage using only three whole worm RNA-seq libraries (SRR867386 - SRR867388). X1 ChIP-seq library tracks are shown H3K4me3, H3K27me3 and H3K9me3 as log2 ratio of normalization to input. ATAC-seq libraries for X1, X2, and Xins merged replicates are shown as normalized RPKM coverage. **B.** Two DNA TEs, belonging to the TcMariner superfamily that are enriched in expression in NBs. Both DNA TEs are derived as indicated by their high Kimura distance

from consensus sequence. Both are marked with H3K4me3. **C.** Alignment of the two planarian DNA TEs with protein sequences of known expressed Tc1-like DNA elements from plaice (Genbank: CAB51372), frog (Genbank: DAA0156), *C. elegans* (Genbank: NP_741053) as well as the human mariner-2 TE consensus sequence GenBank: U49974.1). Residues of the DNA binding domain are shown in blue (2x HTH domains consisting of alpha helices H1-H3) and catalytic DDE motifs are shown in red. The GRPR motif characteristic of many homeodomain-containing proteins and the Tc1/mariner transposases is not found in the two *S. mediterranea* DNA TEs, highlighted in yellow, indicating possible divergence of function.

5.5 ATAC-seq identifies changes in accessibility at Xins-enriched gene promoters, but does not X1-enriched gene promoters

The Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) can be used to identify both promoters and enhancers that change in chromatin accessibility between cell-types, and which broadly correlate to the transcriptional activity of associated genes. Consequently, we reasoned that we could employ ATAC-seq for the first time in planarians to ascertain promoter-proximal cis-regulatory elements that are active in particular FACS isolated compartments, and therefore indicative of either a NB (X1), post-mitotic (X2), or differentiated cell-type (Xins) specific function. This methodology relies on the incubation of a hyperactive form of Tn5 transposase, which introduces two cuts (9bp apart) at a particular site as well as the simultaneous insertion of sequencing adaptors within the DNA (Buenrostro et al. 2013, 2015a). Where chromatin is more accessible, there will be a greater probability of Tn5 insertion, compared with less accessible chromatin such as in condensed heterochromatin (**Figure 6A**). The fragments excised out of the DNA will be the products of two adjacent Tn5 reactions, and will be the input to PCR amplification followed by sequencing. Using this methodology, we were able to generate >12million uniquely mapping paired-end fragments per library, with two replicates in each of the X1, X2 and Xins samples. Initially, we analyzed the fragment distribution of each library and noted that the majority of fragments were <100bp, and therefore likely correspond to nucleosome-free regions (NFR) (**Figure 6B**). Given that the size of DNA wrapped around nucleosomes is around ~147bp, we also observe periodicity in fragment size, at least for our X2 and Xins samples that corresponds to the presumptive mono- and di-nucleosome length that the Tn5 enzyme cuts around.

We subsequently filtered out our mapped fragments to include only those that had a length of <100bp, as we wished to only look at NFRs distributed across the genome. We used MACS2 to call peaks from these NFR reads for each of the FACS isolated populations, in order to ascertain genomic sites with the greatest amount of chromatin accessibility (Zhang et al. 2008). We found a total of 18,331 consensus peaks in X1 cell replicates, 32,440 peaks in X2 cells, and 22,092 peaks in Xins cells that were present in both of replicates (**Figure 6C**). We utilized our annotation of the genome to estimate that the majority of called peaks for each of the three samples were in the intergenic regions of the genome, with comparatively fewer peaks being called at promoter-TSS annotated sites.

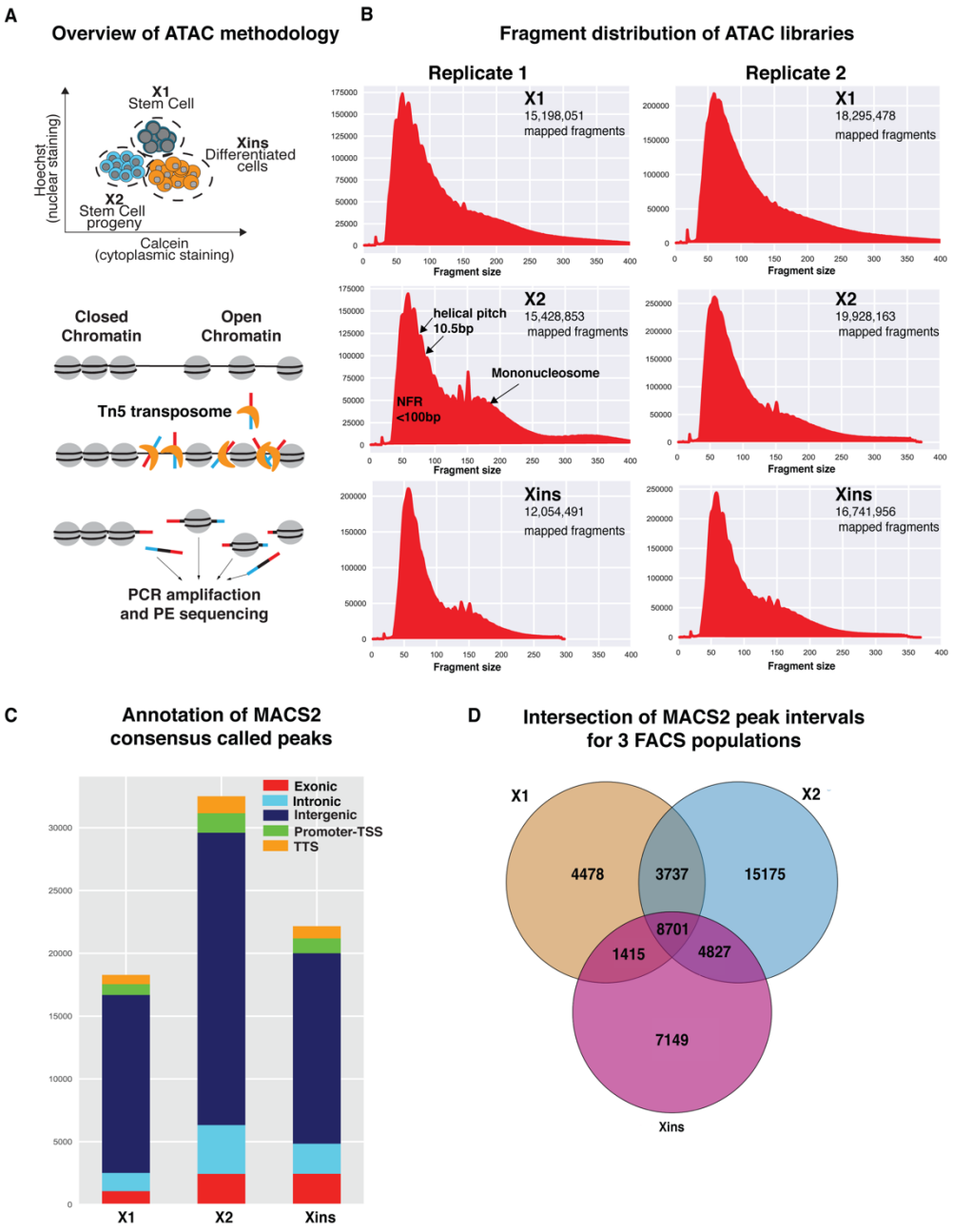


Figure 6: **A**. Overview of ATAC-seq methodology. **B.** Mapped fragment distribution sizes for replicate ATAC-seq libraries. **C.** Annotation of MACS2 called peaks (consensus between replicates) to nearest genomic feature. **D**. Intersection of BED intervals for MACS2 consensus peaks of X1, X2, Xins libraries. An overlap of 1bp or more is taken as a common overlap between peaks.

Despite attaining fragment distributions similar to those of other published ATAC-seq libraries from other model systems, further evidence was needed to verify the quality and usability of our ATAC-seq libraries for identifying enhancer regions. We expected that the extent of accessibility at promoters should correlate with the transcriptional activity of corresponding genes. Given the

relative paucity of MACS2 called promoter-TSS peaks in our ATAC-seq libraries, we plotted normalized read coverage in X1, X2, Xins ATAC-seq libraries across the top 700 genes enriched by transcriptional expression in the X1, X2 and Xins FACS categories. We found that for genes categorized as being X1 enriched, a high ATAC-seq coverage (<100bp fragments) was apparent across the gene bodies of these genes, with no obvious enrichment at the promoter region (**Figure 7A**). Moreover, contrary to our expectations, ATAC-seq coverage increased downstream of the promoter between X1 and Xins ATAC-seq experiments, indicative of the fact that these genes either remained as accessible upon differentiation or increased in accessibility (**Figure 7A and 7D**). To verify this, we plotted the individual ATAC-seq profiles of TFs that we identified as being X1 enriched in the preceding section (**Figure 8**). We further included ChIP-seq tracks for X1 cells to show that whilst these X1-enriched TFs have a high H3K4me3 and low H3K27me3 indicative of active promoters in the X1 compartment, the nucleosome-free region immediately preceding H3K4me3 enrichment did not change in the Xins compartment following differentiation. To be sure that we were identifying promoter regions, we also plotted a RNA-seq coverage from whole worm RNA-seq libraries, and verified that our promoter ATAC-seq peak also preceded the start of transcriptional coverage at the 5'UTR of genes.

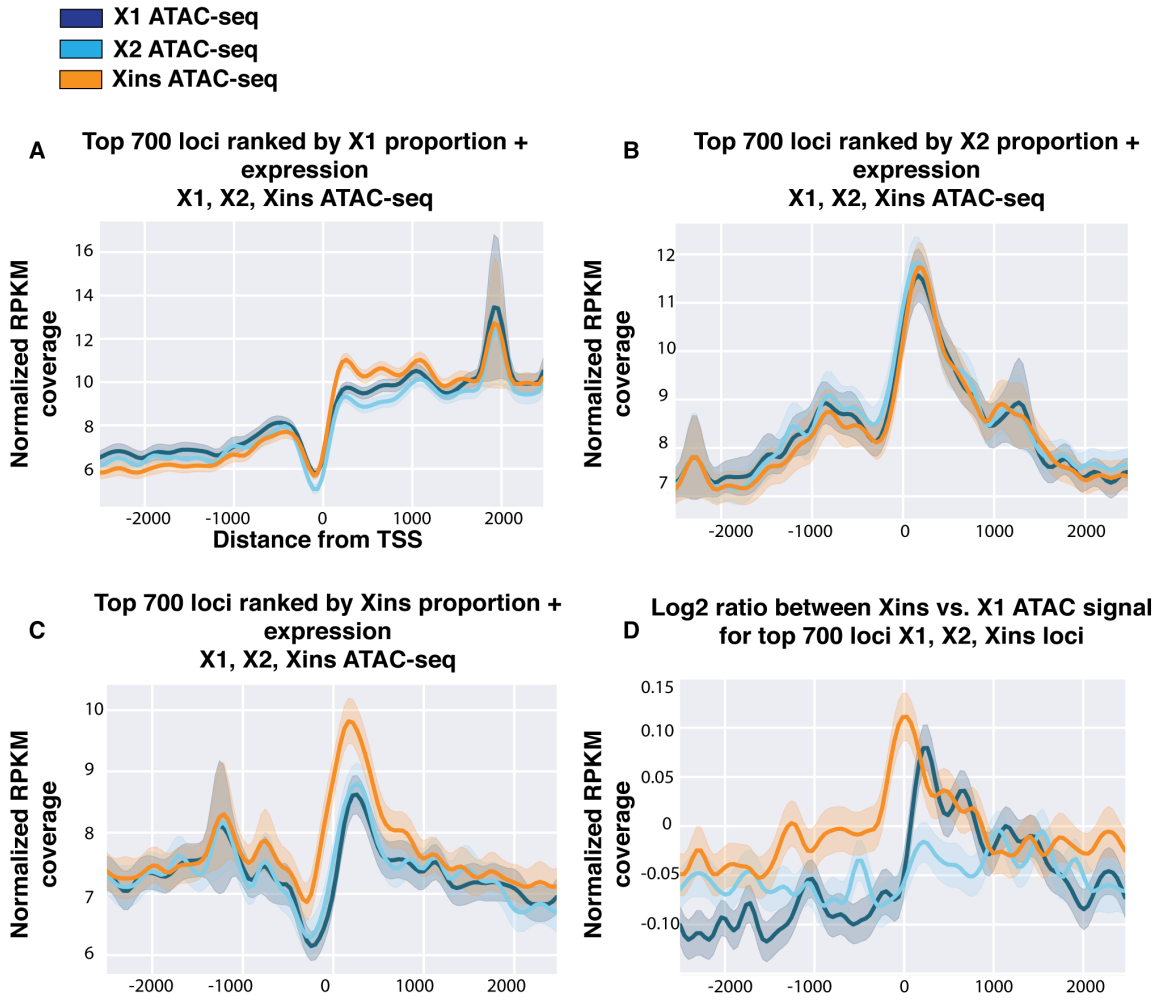


Figure 7: Dark blue represents average X1 ATAC-seq libraries, light blue X2 ATAC-seq libraries and orange Xins ATAC-seq libraries. **A.** Gene profiles for the 700 loci ranked by X1 proportion and then TPM expression in X1, X2, Xins FACS ATAC-seq libraries. Note that ATAC-seq coverage is equally distributed from the TSS to gene bodies for X1 enriched genes in all three FACS ATAC-seq libraries. **B.** Gene profiles for the 700 loci ranked by X2 proportion and then TPM expression in X1, X2, Xins FACS ATAC-seq libraries. An average peak around the TSS of X2 enriched genes is observed that does not change in magnitude between FACS ATAC-seq libraries. **C.** Gene profiles for the 700 loci ranked by Xins proportion and then TPM expression in X1, X2, Xins FACS ATAC-seq libraries. A clear peak for Xins genes near the TSS is observed, which is greater in magnitude in the Xins FACS library compared to X1 and X2 ATAC-seq libraries. **D**. Log2 ratio between the coverage in Xins versus X1 ATAC-seq libraries. There is an average increase in the accessibility of Xins genes close to the TSS in the Xins ATAC-seq libraries compared to X1. The accessibility of X1 genes also increases, but is downstream of the TSS in comparison.

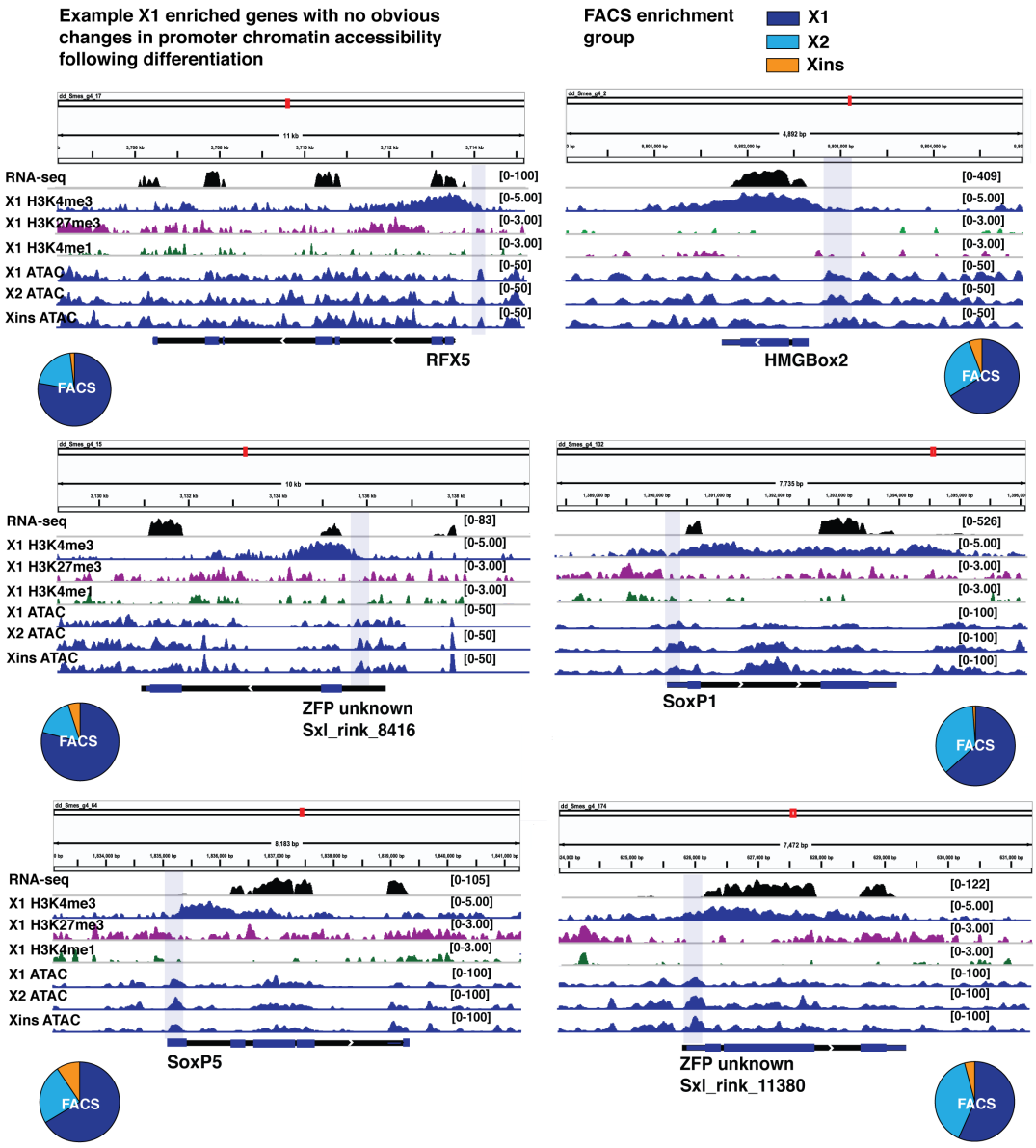


Figure 8: A selection of individual genes enriched in X1 NBs that have promoters (highlighted blue box) which do not obviously change in accessibility in the Xins FACS compartment as assayed by ATAC-seq. RNA-seq track is represented as normalized RPKM coverage using whole worm RNA-seq libraries (SRR867386 - SRR867388). X1 ChIP-seq library tracks for H3K4me3, H3K27me3 and H3K4me1 is shown as a log2 ratio of normalization to input. ATAC-seq libraries for X1, X2, and Xins merged replicates is shown as normalized RPKM coverage.

Conversely, for Xins enriched genes we noted an increase in accessibility at the promoter-proximal region between X1 and Xins ATAC-seq (**Figure 7C**). This is consistent with these genes gaining accessibility upon differentiation from the stem cell compartment. Indeed, the individual profiles of genes known to be exclusively expressed in differentiated cells corroborated this result. For example, both *calmodulin* and *myosin heavy chain-1* are known to be expressed in the muscle, with little or

no expression in NBs, and read coverage at the promoter region of these genes is greatest in Xins, compared with a much lower coverage in X1 ATAC-seq samples (**Figure 9**). We also noted that some Xins enriched genes (such as *Zinc finger 609a*, *calmodulin*, *myosin heavy chain* 1) may become permissive to transcription during the differentiation process from stem cell, as indicated by an increased read coverage at the promoters of these genes in the X2 ATAC-seq samples compared with the X1 ATAC-seq samples. Given that X2 is a highly heterogeneous compartment, these promoter peaks may represent late differentiating cells transitioning to the Xins compartment, and which are permissive to RNA-Pol II docking and transcription.

Figure 9: A selection of individual genes enriched in differentiated cells that have promoters (highlighted blue box) which increase in accessibility in the Xins FACS compartment compared to X1 as assayed by ATAC-seq. Note that for genes such as *calmodulin*, *LPPR5*, and *myosin-heavy chain 1*, chromatin accessibility begins in the X2 compartment even though these genes have little transcriptional representation in X2 compartment. This may reflect that these genes are beginning to become permissive to the transcriptional machinery during differentiation. RNA-seq track is represented as normalized RPKM coverage using whole worm RNA-seq libraries (SRR867386 - SRR867388). X1 ChIP-seq library tracks for H3K4me3, H3K27me3 and H3K4me1 are shown as log2 ratio of normalization to input. ATAC-seq libraries for X1, X2, and Xins merged replicates is shown as normalized RPKM coverage.

5.6 ATAC-seq identifies potential enhancers that correspond with neighbouring gene expression changes

In the absence of transgenic reporters in planarians to verify the cellular co-localization of enhancer and target gene expression, we reasoned that one way in which we can provide evidence for the existence of enhancers is to correlate their accessibility with target gene expression, inferring target genes simply by their proximity. We therefore filtered our list of MACS2 called consensus peaks for each FACS category to contain only intronic peaks, and intergenic peaks that were within 7.5kb of the nearest gene. By this methodology, we were able to find 889 promoter-proximal X1 peaks, 3547 promoter-proximal X2 peaks, and 1471 promoter-proximal Xins peaks. From these FACS cell-type specific peaks only a small proportion positively correlated with neighbouring gene expression: e.g. 146/889 (16.4%) of X1-specific peaks had neighbouring genes that were X1 enriched, 269/3547 (7.6%) represented X2-specific peaks that neighbored X2 enriched genes, and 503/1471 (34%) Xins peaks neighbored genes with Xins enrichment. The remaining peaks may be indicative of regions accessible to negative regulators, or false-positives called by MACS2. We looked at likelihood of nucleosome occupancy (inferred by the NucleoATAC using reads from all 3 FACS samples) either side of the MACS2 called peak regions as well as the density of H3K4me1 normalized read coverage (**Figure 10A**). We found that these putative enhancer regions have a strong nucleosome positioning either side of the -100bp open chromatin window, and these flanking nucleosomes are likely to be marked with H3K4me1 **(Figure 10B and 10C**). In general, the presence of H3K4me1 is generally used to distinguish enhancers from proximal promoters (Heintzman et al. 2007; Ernst and Kellis 2015; Kharchenko et al. 2011; Daugherty et al. 2017; Jänes et al. 2018), and as such our open-chromatin regions are likely planarian regulatory elements.

Given the small number of peaks predicted to be positive enhancers of genes, we also manually looked at the ATAC-read distribution across individual scaffolds. For instance, the planarian orthologue of *SPSB-1* (SPRY domain-containing SOCS box protein 1) is enriched for expression in the X2 compartment, and additionally contains an X2-specific MACS2 called peak in its intron (**Figure 11**). There is, however, a smaller density of fragments mapping to this same region in the

X1 and Xins ATAC-seq libraries. Intriguingly, *SPSB-1* is also marked by H3K4me3 in X1 cells and together with the absence of H3K27me3 at the promoter-proximal region is indicative of a transcriptional output in NBs. The expression of *SPSB-1* may be modulated by the highlighted putative enhancer to increase expression upon differentiation in the X2 compartment. Similarly, a planarian solute carrier gene, *slc2a-2,* is highly expressed in the Xins compartment, consistent with its single-cell profile of neuronal enrichment, and also contains a highly accessible distal intronic region in Xins cells, but which is less accessible in the X1 and X2 cell types (**Figure 11**). An example of a putative enhancer regulating expression in NBs is highlighted in the intronic region of *BAG6-like* - a gene enriched for expression in X1 and X2 FACS compartments, and whose orthologue in humans is responsible for DDR signaling and damage-induced cell death. In all of our highlighted examples, we observed an enrichment of H3K4me1 around these putative enhancer regions.

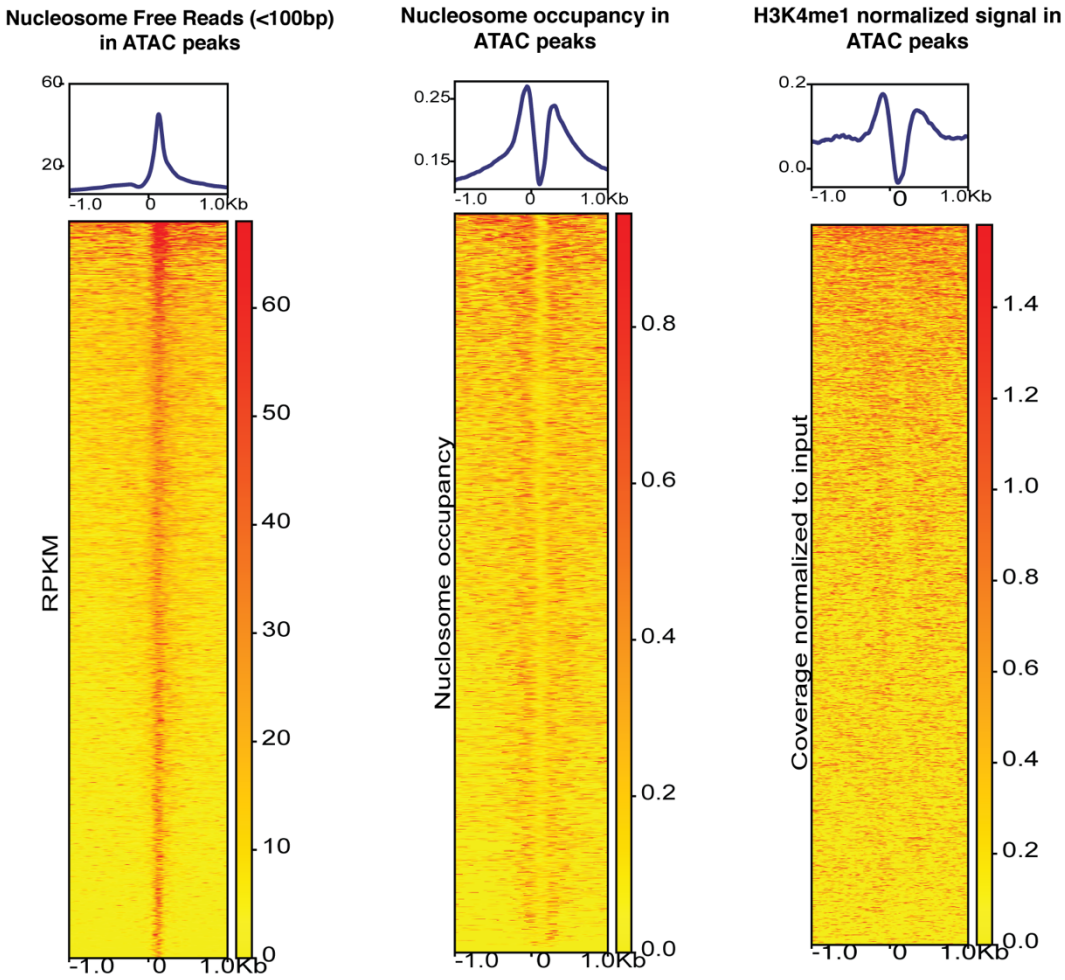


Figure 10: (A) Density of nucleosome free reads (<100bp) in MACS2 called ATAC peaks indicates that peaks are generally narrow (-100bp) and not concentrated in broad 'open-chromatin' regions. (B) NucleoATAC was used to discern likelihood of nucleosome occupancy (on scale of 0-1) around ATAC-peaks. Open chromatin regions are on average flanked by nucleosomes either side. (C) H3K4me1 normalized coverage to input indicates a higher coverage either site of ATAC-seq peaks. In all three heat maps, X-axis represents distance from ATAC-seq peak center.

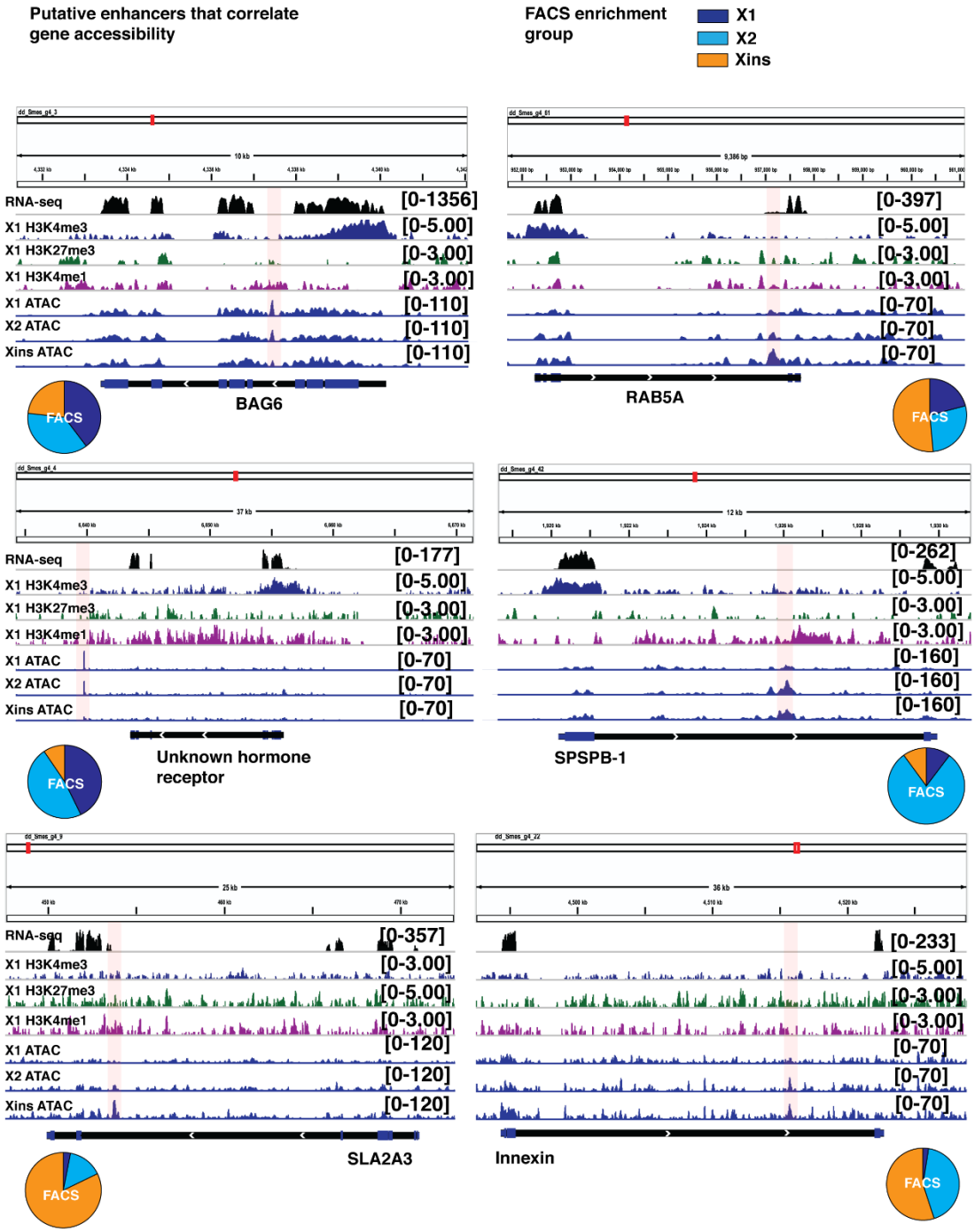


Figure 11: A selection of distal open chromatin regions (either within introns or 10kb intergenic distance from promoter) that correlate with nearby gene expression (highlighted pink box) as assayed by ATAC-seq. These regions are putative enhancers. RNA-seq track is represented as normalized RPKM coverage using whole worm RNA-seq libraries (SRR867386 - SRR867388). X1 ChIP-seq library tracks for H3K4me3, H3K27me3 and H3K4me1 are shown as as log2 ratio of normalization to input. ATAC-seq libraries for X1, X2, and Xins merged replicates is shown as normalized RPKM coverage.

We also searched for potential TF target motifs in our consensus MACS2 called peaks specific for the X1, X2 and Xins populations. We performed a search of known motifs of TFs in the HOMER database, with the majority of motifs being inferred from publicly available ChIP-seq data from mammals. HOMER discovered an enrichment for the Sox3 human motif ($p<0.01$) in the 4478 X1-specific peaks, utilizing the 7149 Xins-specific peaks as background peaks. All metazoan Sox genes are known to bind to a highly conserved WWCAAW (W=T or A) motif, and target gene selectivity can be achieved through differential affinity for flanking residues (Harley et al. 1994; Wegner 2010). Given that Sox3 has no obvious orthologue in planarians, the detection of an enriched Sox3 motif may simply be reflective of Sox motifs in general. Given that 4 planarian Sox genes (*soxP-5, soxP-1, soxP-2, soxB1-1*) are enriched in the X1 compartment of *S. mediterranea,* we reasoned that any of these genes can potentially bind to the Sox motif in X1 NBs. We found that the 239 Sox-containing X1-specific peaks may reflect the function of these 4 TFs in NBs. However, only 78 of these X1-specific peaks lie within an intron or 10kb distance of a gene, and from these 15/78 are X1-enriched genes (**Figure 12A**). Consequently, planarian Sox genes may act as negative regulators of target genes, and as such open chromatin regions containing Sox motifs will be more open in NBs to allow for this function. Moreover, motif detection by homology with mammalian ChIP-seq data may simply not allow for the detection of planarian TFs.

Given that there are four NB enriched Sox TFs, we cannot easily tease out the unique targets of these genes, and neither are we sure whether their predominant functions are as activators or repressors. Consequently, we realized that the bHLH TF, Collier/Olfactory-1/Early B-cell factor (COE), has only one orthologue in invertebrates, whereas vertebrates contain four paralogues. In planarians, *in situ* hybridization has shown that *coe* is expressed in a small population of neural committed NBs, and is necessary for the expression of genes involved in the maintenance and identity of neuronal subtypes (i.e. neurotransmitters, ion channels, and neuropeptide-encoding genes) (Cowles et al. 2014). We therefore looked for the EBF1 motif, the target for COE TF in humans, in our planarian peak dataset across all three FACS compartments. Only 37 peaks were identified with an underlying

COE motif, and we used the single-cell database to find only 9 peaks whose closest gene (within 10kb) was enriched in expression during neuronal differentiation and commitment (**Figure 12B)**.

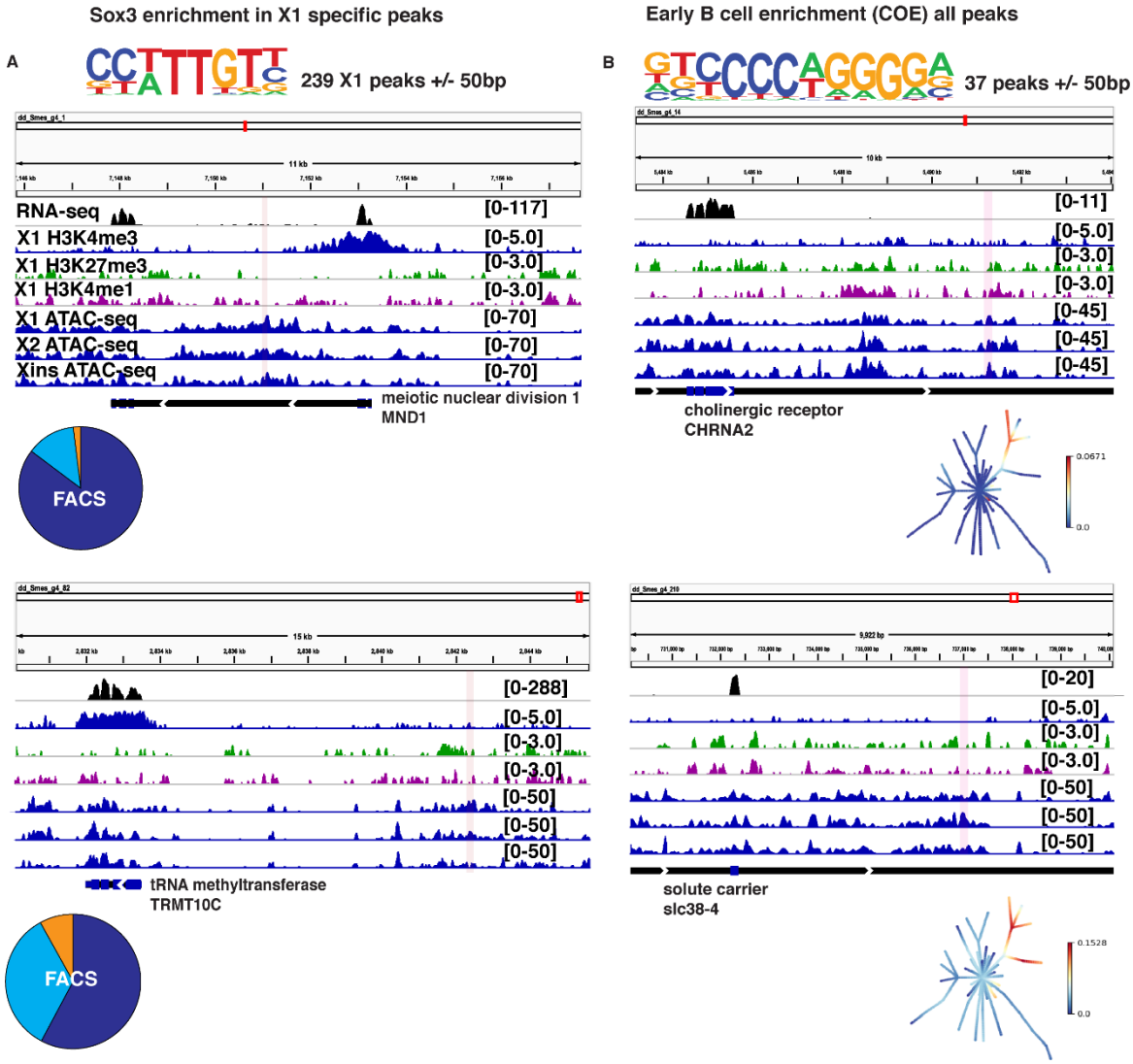


Figure 12: (A) *mnd1* and *trm10c* are X1 enriched genes with Sox3 motifs within X1-specific MACS2 called peaks highlighted in red. (B) *chrna2* and *sclc38-4* are both genes expressed in neuronal subtypes as shown by the lineage tree generated by Plass et al. 2018. Both of these genes have an intronic peak with a motif for Early B factor (EBF) whose orthologue in planarian is *coe* – a gene known necessary for neurogenesis. The peak called for *chrna2* is in the Xins ATAC-seq libraries, whereas the peak for *scl38-4* is X2 specific.

5.7 Discussion

We wish to understand the TF-mediated cis-regulatory elements and broader gene regulatory networks (GRNs) that enable NB maintenance and differentiation, and this chapter has made a preliminary attempt towards achieving this ultimate goal.

Firstly, we have documented the potential TFs present expressed in planarian cells, using an expression-based annotation of the highly contiguous sexual *S. mediterranea* genome. In particular, we have for the first time identified NB-enriched TFs that play a presumptive role in either maintaining NB pluripotency or early lineage commitment. Interestingly, we found a number of uncharacterized genes, such as those belonging to the C2H2-ZNF family, which will be targets for downstream *in vivo* experiments that include *in situ* hybridization (to verify stem cell expression), and RNAi (to ascertain whether the NB population or lineage-committed NB subsets are affected). Moreover, we identify TEs that are active within the NB population, and show that at least two of these elements are highly derived, and display high levels of the active H3K4me3 and low levels of suppressive H3K9me3/H3K27me3 marks, indicative of the planarian epigenetic machinery actively maintaining their expression. In the future, molecular phylogenetic analyses will establish the evolutionary origin of these 'unknown' ZNFs and active TEs, and whether or not they are flatworm specific.

Secondly, we have presented preliminary results for ATAC-seq on the three FACS-isolated populations of planarians. We achieve the expected result in terms of the size distribution of ATAC-seq paired-end fragments, and for the X2 and Xins replicate libraries these display a clear periodicity indicative of Tn5 insertion between mono- and di-nucleosomes. There is, however, a distinct lack of an obvious nucleosome peak in the X1 replicate libraries. We separated out nucleosome-free fragments from our samples and used MACS2 to call peaks to ascertain regions of the genome that are highly accessible to Tn5 insertion and therefore are likely to be regions of open chromatin. We obtained the highest number of MACS2-called peaks in our X2 samples (32,440) compared with

Xins (22,092) and X1 (18,331). In biological terms, this may relate to the fact that cells undergoing lineage commitment have more many more accessible regulatory sites compared to stem cells. However, for most putative enhancers in all FACS populations, the expression dynamics of these elements did not correlate with the expression of the nearest gene promoters, perhaps indicative of more distal gene regulation.

Given the poor periodicity in our X1 libraries, relative to X2 and Xins, together with comparatively fewer called peaks by MACS2, we can posit that our X1 libraries are over-tagmented. If this were true, we would expect there to be many more 'false-positive' fragments that do not originate from nucleosome-free regions (NFR), but which instead are generated from the over-tagmentation of larger nucleosome-spanning fragments. This would increase the overall level of fragment noise surrounding truly enriched ATAC-seq regions, and as such MACS2 would have difficulty in calling peaks. This would also explain the smoothness and lack of distinct nucleosome peak observed in the fragment distribution pattern for the X1 libraries. In order to test this possibility, in the future, we will vary the time of incubation of Tn5 as well as utilize different cell numbers in case NBs are more sensitive to tagmentation compared to post-mitotic cell types.

Additionally, it is known that chromatin configurations change during the cell-cycle (in particular, in S-phase during DNA replication, and in mitosis as chromatin condenses). Given that X1 FACS isolated planarian cells reflect a heterogeneous population of cell-cycle stages, our ATAC-seq results for X1 NBs may simply be a reflection of an average of distinct chromatin configurations that occur at specific cell cycle phases (Hogan et al. 2006). Indeed, MNase-seq derived nucleosome maps in *S. cerevisiae* showed that nucleosome 'fuzziness' increases around the TSS in S and M phase as a result of the nucleosome sliding along the DNA during these cell-cycle stages (Deniz et al. 2016).

In order to validate that our ATAC-seq methodology can detect open-chromatin regions that change dynamically according to cell-type, we first correlated the accessibility of promoters with the transcriptional activity of associated genes in each FACS compartment. We noted that for Xins-

enriched genes, chromatin accessibility at the promoter increased during differentiation, consistent with these genes becoming transcriptionally active in differentiated cells. However, counterintuitively, we observed that X1 genes, remained accessible at the promoter region during differentiation, and in some cases even increased accessibility. One intriguing explanation is that X1-enriched genes remain permissive to transcription at the promoter region, but enhancers for these genes are not active thus reducing overall transcription in the differentiated cell compartments. However, this may simply by an artefact of the X1 ATAC-libraries being a collection of cells in different stages of the cell-cycle – thus, the 'open-ness' of NB gene promoters is not fairly represented. It is clear, therefore, that single-cell ATAC-seq on the X1 compartment would help in partitioning out the heterogeneity originating from cells being in different cell-cycle stages. Moreover, single-cell ATAC-seq would also help to discern the accessibility landscape of lineage-committed NB subsets, as well as different cell-types in the animal as a whole.

We were able to make attempts to identify putative intergenic and intronic enhancers that were located within 10kb of a gene promoter, and which correlated with this gene's activity. However, the majority of called peaks by MACS2 did not correlate with neighbouring gene expression, either because these putative non-coding 'open-chromatin' regions are transcriptional repressors or they may be involved in the activation of distally located genes. It is clear that further replicates of ATAC-seq need to be attempted, most likely utilizing a range of Tn5 incubation times, to be sure that our findings are not simply artefacts of technical bias. We can also utilize alternative methodologies such as DNase-seq, which, has been documented to identify a far greater number of open promoters and gene bodies, but which requires a higher cell starting material (Clark et al. 2018). Moreover, enhancers can also be identified by ChIP-seq against the transcriptional cofactor p300, which has been shown to be a ubiquitously expressed component of protein assemblies at enhancers in animals including *C. elegans* and the sea anemone *Nematostella vectensis* (Visel et al. 2009; May et al. 2012; Schwaiger et al. 2014; Ho et al. 2017). Indeed, using complementary methodologies would support our identification of putative enhancers in the planarian genome.

Once we have consistently identified enhancers in planarian NBs that correlate with the activity of proximal promoters, there is an overt challenge in verifying their function and targets given that transgenic reporters are not available. Searching for TF binding motifs under individual ATAC-seq peaks is one way of identifying potential TF interactors, but these motifs will be inferred from ChIP-seq datasets from other animals, likely mammals. However, as is the case for the Sox gene family, TFs will be divergent between species, and only potential candidates can be inferred. For instance, the identification of Sox3 motifs under X1 ATAC-seq peaks enables us to narrow down our list of potential TF mediators to *soxP-5*, *soxP-1*, *soxP-2* and *soxB1-1*, as these genes are all expressed in the X1 compartment. If either of these TFs is a pioneer factor, RNAi knock-down of one of these TFs can potentially alter the chromatin configuration of corresponding ATAC-seq peak, and as such will be less accessible to Tn5. Consequently, we would be able to infer a putative gene regulatory network (GRN) in planarian NBs, by identifying a NB-specific TF and its epigenetic and transcriptional effect on target genes. More broadly, as discussed in Chapter VI, we can identify potential promoter-enhancer interactions using techniques such as promoter-capture HiC (Schoenfelder et al. 2015a) and single-cell ATAC-seq (Pliner et al. 2018).

Chapter VI

Thesis Discussion and Future Directions

6.1 Building a landscape map of histone modifications in planarian cell types

In this thesis, we have shown that ChIP-seq of planarian NBs works robustly to generate histone modification profiles for genes that also correlate with their transcriptional output. We have utilised the marks for H3K4me3 and H3K36me3, which correlate with active NB gene expression, as well as the marks for H3K27me3, H3K4me1, and H3K9me3 that correlate with transcriptional suppression and heterochromatin. In the future, this can also be extended to performing ChIP-seq for both post-mitotic progeny (X2) and differentiated (Xins) cells. However, both the X2 and Xins compartments are highly heterogeneous, and so the data arising from such experiments will likely not be biologically informative. For instance, signal arising from the presence of specific histone modifications for genes expressed in the neuronal lineage will be masked by opposing histone modifications for the same genes in cells of the gut. Given that single-cell ChIP-seq technologies are still in their infancy (Rotem et al. 2015), conducting ChIP-seq experiments on the X1 compartment seem our only reasonable avenue to yield usable epigenetic information in planarians, as unambiguous conclusions can be reached about genes known to be expressed in NBs and those with minimal expression in this compartment.

However, the presence of lineage-committed NBs in the X1 compartment, means that genes can be marked by opposing histone modifications if they play a role in early NB commitment and are X1-enriched, such as in the case of *zfp-1* which is specific to the ζ NBs. The recent discovery and production of an antibody against a cell surface marker, TSPAN-1, has been shown to enrich for individually pluripotent cNBs (Zeng et al. 2018), and may be an attractive solution to circumventing the issues associated with NB heterogeneity. By prospectively isolating NBs using the TSPAN-1 antibody, a truer and more direct representation of histone modification landscape of pluripotent NBs can be gained, without contamination of lineage-committed NBs.

Given the recent release of the highly contiguous sexual *S. mediterranea* genome, it would be worth re-mapping our ChIP-seq data to the new genome assembly and making this information accessible to the planarian community via the online UCSC genome browser (Grohme et al. 2018; Rozanski et al. 2019)(http://planmine.mpi-cbg.de/planmine/genome.do). We have accomplished this in part in Chapter VI, but allowing this information to be available as an online resource, together with a curated set of genomic annotations, will undoubtedly help planarian biologists easily ascertain the epigenetic status of their genes of interest without requiring specific bioinformatics expertise. We have shown that RNA-seq data from FACS compartments can be variable between different labs, and we predict that this will also be the case for ChIP-seq - an assay known to be intrinsically noisier and variable than RNA-seq. Therefore, efforts must be made to ensure consistency of both RNA-seq and ChIP-seq results in the future, if planarians are to be used as a model system for genomics-based research. Eventually, given the generation of more datasets, an ENCODE styled curation will be necessary to establish a consensus status of both epigenetic and transcriptomic information of cells in the three planarian FACS compartments (Landt et al. 2012; Ecker et al. 2012).

6.2 Investigating the function of bivalent histone modifications in planarian cells

We have shown that genes involved in the differentiation process, and as such enriched in the X2 compartment, are marked by bivalent histone modifications allowing them to be kept in a poised transcriptional state in NBs. We have shown that these genes have a preponderance of paused RNA-Pol II at their promoter-proximal regions, and that these genes have little or no transcription in NBs as assayed by both bulk-cell and single-cell methodologies. We did attempt to prove this definitively, performing re-ChIP on NB cells – whereby chromatin was sequentially passaged through immunoprecipitations for the opposing H3K4me3 and H3K27me3 marks. The resultant DNA concentrations from our elutes were too small to be sequenced without additional PCR amplifications, and our resultant libraries indicated high read duplication and low coverage across the genome. Given that mammalian re-ChIP protocols, including the one we followed (Weiner et al. 2016), typically use at least 60x greater input chromatin material than that of our planarian single-

ChIP-seq experiments, future attempts to carry this out should be weary of requiring a colossal number of worms and dedicated FACS time in the order of weeks.

We can, however, begin to perform RNAi knockdowns against genes known to mediate both the H3K4me3 and H3K27me3 marks on bivalent promoters in other model systems. For instance, a crucial component of silencing bivalent genes in ESCs is the Polycomb repressive complex 2 (PRC2) and analysis of cells lacking the *Eed* subunit of PRC2 showed that the deficiency results in an almost complete absence of H3K27me3 genome-wide, as well as the activation of several genes with H3K4me3 deposition at promoters (Azuara et al. 2006). Moreover, RNAi of the histone deacetylase *HDAC-1* has shown to result in differentiation defects in planarians, and this phenotype may be the result of premature activation of bivalent loci owing to the incomplete conversion of H3K27ac to H3K27me3 (Reynolds et al. 2012b; Robb and Alvarado 2014). Similarly, knockout of the methyl-transferase complex MLL2 in mouse ESCs (mESCs) led to a depletion of H3K4me3 on promoters that have very little transcriptional output and that are also typically marked by H3K27me3 in wild-type mESCs. Interestingly, MLL1/2 RNAi knockdown in planarians results in the premature expression of cilia-related genes in NBs (Duncan et al. 2015). Given our knowledge of the likely existence of bivalent promoters in planarians, we can hypothesize cilia-related genes at least are maintained in a poised transcriptional state by bivalent histone modifications and are de-silenced following MLL1/2 knockdown. It would be informative for us to replicate the MLL1/2 phenotype and perform ChIP-seq on sorted NBs against the H3K4me3 and H3K27me3 marks, to investigate whether bivalent domains are affected and whether these domains are specific to cilia-related loci.

Moreover, if the FACS isolation of distinct planarians differentiated cellular subtypes becomes a possibility in the future, probing for the existence of bivalent domains beyond the NB compartment will be interesting. For instance, bivalent promoters, occupied by both PRC2 and paused RNA Pol II have been observed on non-neuronal TFs in neuronal cells (Ferrai et al. 2017). These bivalently marked genes are proposed to act as the major drivers of trans-differentiation towards non-neuronal fates (e.g. heart, bone, muscle). Thus, transcriptional poising in neuronal cells may be important in

the regulation of trans-differentiation to non-neuronal cellular fates. Indeed, bivalent histone modifications may also be important for the trans-differentiation of hindgut cells in *C. elegans* into motor neurons (Zuryn et al. 2014). It is generally difficult to prove the existence of trans-differentiation between differentiated cell types without transgenic reporters, but the existence of bivalent domains in differentiated planarian cell-types can provide a window in lending evidence for this.

6.3 Identifying gene regulatory networks in planarian cell types

As yet we have little understanding of the role regulatory elements presumably play in controlling the temporal expression of genes in different planarian cell-types. Consequently, in Chapter V, we describe preliminary efforts to identify 'open-chromatin' regions of the genome that have the potential to be enhancers using ATAC-seq. In the future, alternative methodologies such as DNase-seq and ChIP-seq against p300 will independently verify their existence. The aim would then be to elucidate the TFs that bind to these enhancers, and the target genes whose expression is controlled by them.

Firstly, one way of identifying TF binding sites it to perform ChIP-seq using monoclonal antibodies to specific planarian TFs. For example, ChIP-seq mapping profiles generated for the *C. elegans* TF EOR-1, were used to determine its dimeric GAGA DNA-binding motif. This motif was then found to significantly enriched in ATAC-seq peaks at the L3 developmental stage, where it is hypothesized to play a role in opening up chromatin configurations by the recruitment of the SWI/SNF and RSC chromatin remodeling complexes (Daugherty et al. 2017). The generation of planarian specific antibodies against TFs would allow for the unambiguous identification of cognate DNA motifs, which we would then search for in our ATAC-seq peaks. Alternatively, Protein Binding Microarrays (PBMs) can also be used to identify *in vitro* DNA binding sites of cloned and expressed putative TFs, by detecting the level of hybridization between TF protein and alternative k-mers of a defined length (Berger et al. 2006; Narasimhan et al. 2015). As a result, we would be able to correlate the

activity of specific TFs with the accessibility of their enhancers over the course of differentiation. Moreover, we would be able to test specific hypotheses concerning the evolution of enhancers in the planarian genome. Given that the genome is replete with TEs, one attractive theory explaining their retention in large numbers would be that they are contributing TF binding sites, therefore contributing to building complex gene regulatory networks that may relate to their stem cell biology and regenerative capacity (Britten et al. 1988; Bannert and Kurth 2004). In mammalian genomes, there is evidence that promoters and enhancers have been born out of this exaptation (Cordaux and Batzer 2009). For instance, primate specific LTRs contribute >30% of p53 binding sites found from a genomic analysis of p53 human ChIP-seq data (Wang et al. 2007), and, more broadly, a total of 20% of binding sites for 26 TFs found in mice and humans are embedded within TE sequences (Sundaram et al. 2014) .

Enhancers can be found at considerable distances away from their target gene promoters, and may not necessarily control their nearest genes (Spitz et al. 2003; Sagai 2005; Freire-Pritchett et al. 2017; Novo et al. 2018). It is generally accepted that long-range enhancer-promoter interactions are facilitated by DNA-looping interactions mediated by cohesin (Kagey et al. 2010; Dowen et al. 2014). However, the nature of the molecular determinants of chromosomal interactions are not well-understood, and *in silico* predictions of regulatory interactions are nontrivial. Moreover, multiple enhancers may affect a single gene target.

To facilitate the identification of enhancer targets, chromosome confirmation capture (3C) and derivatives of this technique have enabled the biochemical mapping of these DNA looping interactions (Dekker et al. 2002; Richmond et al. 2006). In 3C, chromatin is cross-linked with formaldehyde, then digested, and re-ligated in such a way that only DNA fragments that are covalently linked together form ligation products. The ligation products therefore not only contain information of where they originated in the genome, but also where in the 3D organization of the genome they reside, and as such enables the identification of topologically associated domains (TADs) (de Wit and de Laat 2012). In HiC, a modification is included in the 3C protocol such that

biotin is included in the ligation junction, which is then pulled-down with streptavidin beads, thus enabling selective purification of the chimeric DNA ligation junctions of TADs. Given the fact that HiC will identify all genome-wide DNA interactions, sequencing to sufficient depths (over -1billion reads for the human genome) is necessary to gain appropriate coverage to identify statistically significant TADs. To get around this shortcoming, biotinylated RNA probes complementary to annotated promoters in the genome of interest can be used to enrich for promoter-containing fragments in HiC libraries (Schoenfelder et al. 2015b; Mifsud et al. 2015; Schoenfelder et al. 2015a). In this way, the enhancers located both distally and more proximally can be linked to their target promoters. Utilizing Promoter Capture Hi-C as a complementary methodology to ATAC-seq can identify, with statistical significance, all promoter-enhancer interactions in FACS isolated populations in planarians.

6.4 Using single cell technologies to unpick regulatory interactions in heterogeneous cell populations

Single cell technologies in planarians have allowed gene expression to be better resolved at the cell-type level, rather than being reliant on the broad definitions offered by bulk FACS isolation. For instance, single-cell qPCR and RNA-seq has identified a large degree of cellular heterogeneity within the NB population, and more recently, Drop-seq has further elucidated that NBs can differentiate into at least 23 independent cell lineages (Plass et al. 2018).

Outside of planarians, single-cell ATAC-seq (scATAC-seq) has resolved the regulatory landscape of individual cells in various tissues as well as in entire animals (Buenrostro et al. 2015b; Cusanovich et al. 2015, 2018b; Pliner et al. 2018; Cao et al. 2018). These experiments are reliant on combinatorial indexing strategies that allow for thousands of nuclei to be uniquely barcoded, whilst not requiring the physical isolation of cells during the library construction process. Using traditional clustering methodologies, cells with similar accessibility profiles can be grouped and subsequently identified on the basis of known gene promoters being accessible. Moreover, Cicero – a recently developed

scATAC-seq analysis package implemented in R - can be used to identify promoter-enhancer interactions in individual cells (Pliner et al. 2018; Cusanovich et al. 2018a). Essentially, Cicero reports a co-accessibility score based on how correlated two sites are in the cells comprising distinct clusters. Moreover, the presence of a particular motif at a promoter can serve as a predictor for the presence of another motif at a Cicero-linked distally located enhancer. In this way, both individual enhancer-promoter interactions as well as interactions between TFs can be identified within the gene regulatory networks of single cells.

Consequently, we propose that the regulatory landscape of single planarian cells can be ascertained using the aforementioned scATAC-seq methodologies. This will yield better insights into the regulatory logic underlying the transcriptional heterogeneity within the NB cluster, by the identification, for example, of GRNs specific to different lineage-committed subsets. Moreover, scATAC-seq can be used to identify regulatory differences between closely related cell types (i.e. between neuronal subtypes) and when these regulatory differences arise during cellular commitment. Given that planarians represent an *in vivo* stem cell system, a range of experimental paradigms can be imagined in which to both observe and perturb the regulatory logic of individual cells including during regeneration, starvation, growth, and disease (González-Estévez et al. 2012; Mangel et al. 2016; Mihaylova et al. 2018)

Chapter VII

Materials and Methods

7.1 Generation of reference annotations for asexual *S. mediterranea* genome

Previous transcriptome assemblies - Oxford (ox_Smed_v2), Dresden (dd_smed_v4), SmedGD Asexual Transcriptome and Smed GD Unigenes - were downloaded from and PlanMine (Brandl et al. 2016) and Smed GD 2.0 (Robb et al. 2015). NCBI complete CDS sequences for *Schmidtea mediterranea* were also downloaded. Sequences were aligned to the SmedGD Asexual 1.1 genome with GMAP (Wu and Watanabe 2005) and consolidated with PASA. An independent reference assembly was also performed by mapping 164 available RNA-seq datasets with HISAT2 (Sirén et al. 2014) and assembly was performed with StringTie. The PASA consolidated and StringTie annotations were merged with StringTie.

An intron jaccard score (intersection of introns / union of introns) was calculated for all overlapping transcripts. Pair-wise jaccard similarity scores of 0.9 or greater were used to create a graph of similar annotations. From the resultant cliques of transcripts, one was chosen to be the representative transcript for the locus, by prioritizing transcript length, ORF length, and BLAST homology.

Strand information for annotations was assigned by utilizing strand-specific RNA-seq libraries generated by the Aboobaker lab, BLAST homology, and longest ORF length. TransDecoder was run utilizing Pfam and Uniprot as ORF retention criteria, to identify protein-coding transcripts in the annotations.

iPython notebook pertaining to asexual genome annotation is available at https://github.com/anishdattani12/Planarian-ChIP-seq-iPython-Notebooks/ and is contained in SupplementalFile2.html

7.2 Generation of reference annotations for sexual *S. mediterranea* genome

The new assembly of the *S. mediterranea* sexual genome was shared with us prior to its publication and associated annotations (Jochen Rink, personal communication). The Oxford (ox_Smed_v2), Dresden (dd_smed_v4), SmedGD Sexual Transcriptome, and Smed GD Unigenes were individually mapped to the sexual genome assembly with GMAP. Additionally, 183 publicly available RNA-seq datasets were mapped to the sexual genome assembly using hisat2, and assembled into potential transcripts and merged with StringTie. Additionally, 4 strand-specific libraries, generated by the Aboobaker lab, were used also separately mapped and assembled intro transcripts. The GTF/GFF files generated from the above processes were used as input into Mikado, which used a novel algorithm to generate genome annotations from multiple transcriptome assemblies, as well as leveraging additional data such as the position of ORFs and reliable splice junctions. We generated a BED file containing information on splice junctions using Portcullis on 11 mapped RNA-seq libraries. TransDecoder was used on the mapped de novo transcriptome assemblies to obtain ORF information. These multiple files were used to create a SQLite database to feed into the Mikado pipeline, whose output is to pick the best transcript models for a particular gene locus, and additionally bring back splice variants compatible with the primary isoform. A custom python script was used to rename loci in the output gff file.

We selected the longest isoform using the Perl script gff3_sp_keep_longest_isoform.pl in the Genome Assembly Annotation Service (GAAS) (https://github.com/NBISweden/GAAS). Representative transcripts were then used as the input to TransDecoder to select for protein coding-genes with >100aa length. These candidate proteins served as the input to a local copy of InterProScan to search for protein signatures and associated GO terms.

7.3 Repeat Masker annotation of genome and generation of Kimura histograms

In order to provide information about repeats, Schmidtea mediterranea both de novo and known repeats were searched for in the dd_Smed_g4 assembly. RepeatModeler 1.0.11 was run using standard parameters in order to assemble de novo repetitive elements using the results from RECON, RepeatScout and Tandem Repeats Finder (TRF). The consensi.fa.classified output file was concatenated with a *S. mediterranea* repeat library downloaded from RepBase to produce a combined repeat library. This served as the input query to the RepeatMasker to identify corresponding repeats in the sexual genome assembly. A perl script, parseRM.pl (available at https://github.com/4ureliek/Parsing-RepeatMasker-Outputs), was used to parse the raw alignment outputs from RepeatMasker (.align files) that contains the Kimura 2-paramater divergence metric (accounting for the extremely high rate of mutations at CpG sites). In case of overlaps (when a position could be aligned to more than one consensus sequence), the smallest percentage divergence is chosen for that position. This process allowed us to use the corrected percentage of divergence of each copy to the consensus from .alignfiles, and allowed us to split Kimura distance values into bins and plot based on TE type.

7.4 FACS proportional expression values generated for annotated loci

Kallisto (Bray et al. 2016) was used to pseudo-align RNA-seq libraries originating from four labs (Önal et al. 2012; Labbé et al. 2012; Van Wolfswinkel et al. 2014; Zhu et al. 2015; Duncan et al. 2015) to our expression-based annotation of the asexual *S. mediterranea* asexual and sexual genomes. This generated TPM values for each annotated locus. Sleuth was used to calculate a normalization factor for each library. For each locus, the TPM values of member transcripts (potential isoforms) were summed to generate a consensus TPM value and then normalized accordingly. Replicates within each lab dataset were then averaged.

Normalized TPM values for each lab dataset were converted to a proportional value as a representation of expression in FACS categories. We next calculated three sets of pairwise ratios (X1:X2, X1:Xins, X2:Xins) using these proportional values. Given two of the three ratios, a third ratio can be 'predicted'. Consequently, we calculated 3 'observed' ratios and 3 'predicted ratios'. A good Spearman's rank correlation was observed for the X2:Xins ratio and as such we kept these observed proportions, and calculated an inferred X1 proportion.

iPython notebook pertaining to FACS proportional analysis pipeline is available at https://github.com/anishdattani12/Planarian-ChIP-seq-iPython-Notebooks/ and is contained in SupplementalFile3.html

7.5 FACS isolation of planarian cell types

Prior to dissociation, 10x CMF (25.6 mM NaH_2P0_4. H_2O, 142.8 mM NaCl, 102.1 mM KCl, 94.2 mM $NaHCO_3$) and $CMFHE^{2+}$ (1x CMF, 3 mM EDTA pH 8.0, 0.5% Glucose, 15 mM Hepes pH 7.5) were prepared. Briefly, planarian cells were dissociated in batches of 40 worms. Worms (at least 7 mm in length) were placed on foil-covered petri-dish full with ice. All instant ocean residues were removed and worms were cut finely, trying to achieve a slurry. Worm pieces were transferred to a DNA lo-binding tube with the addition of 400 µl $CMFHE^{2+}$. An equal volume of papain digestion solution (30 U/ml) was added to the tube, and left to incubate at 25°C for 1 hr. Following incubation, worm pieces were mechanically dissociated to a cell suspension with a pipette. Suspension was centrifuged at 4°C for 5 mins at 500 g to pellet cells, and supernatant was replaced with 1ml of $CMFHE^{2+}$. One further wash ensured complete removal of papain solution. Following re-suspension, solution was passed sequentially through a 100 µm and 35 µm mesh to a pre-cooled tube. 50 µl of Hoechst 34580 (1 mg/ml) was added to a 1ml cell suspension, adjusting the concentration depending on visual inspection of cell density. 1 µl of calcein (0.2µg/ml) was added to a 1ml cell suspension, and incubated for 1hr at least in the dark. Prior to FACS analysis, 1µl of Propidium Iodide (10 µg/ml) was added prior to FACS gating on the FACS ARIA III equipped with violet laser for sorting. BD

FACS DIVA software was used for the sequential gating of cell populations as outlined in (Romero et al. 2012).

7.6 ChIP-seq and library preparation

For each experimental replicate, 600,000-700,000 planarian X1 cells were isolated, (sufficient for ChIP-seq of 3 histone marks and an input control) by utilisation of a published FACS protocol. We dissociated cells from an equal number of head, pharyngeal, and tail pieces from 3-day regenerating planarians or whole worms. All animals were starved for 2 weeks prior to dissociation.

Following FACS, cells were pelleted. The pellet was re-suspended in Nuclei Extraction Buffer (0.5% NP-40, 0.25% Triton X-100, 10 mM Tris-Cl pH 7.5, 3 mM $CaCl_2$, 0.25 mM Sucrose, 1 mM DTT, phosphatase cocktail inhibitor 2, phosphatase cocktail inhibitor 3). A 3% *Drosophila* S2 cell spike-in was added at this point. This was followed by 1% formaldehyde fixation for 7 mins, which was quenched with the addition of glycine to a final 125 mM concentration. The nuclei pellet was re-suspended in SDS lysis buffer (1% SDS, 50 mM Tris-Cl pH8.0, 10 mM EDTA) and incubated on ice, followed by the addition of ChIP dilution buffer. Samples were sonicated and $1/10^{th}$ volume of Triton X-100 was added to dilute SDS in the solution. Samples were pelleted, and supernatant was collected that contained the sonicated chromatin. Test de-crosslinking was performed on $1/8^{th}$ of the sonicated chromatin, and analysed using a TapeStation DNA HS tape to verify the DNA fragment range was between 100-500bp. If S2 cells had not been added earlier before chromatin preparation, a commercial *Drosophila* S2 chromatin (Active Motif 53083) spike-in was added at this point (at 3% of the amount of amount of *S. mediterranea* prepared chromatin)

Protein A-covered Dynabeads were used for immunoprecipitation (IP). 50 µl of Dynabeads were incubated overnight at 4°C with 7 µg of antibody (H3K4me3 Abcam ab8580; H3K36me3 Abcam ab9050; H3K4me1 Abcam ab8895; H3K27me3 Abcam ab6002; RNAPII-Ser5P ab5131) diluted in 0.5% BSA/PBS. Following incubation, Dynabeads were washed with 0.5 % BSA/PBS, and ¼ of the

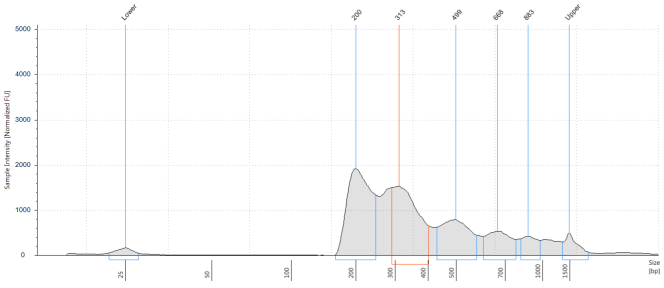
total isolated chromatin was added per IP. Following overnight incubation, washes were performed 6 times with RIPA buffer (50 mM HEPES-KOH pH 8.0, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% DOC, protease inhibitors). Dynabeads were washed with TE buffer and re-suspended in Elution Buffer (50 mM Tris-Cl pH 8.0, 10 mM EDTA, 1% SDS). Protein was separated from Dynabeads by incubating for 15 mins at 65°C on a shaking heating block at 1400rpm. Eluates were de-crosslinked at 65°C overnight. Input chromatin ($1/8^{th}$ of the total chromatin amount) was also de-crosslinked at this point. Following incubation, RNaseA (0.2 µg) and Proteinase K (0.2 µg) was added to each sample and incubated for 1 hr. DNA was purified with phenol:chloroform extraction followed by ethanol precipitation. DNA is re-suspended in TE and quantified with Qubit dsDNA HS kit. A NEB Ultra II kit was used for library preparation, and clean-up was performed with Agencourt AMPure XP beads. Samples were paired-end sequenced on the Illumina NextSeq.

7.7 ATAC-seq library preparation

For each experimental replicate, approximately 120,000-250,000 planarian X1, X2 or Xins cells were isolated by FACS. The primary protocol was based on the Buenrostro et al. (2013) paper. We first washed collected cells in 1X PBS and centrifuged at 1200 RPM, and supernatant was discarded. For cells undergoing lysis we added 50 µl of cold lysis buffer and pipetted up and down to re-suspend the cells (for 10 mM Tri-Cl (pH7.5), 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40). For lysed cells we centrifuged at 500RPM for 10 mins at 4°C. We discarded the cytoplasmic content within the supernatant and kept the nuclei pellet (very faint white dot, so we were careful not to remove all supernatant). For cells undergoing transposition without lysis, we centrifuged in PBS and removed supernatant, keeping the cell pellet. We next added 25 µl 2X TD Buffer, 2.5 µl Tn5 Transposase and 22.5 µl of nuclease-free H_2O. We mixed the solution up and down to re-suspend the pellets/nuclei and incubated at 37°C for 60 mins. We next isolated the DNA using a Zymogen Clean and Concentrator Kit, and eluted the transposed DNA in 10 µl of EB buffer. At this point we stored at -20°C or proceeded to PCR amplification and library purification. For each sample, we combined 10 µl of purified transposed DNA, 10 µl of nuclease-free water, 15 µl Nextera PCR Master Mix, 5 µl of

PCR primer cocktail, and 5 µl Index Primer 1, 5 µl Index Primer 2. We amplified samples with 14 PCR cycles (72°C 3 mins, 98°C 30 secs, [98°C 10 secs, 63°C 30 secs, 72°C 1 min]). We performed cleanup of libraries by AMPure bead purification. Briefly, we transferred the PCR sample to an Eppendorf tube, and added 1.8X volume of Agencourt AMPure XP beads by pipetting up and down 10X to mix thoroughly. We placed PCR-bead mixture on magnetic rack for 5 mins, discarded supernatant, and washed once with 200 µl 80% EtOH. After drying on rack, to ensure all EtOH removal, we re-suspended in 20 µl H_2O. We then checked for appropriate fragmentation patterns on the TapeStation (Agilent). Given this is a protocol that may need further adjustments, **Figure 1A** shows is an example of optimum fragmentation patterns obtained for the X2 FACS cells, and **Figure1B** shows a sub-optimal (but the best we can achieve thus far) tagmentation for the X1 FACS isolated cells.

A: X2 Whole Cells (approximately 219K)



B: X1 Whole Cells (approximately 130K)

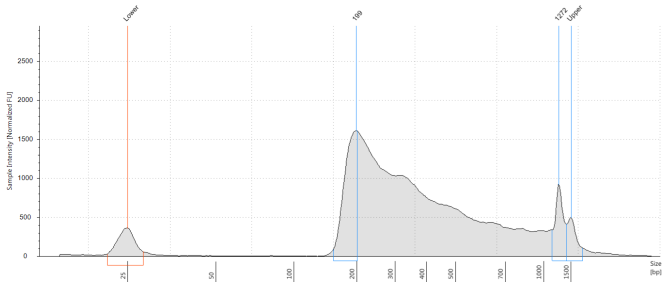


Figure 1: **A**. TapeStation profile of tagmented DNA library for approximately 219K FACS isolated X2 cells. Peak size takes into account the addition of a 135bp adapter sequence. Profile shows clear nucleosome phasing. **B.** TapeStation profile of tagmented DNA library for approximately 130K FACS isolated X1 cells. Peak size takes into account the addition of a 135bp adapter sequence. Profile for X1 cells shows weaker nucleosome phasing.

7.8 ChIP-seq analysis

Reads were trimmed with Trimmomatic (Bolger et al. 2014) and aligned to a concatenated file containing both the asexual *Schmidtea mediterranea* genome as well as the *Drosophila melanogaster* release 6 reference genome using BWA-MEM (Li and Durbin 2010) . Only uniquely mapping reads were considered further and those with a mapping score of greater than 10. Paired reads that map to both species were also removed. Picard tools-1.115 (https://broadinstitute.github.io/picard/) was used to remove duplicate reads. Reads were separated into sets that mapped to *Drosophila* or *S. mediterranea* using custom Python scripts. The number of reads aligning to the *Drosophila* genome were calculated for use in normalization calculations. For each paired or single map read, coordinates representing 100 bp at the center of the sequence were parsed and written to a BED file.

The genomecov function was used in BEDTools 2.27.0 (Quinlan and Hall 2010) to generate coverage tracks in bedgraph format. The resultant bedgraph file was converted to bigwig format using UCSC's bedgraphtoBigWig tool (Kent et al. 2002). deepTools2's computeMatrix was used to extract coverage around 2.5 kb or 5 kb either side of the annotated TSS for each annotated locus in 50 bp windows for each sample and corresponding input (Ramírez et al. 2016). A normalization factor was calculated using the number of mapped reads corresponding to the *Drosophila* spike-in to control for between IP technical variation (Orlando et al. 2014). A scaling factor for input ChIP-seq libraries was calculated (Diaz et al. 2012)using the deepTools2 Python API that uses the SES method. The mean normalized coverage was calculated for each sample and input. The normalized input coverage was subtracted from the normalized sample coverage to generate a final coverage track for downstream visualization and analyses.

To calculate the correlation of ChIP-seq signal coverage to proportional FACS expression, two vector values were calculated. The first vector was proportional FACS expression for all genomic loci. The second vector was ChIP-seq coverage at each 50 bp position 2.5 kb either side of the TSS.

A Spearman's rank correlation was performed on both vectors yielding a correlation value for the assayed position. The correlation value for each non-overlapping 50-bp window was then plotted on a graph.

For comparison of profiles between different epigenetic marks, a percentage coverage was calculated for each mark. The maximum coverage was found across all 5- or 10-kb regions for all loci. Absolute normalized coverage for each 50-bp window was then divided by the maximum coverage observed for that mark in the genome, resulting in a percentage coverage in each 50-bp window for each mark.

For calculation of a pausing index for individual genes, we divided normalized coverage to input 500 bp either side of the annotated TSS by the coverage between 500 bp and 2.5 kb downstream from the TSS.

iPython notebook pertaining to ChIP-seq analysis pipeline is available at https://github.com/anishdattani12/Planarian-ChIP-seq-iPython-Notebooks/ and is contained in SupplementalFile4.html

7.9 ATAC-seq peak calling and IGV visualization

Paired-end reads were trimmed with Trimmomatic 0.32 (Bolger et al. 2014) and mapped to the sexual genome assembly with BWA-MEM (Li and Durbin 2010). Only uniquely mapped reads and those with a quality score of greater than 10 were considered further. Duplicates were removed using Picard tools-1.115 (https://broadinstitute.github.io/picard/). The resultant BAM file was filtered to include reads of a user-defined size using samjdk.jar (http://lindenb.github.io/jvarkit/SamJdk.html). MACS2 was used to call peaks using the following option to find enriched Tn5 'cutting sites', and also to extend 5'ends of sequenced reads in both directions to smooth pile-up signals in a window of 200bps: --nomodel –shift -100 –extsize 200. Shared bed intervals between peak files of replicates were identified using bedtools's IntersectBed and unique and common intervals between FACS isolated libraries were identified using Intervene (Khan and Mathelier 2017). The bamCoverage tool in deepTools2 (Ramírez et al. 2016) was used to calculate coverage using a genome binsize of 10bp,

normalizing to RPKM, and with a smoot length of 50 bp. Resultant bigwig files were visualized in IGV. Motif and enrichment analysis was conducted with Homer (Benner et al. 2017).

7.10 Western Blot

To test the reactivity of conserved histone modification epitopes against commercial antibodies, a western blot was always carried out. 10 asexual planarians (5–7 mm) were thoroughly rinsed, the water removed, before addition of 90 µL PBS. 4 µL phenylmethylsulfonyl fluoride (PMSF) and 4µL 50x complete protease inhibitor (Roche) were added and the animals homogenized with a Kontes pellet pestle motor. 100 µL 2x Laemmli sample buffer (Sigma) and 10 µL Dithiothreitol (DTT) were added, the sample boiled at 100°C for 5 minutes, and then centrifuged at 13000 g for 5 minutes. Membranes were blocked in 5% dry skimmed milk in PBS with 0.05% Tween-20 for 1 h, and incubated with antibodies (usually at a concentration of 1:5000, or according to manufacturer's recommendations). Bands were detected using the SuperSignal West Pico kit (ThermoFisher).

7.11 In situ hybridisation

Whole mount in situ hybridization was performed as described by King and Newark (2013). Alkaline phosphatase or Peroxidase coupled anti-DIG/anti-FITC antibodies were diluted in blocking solution (1:200) and incubated at 4 °C overnight. Development was performed using either a NBT-BCIP (colorimetric) or TSA (fluorescent) reaction. Nuclei were stained with Hoechst 33342 (Sigma).

7.12 Immunochemistry for TUD-1

As in the King and Newmark (2013) protocol above, animals were washed with dH_2O, and incubated in 7.5% N-acetyl-L-cysteine PBS solution for 10 mins. Animals were then fixed in 4% formaldehyde solution and dehydrated in MeOH. Next, animals were bleached using 5% H_20_2/MeOH for 20h. Animals were rehydrated with consecutive washes with 75%, 50%, and 25% MeOH/PBSTx.

Following 2x 10min washes with PBSTx, animals were blocked with 1% Bovine serum albumin (BSA)/ PBSTx for 2 h at room temperature. Anti-SmedTUD-1 (rabbit polyclonal 1:200) was diluted in blocking solution and incubated at 4 °C overnight. Following 7-8hr washed with PBSTx, secondary antibodies were added diluted in blocking solution. Goat anti-rabbit IgG Alexa Fluor® 488 coupled (ThermoFisher) (1:400) was used for anti-SmedTUD1. Nuclei were stained with Hoechst 33342 (Sigma).

Bibliography

Bibliography:

Aboobaker AA. 2011. Planarian stem cells: A simple paradigm for regeneration. *Trends Cell Biol* **21**: 304–311.

Adamo A, Sesé B, Boue S, Castaño J, Paramonov I, Barrero MJ, Belmonte JCI. 2011. LSD1 regulates the balance between self-renewal and differentiation in human embryonic stem cells. *Nat Cell Biol* **13**: 652–661.

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**: 720–731.

Adler CE, Seidel CW, McKinney SA, Sánchez Alvarado A. 2014. Selective amputation of the pharynx identifies a FoxA-dependent regeneration program in planaria. *Elife* **3**: e02238.

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**: 489–493.

Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Françoijs KJ, Stunnenberg HG, Veenstra GJC. 2009. A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in Xenopus Embryos. *Dev Cell* **17**: 425–434.

Albalat R. 2008. Evolution of DNA-methylation machinery: DNA methyltransferases and methyl-DNA binding proteins in the amphioxus Branchiostoma floridae. *Dev Genes Evol* **218**: 691–701.

Albalat R, Martí-Solans J, Cañestro C. 2012. Dna methylation in amphioxus: From ancestral functions to new roles in vertebrates. *Brief Funct Genomics* **11**: 142–155.

Alié A, Hayashi T, Sugimura I, Manuel M, Sugano W, Mano A, Satoh N, Agata K, Funayama N. 2015. The ancestral gene repertoire of animal stem cells. *Proc Natl Acad Sci* 201514789.

Almuedo-Castillo M, Crespo X, Seebeck F, Bartscherer K, Salò E, Adell T. 2014. JNK Controls the Onset of Mitosis in Planarian Stem Cells and Triggers Apoptotic Cell Death Required for Regeneration and Remodeling ed. A.A. Aboobaker. *PLoS Genet* **10**: e1004400.

Ambrosone A, Marchesano V, Tino A, Hobmayer B, Tortiglione C. 2012. Hymyc1 downregulation promotes stem cell proliferation in hydra vulgaris. *PLoS One*.

Anava S, Posner R, Rechavi O. 2014. The soft genome. *Worm* **3**: e989798.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Aranda S, Mas G, Di Croce L. 2015. Regulation of gene transcription by Polycomb proteins. *Sci Adv* **1**: e1500737–e1500737.

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**: 203–207.

Ardehali MB, Mei A, Zobeck KL, Caron M, Lis JT, Kusch T. 2011. Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *EMBO J* **30**: 2817–2828.

Armstrong L, Saretzki G, Peters H, Wappler I, Evans J, Hole N, von Zglinicki T, Lako M. 2005. Overexpression of Telomerase Confers Growth Advantage, Stress Resistance, and Enhanced Differentiation of ESCs Toward the Hematopoietic Lineage. *Stem Cells* **23**: 516–529.

Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M, et al. 2006. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**: 532–538.

Baguñà J, Carranza S, Pala M, Ribera C, Giribet G, Arnedo MA, Ribas M, Riutort M. 1999. From morphology and karyology to molecules. New methods for taxonomical identification of asexual populations of freshwater planarians. A tribute to Professor Mario Benazzi. *Ital J Zool* **66**: 207–214.

Baguñá J, Romero R. 1981. Quantitative analysis of cell types during growth, degrowth and regeneration in the planarians Dugesia mediterranea and Dugesia tigrina. *Hydrobiologia* **84**: 181–194.

Bannert N, Kurth R. 2004. Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci U S A* **101 Suppl**: 14572–9.

Bar-Nur O, Russ HA, Efrat S, Benvenisty N. 2011. Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet beta cells. *Cell Stem Cell* **9**: 17–23.
Barrett SCH, Charlesworth D. 1991. Effects of a change in the level of inbreeding on the genetic load. *Nature* **352**: 522–524.
Basyuk E, Cross JC, Corbin J, Nakayama H, Hunter P, Nait-Oumesmar B, Lazzarini RA. 1999. Murine Gcm1 gene is expressed in a subset of placental trophoblast cells. *Dev Dyn*.
Baubec T, Ivánek R, Lienert F, Schübeler D. 2013. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **153**: 480–92.
Baum CM, Weissman IL, Tsukamoto AS, Buckle AM, Peault B. 1992. Isolation of a candidate human hematopoietic stem-cell population. *Proc Natl Acad Sci* **89**: 2804–2808.
Becker AJ, McCulloch EA, Till JE. 1963. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197**: 452–4.
Benner C, Heinz S, Glass CK. 2017. HOMER - Software for motif discovery and next generation sequencing analysis. *Http://HomerUcsdEdu/*.
Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**: 1429–1435.
Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**: 315–326.
Bhaumik SR, Smith E, Shilatifard A. 2007. Covalent modifications of histones during development and disease pathogenesis. *Nat Struct Mol Biol* **14**: 1008–1016.
Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6–21.
Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**: 91–99.
Blassberg RA, Felix DA, Tejada-Romero B, Aboobaker AA. 2013. PBX/extradenticle is required to re-establish axial structures and polarity during planarian regeneration. *Development* **140**: 730–739.
Bledau AS, Schmidt K, Neumann K, Hill U, Ciotta G, Gupta A, Torres DC, Fu J, Kranz A, Stewart AF, et al. 2014. The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Development* **141**: 1022–1035.
Boeger H, Griesenbeck J, Kornberg RD. 2008. Nucleosome Retention and the Stochastic Nature of Promoter Chromatin Remodeling for Transcription. *Cell* **133**: 716–726.
Boettiger AN, Bintu B, Moffitt JR, Wang S, Beliveau BJ, Fudenberg G, Imakaev M, Mirny LA, Wu CT, Zhuang X. 2016. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529**: 418–422.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.
Bonuccelli L, Rossi L, Lena A, Scarcelli V, Rainaldi G, Evangelista M, Iacopetti P, Gremigni V, Salvetti A. 2010. An RbAp48-like gene regulates adult stem cells in planarians. *J Cell Sci* **123**: 690–698.
Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, et al. 2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**: 349–353.
Brandl H, Moon HK, Vila-Farré M, Liu SY, Henry I, Rink JC. 2016. PlanMine - A mineable resource of planarian biology and biodiversity. *Nucleic Acids Res* **44**: D764–D773.
Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* **128**: 1089–1103.
Britten RJ, Baron WF, Stout DB, Davidson EH. 1988. Sources and evolution of human Alu repeated sequences. *Evolution (N Y)* **85**: 4770–4774.
Brookes E, De Santiago I, Hebenstreit D, Morris KJ, Carroll T, Xie SQ, Stock JK, Heidemann M, Eick D, Nozaki N, et al. 2012. Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell* **10**: 157–170.

Brookes E, Pombo A. 2009. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Rep* **10**: 1213–1219.
Bueno A, Russell P. 1992. Dual functions of CDC6: a yeast protein required for DNA replication also inhibits nuclear division. *EMBO J* **11**: 2167–76.
Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.
Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015a. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. In *Current Protocols in Molecular Biology*, pp. 21.29.1-21.29.9, John Wiley & Sons, Inc., Hoboken, NJ, USA.
Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015b. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**: 486–490.
Bulmer M. 1986. Neighboring base effects on substitution rates in pseudogenes. *Mol Biol Evol* **3**: 322–329.
Bulut-Karslioglu A, De La Rosa-Velázquez IA, Ramirez F, Barenboim M, Onishi-Seebacher M, Arand J, Galán C, Winter GE, Engist B, Gerle B, et al. 2014. Suv39h-Dependent H3K9me3 Marks Intact Retrotransposons and Silences LINE Elements in Mouse Embryonic Stem Cells. *Mol Cell* **55**: 277–290.
Calo E, Wysocka J. 2013. Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell* **49**: 825–837.
Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, et al. 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (80-)* **361**: 1380–1385.
Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, et al. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**: 661–667.
Cao R, Tsukada Y, Zhang Y. 2005. Role of Bmi-1 and Ring1A in H2A Ubiquitylation and Hox Gene Silencing. *Mol Cell* **20**: 845–854.
Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y. 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**: 1039–43.
Capy P, Langin T, Higuet D, Maurer P, Bazin C. 1997. Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica* **100**: 63–72.
Carpenter KS, Morita M, Best JB. 1974. Ultrastructure of the photoreceptor of the planarian Dugesia dorotocephala. *Cell Tissue Res* **148**.
Carranza S, Littlewood DTJ, Clough KA, Ruiz-Trillo I, Baguñà J, Riutort M. 1998. A robust molecular phylogeny of the Tricladida (Platyhelminthes: Seriata) with a discussion on morphological synapomorphies. *Proc R Soc London Ser B Biol Sci* **265**: 631–640.
Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia W-J, Anderson S, Yates J, Washburn MP, et al. 2005. Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell* **123**: 581–592.
Cebrià F, Kudome T, Nakazawa M, Mineta K, Ikeo K, Gojobori T, Agata K. 2002. The expression of neural-specific genes reveals the structural and molecular complexity of the planarian central nervous system. *Mech Dev* **116**: 199–204.
Cebrià F, Vispo M, Newmark P, Bueno D, Romero R. 1997. Myocyte differentiation and body wall muscle regeneration in the planarian Girardia tigrina. *Dev Genes Evol* **207**: 306–316.
Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, et al. 2010. The dynamic genome of Hydra. *Nature* **464**: 592–596.
Chen C-CG, Wang IE, Reddien PW. 2013. pbx is required for pole and eye regeneration in planarians. *Development* **140**: 719–729.
Cheng J, Blum R, Bowman C, Hu D, Shilatifard A, Shen S, Dynlacht BD. 2014. A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Mol Cell* **53**: 979–992.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71–86.
Clark SJ, Argelaguet R, Kapourani C-AA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, et al. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**: 781.
Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W. 2016. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* **17**: 72.
Cleary SP, Jeck WR, Zhao X, Chen K, Selitsky SR, Savich GL, Tan TX, Wu MC, Getz G, Lawrence MS, et al. 2013. Identification of driver genes in hepatocellular carcinoma by exome sequencing. *Hepatology* **58**: 1693–1702.
Colloms SD, Van Luenen HGAM a m, PIasterk RHAA. 1994. DNA binding activities of the caenorhabditis elegans Tc3 transposase. *Nucleic Acids Res* **22**: 5548–5554.
Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691–703.
Corden JL. 2013. RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. *Chem Rev* **113**: 8423–8455.
Coward SJ. 1974. Chromatoid bodies in somatic cells of the planarian: Observations on their behavior during mitosis. *Anat Rec* **180**: 533–545.
Cowles MW, Omuro KC, Stanley BN, Quintanilla CG, Zayas RM. 2014. COE Loss-of-Function Analysis Reveals a Genetic Program Underlying Maintenance and Regeneration of the Nervous System in Planarians ed. A.A. Aboobaker. *PLoS Genet* **10**: e1004746.
Cramer JM, Pohlmann D, Gomez F, Mark L, Kornegay B, Hall C, Siraliev-Perez E, Walavalkar NM, Sperlazza MJ, Bilinovich S, et al. 2017. Methylation specific targeting of a chromatin remodeling complex from sponges to humans. *Sci Rep* **7**: 40674.
Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-)* **348**: 910–914.
Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. 2018a. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**: 1309-1324.e18.
Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, et al. 2018b. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**: 538–542.
Dattani A, Kao D, Mihaylova Y, Abnave P, Hughes S, Lai A, Sahu S, Aboobaker AA. 2018. Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells. *Genome Res* **28**: 1543–1554.
Dattani A, Sridhar D, Aziz Aboobaker A. 2019. Planarian flatworms as a new model system for understanding the epigenetic regulation of stem cell pluripotency and differentiation. *Semin Cell Dev Biol* **87**: 79–94.
Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. 2017. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res* **27**: 2096–2107.
De Robertis EM. 2010. Wnt Signaling in Axial Patterning and Regeneration: Lessons from Planaria. *Sci Signal* **3**: pe21–pe21.
de Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol* **8**.
de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**: 11–24.
Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–1022.
Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**: 1306–11.
Denissov S, Hofemeister H, Marks H, Kranz A, Ciotta G, Singh S, Anastassiadis K, Stunnenberg

HG, Stewart AF. 2014. Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development* **141**: 526–537.
Deniz Ö, Flores O, Aldea M, Soler-López M, Orozco M. 2016. Nucleosome architecture throughout the cell cycle. *Sci Rep* **6**: 19729.
Denslow SA, Wade PA. 2007. The human Mi-2/NuRD complex and gene regulation. *Oncogene* **26**: 5433–5438.
Diaz A, Park K, Lim DA, Song JS. 2012. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol Biol* **11**.
Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.
Dorighi KM, Swigut T, Henriques T, Bhanu N V., Scruggs BS, Nady N, Still CD, Garcia BA, Adelman K, Wysocka J. 2017. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**: 568-576.e4.
Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, et al. 2014. Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell* **159**: 374–387.
Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I, et al. 2014. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res* **24**: 1854–1868.
Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**: 560–561.
Duncan EM, Chitsazan AD, Seidel CW, Sánchez Alvarado A. 2015. Set1 and MLL1/2 Target Distinct Sets of Functionally Different Genomic Loci In Vivo. *Cell Rep* **13**: 2741–55.
Dunwell TL, Pfeifer GP. 2014. Drosophila genomic methylation: new evidence and new questions. *Epigenomics* **6**: 459–61.
Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E. 2012. ENCODE explained. *Nature* **489**: 52–54.
Egger B, Lapraz F, Tomiczek B, Müller S, Dessimoz C, Girstmair J, Škunca N, Rawlinson KA, Cameron CB, Beli E, et al. 2015. A Transcriptomic-Phylogenomic Analysis of the Evolutionary Relationships of Flatworms. *Curr Biol* **25**: 1347–1353.
Eisenhoffer GT, Kang H, Alvarado AS. 2008. Molecular Analysis of Stem Cells and Their Descendants during Cell Turnover and Regeneration in the Planarian Schmidtea mediterranea. *Cell Stem Cell* **3**: 327–339.
Eissenberg JC, Shilatifard A. 2010. Histone H3 lysine 4 (H3K4) methylation in development and differentiation. *Dev Biol* **339**: 240–249.
Emerling BM, Bonifas J, Kratz CP, Donovan S, Taylor BR, Green ED, Le Beau MM, Shannon KM. 2002. MLL5, a homolog of Drosophila trithorax located within a segment of chromosome band 7q22 implicated in myeloid leukemia. *Oncogene* **21**: 4849–4854.
Ernst C, Odom DT, Kutter C. 2017. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nat Commun* **8**: 1411.
Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**: 364–376.
Fang F, Xia N, Angulo B, Carey J, Cady Z, Durruthy-Durruthy J, Bennett T, Sebastiano V, Reijo Pera RA. 2018. A distinct isoform of ZNF207 controls self-renewal and pluripotency of human embryonic stem cells. *Nat Commun* **9**: 4384.
Ferrai C, Torlai Triglia E, Risner-Janiczek JR, Rito T, Rackham OJ, de Santiago I, Kukalev A, Nicodemi M, Akalin A, Li M, et al. 2017. RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation. *Mol Syst Biol* **13**: 946.
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405.
Fincher CT, Wurtzel O, de Hoog T, Kravarik KM, Reddien PW. 2018. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science (80-)* **360**: eaaq1736.
Francis NJ, Kingston RE, Woodcock CL. 2004. Chromatin compaction by a polycomb group protein complex. *Science* **306**: 1574–7.

Freire-Pritchett P, Schoenfelder S, Várnai C, Wingett SW, Cairns J, Collier AJ, García-Vílchez R, Furlan-Magaril M, Osborne CS, Fraser P, et al. 2017. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6**.
Gaiti F, Jindrich K, Fernandez-Valverde SL, Roper KE, Degnan BM, Tanurdžić M. 2017. Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. *Elife* **6**.
Gan Q, Schones DE, Ho Eun S, Wei G, Cui K, Zhao K, Chen X. 2010. Monovalent and unpoised status of most genes in undifferentiated cell-enriched Drosophila testis. *Genome Biol* **11**: R42.
Gaviño MA, Reddien PW. 2011. A Bmp/Admp regulatory circuit controls maintenance and regeneration of dorsal-ventral polarity in planarians. *Curr Biol* **21**: 294–299.
Geyer KK, Chalmers IW, Mackintosh N, Hirst JE, Geoghegan R, Badets M, Brophy PM, Brehm K, Hoffmann KF. 2013. Cytosine methylation is a conserved epigenetic feature found throughout the phylum Platyhelminthes. *BMC Genomics* **14**: 462.
Geyer KK, Rodríguez López CM, Chalmers IW, Munshi SE, Truscott M, Heald J, Wilkinson MJ, Hoffmann KF. 2011. Cytosine methylation regulates oviposition in the pathogenic blood fluke Schistosoma mansoni. *Nat Commun* **2**: 424.
Gil J, O'Loghlen A. 2014. PRC1 complex diversity: where is it taking us? *Trends Cell Biol* **24**: 632–641.
Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**: 199–202.
Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of RegulatoryElements) isolates active regulatory elements from humanchromatin. *Genome Res* **17**: 877–885.
Gold DA, Jacobs DK. 2013. Stem cell dynamics in Cnidaria: are there unifying principles? *Dev Genes Evol* **223**: 53–66.
Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE, Bestor TH. 2006. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science (80-)* **311**: 395–398.
González-Estévez C, Felix DA, Rodríguez-Esteban G, Aziz Aboobaker A. 2012. Decreased neoblast progeny and increased cell death during starvation-induced planarian degrowth. *Int J Dev Biol* **56**: 83–91.
Gonzalez-Estevez C, Momose T, Gehring WJ, Salo E. 2003. Transgenic planarian lines obtained by electroporation using transposon-derived vectors and an eye-specific GFP marker. *Proc Natl Acad Sci* **100**: 14046–14051.
Grivna ST, Beyret E, Wang Z, Lin H. 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**: 1709–1714.
Grohme MA, Schloissnig S, Rozanski A, Pippel M, Young GR, Winkler S, Brandl H, Henry I, Dahl A, Powell S, et al. 2018. The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. *Nature* **554**: 56–61.
Grondin B, Bazinet M, Aubry M. 1996. The KRAB zinc finger gene ZNF74 encodes an RNA-binding protein tightly associated with the nuclear matrix. *J Biol Chem* **271**: 15458–15467.
Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* **315**: 1587–90.
Günther K, Rust M, Leers J, Boettger T, Scharfe M, Jarek M, Bartkuhn M, Renkawitz R. 2013. Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences. *Nucleic Acids Res* **41**: 3010–3021.
Guo T, Peters AHFM, Newmark PA. 2006. A bruno-like Gene Is Required for Stem Cell Maintenance in Planarians. *Dev Cell* **11**: 159–169.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.
Hainer SJ, McCannell KN, Yu J, Ee L-S, Zhu LJ, Rando OJ, Fazzio TG. 2016. DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells. *Elife* **5**: e21964.
Hallez P. 1892. Catalogue des Turbellariés (Rhabdocoelides, Triclades et Dendrocoelides) du Nord

de la France & de la Cote Boulonnaise. *Rev Biol Nord Fra*.
Hamilton AT. 2006. Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res* **16**: 584–594.
Haren L, Ton-Hoang B, Chandler M. 1999. Integrating DNA: Transposases and Retroviral Integrases. *Annu Rev Microbiol* **53**: 245–281.
Harikumar A, Meshorer E. 2015. Chromatin remodeling and bivalent histone modifications in embryonic stem cells. *EMBO Rep* **in press**: 1609–1619.
Harley VR, Lovell-Badge R, Goodfellow PN. 1994. Definition of a consensus DNA binding site for SRY. *Nucleic Acids Res* **22**: 1500–1501.
Hartwell LH, Mortimer RK, Culotti J, Culotti M. 1973. Genetic Control of the Cell Division Cycle in Yeast: V. Genetic Analysis of cdc Mutants. *Genetics* **74**: 267–86.
Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S, et al. 2010. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**: 479–491.
Hawkins RD, Hon GC, Yang C, Antosiewicz-Bourget JE, Lee LK, Ngo QM, Klugman S, Ching KA, Edsall LE, Ye Z, et al. 2011. Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. *Cell Res* **21**: 1393–1409.
Hayashi T, Asami M, Higuchi S, Shibata N, Agata K. 2006. Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. *Dev Growth Differ* **48**: 371–380.
Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res - Fundam Mol Mech Mutagen* **616**: 46–59.
Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311–318.
Hemmrich G, Khalturin K, Boehm AM, Puchert M, Anton-Erxleben F, Wittlieb J, Klostermeier UC, Rosenstiel P, Oberg HH, Domazet-Lošo T, et al. 2012. Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. *Mol Biol Evol* **29**: 3267–3280.
Hendrich B, Bird A. 1998. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Mol Cell Biol* **18**: 6538–6547.
Hendrich B, Tweedie S. 2003. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* **19**: 269–277.
Hickman AB, Perez ZN, Zhou L, Musingarimi P, Ghirlando R, Hinshaw JE, Craig NL, Dyda F. 2005. Molecular architecture of a eukaryotic DNA transposase. *Nat Struct Mol Biol* **12**: 715–721.
Hirano T. 2012. Condensins: universal organizers of chromosomes with diverse functions. *Genes Dev* **26**: 1659–1678.
Ho MCW, Quintero-Cadena P, Sternberg PW. 2017. Genome-wide discovery of active regulatory elements and transcription factor footprints in Caenorhabditis elegans using DNase-seq. *Genome Res* **27**: 2108–2119.
Hogan GJ, Lee C-K, Lieb JD. 2006. Cell Cycle–Specified Fluctuation of Nucleosome Occupancy at Gene Promoters. *PLoS Genet* **2**: e158.
Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**: 1419–1428.
Hosoya T, Takizawa K, Nitta K, Hotta Y. 1995. Glial cells missing: A binary switch between neuronal and glial determination in drosophila. *Cell* **82**: 1025–1036.
Hu D, Gao X, Morgan MA, Herz H-M, Smith ER, Shilatifard A. 2013a. The MLL3/MLL4 Branches of the COMPASS Family Function as Major Histone H3K4 Monomethylases at Enhancers. *Mol Cell Biol* **33**: 4745–4754.
Hu D, Garruss AS, Gao X, Morgan MA, Cook M, Smith ER, Shilatifard A. 2013b. The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat Struct Mol Biol* **20**: 1093–1097.
Hu G, Wade PA. 2012. NuRD and pluripotency: A complex balancing act. *Cell Stem Cell* **10**: 497–

503.
Huang H, Li Y, Szulwach KE, Zhang G, Jin P, Chen D. 2014. AGO3 Slicer activity regulates mitochondria-nuage localization of Armitage and piRNA amplification. *J Cell Biol* **206**: 217–230.
Hubert A, Henderson JM, Cowles MW, Ross KG, Hagen M, Anderson C, Szeterlak CJ, Zayas RM. 2015. A functional genomics screen identifies an Importin-α homolog as a regulator of stem cell function and tissue patterning during planarian regeneration. *BMC Genomics* **16**: 769.
Hubert A, Henderson JM, Ross KG, Cowles MW, Torres J, Zayas RM. 2013. Epigenetic regulation of planarian stem cells by the SET1/MLL family of histone methyltransferases. *Epigenetics* **8**: 79–91.
Hughes JR, Roberts N, Mcgowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R, Higgs DR. 2014. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* **46**: 205–212.
Hyman LH. 1951. The invertebrates: Platyhelminthes and Rhynchocoela. *New York, McGraw Hill*.
Iglesias M, Gomez-Skarmeta JL, Saló E, Adell T. 2008. Silencing of Smed-betacatenin1 generates radial-like hypercephalized planarians. *Development* **135**: 1215–21.
Ishii S. 1980. The ultrastructure of the protonephridial flame cell of the freshwater planarian Bdellocephala brunnea. *Cell Tissue Res* **206**: 441–449.
Jaber-Hijazi F, Lo PJKP, Mihaylova Y, Foster JM, Benner JS, Tejada Romero B, Chen C, Malla S, Solana J, Ruzov A, et al. 2013. Planarian MBD2/3 is required for adult stem cell pluripotency independently of DNA methylation. *Dev Biol* **384**: 141–153.
Jager M, Quéinnec E, Le Guyader H, Manuel M. 2011. Multiple Sox genes are expressed in stem cells or in differentiating neuro-sensory cells in the hydrozoan Clytia hemisphaerica. *Evodevo*.
Jänes J, Dong Y, Schoof M, Serizay J, Appert A, Cerrato C, Woodbury C, Chen R, Gemma C, Huang N, et al. 2018. Chromatin accessibility dynamics across C. elegans development and ageing. *Elife* **7**.
Jones BW, Fetter RD, Tear G, Goodman CS. 1995. glial cells missing: a genetic switch that controls glial versus neuronal fate. *Cell* **82**: 1013–1023.
Jones DTW, Jäger N, Kool M, Zichner T, Hutter B, Sultan M, Cho Y-J, Pugh TJ, Hovestadt V, Stütz AM, et al. 2012. Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**: 100–105.
Joshi AA, Struhl K. 2005. Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to pol II elongation. *Mol Cell* **20**: 971–978.
Juliano C, Wang J, Lin H. 2011. Uniting Germline and Stem Cells: The Function of Piwi Proteins and the piRNA Pathway in Diverse Organisms. *Annu Rev Genet* **45**: 447–469.
Juliano CE, Reich A, Liu N, Gotzfried J, Zhong M, Uman S, Reenan RA, Wessel GM, Steele RE, Lin H. 2014. PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells. *Proc Natl Acad Sci* **111**: 337–342.
Juliano CE, Swartz SZ, Wessel GM. 2010. A conserved germline multipotency program. *Development* **137**: 4113–4126.
Jung YL, Luquette LJ, Ho JWK, Ferrari F, Tolstorukov M, Minoda A, Issner R, Epstein CB, Karpen GH, Kuroda MI, et al. 2014. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res* **42**: e74–e74.
Jürgens G. 1985. A group of genes controlling the spatial expression of the bithorax complex in Drosophila. *Nature* **316**: 153–155.
Jurkowski TP, Jeltsch A. 2011. On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PLoS One* **6**: e28104.
Juven-Gershon T, Kadonaga JT. 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **339**: 225–229.
Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430–435.
Kang H, Jung YL, McElroy KA, Zee BM, Wallace HA, Woolnough JL, Park PJ, Kuroda MI.

2017. Bivalent complexes of PRC1 with orthologs of BRD4 and MOZ/MORF target developmental genes in Drosophila. *Genes Dev* **31**: 1988–2002.
Kao D, Felix D, Aboobaker A. 2013. The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. *BMC Genomics* **14**: 797.
Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, et al. 2011. DNA Methylation and SETDB1/H3K9me3 Regulate Predominantly Distinct Sets of Genes, Retroelements, and Chimeric Transcripts in mESCs. *Cell Stem Cell* **8**: 676–687.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.
Khan A, Mathelier A. 2017. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics* **18**: 287.
Kharchenko P V, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**: 480–5.
King RS, Newmark PA. 2013. In situ hybridization protocol for enhanced detection of gene expression in the planarian Schmidtea mediterranea. *BMC Dev Biol* **13**.
Knight RD, Shimeld SM. 2001. Identification of conserved C2H2 zinc-finger gene families in the Bilateria. *Genome Biol* **2**: RESEARCH0016.
Kouzarides T. 2007. Chromatin Modifications and Their Function. *Cell* **128**: 693–705.
Krishna S, Palakodeti D, Solana J. 2019. Post-transcriptional regulation in planarian stem cells. *Semin Cell Dev Biol* **87**: 69–78.
Krogan NJ, Dover J, Khorrami S, Greenblatt JF, Schneider J, Johnston M, Shilatifard A. 2002. COMPASS, a histone H3 (lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* **277**: 10753–10755.
Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D. 2002. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev* **16**: 2893–905.
Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DDR, Ross EJ, Voisin V, Bader GD, Blencowe BJ, et al. 2012. A Comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells* **30**: 1734–1745.
Lai AG, Aboobaker AA. 2018. EvoRegen in animals : Time to uncover deep conservation or convergence of adult stem cell evolution and regenerative processes. *Dev Biol* **433**: 118–131.
Lai AG, Kosaka N, Abnave P, Sahu S, Aboobaker AA. 2018. The abrogation of condensin function provides independent evidence for defining the self-renewing population of pluripotent stem cells. *Dev Biol* **433**: 218–226.
Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.
Lander R, Petersen CP. 2016. Wnt, Ptk7, and FGFRL expression gradients control trunk positional identity in planarian regeneration. *Elife* **5**: e12850.
Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**: 1813–1831.
Lapan SW, Reddien PW. 2011. Dlx and sp6-9 control optic cup regeneration in a prototypic eye. *PLoS Genet* **7**.
Lapan SW, Reddien PW. 2012. Transcriptome Analysis of the Planarian Eye Identifies ovo as a Specific Regulator of Eye Regeneration. *Cell Rep* **2**: 294–307.
Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**: 204–220.
Le Thomas A, Stuwe E, Li S, Du J, Marinov G, Rozhkov N, Chen YCA, Luo Y, Sachidanandam R, Toth KF, et al. 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes Dev* **28**: 1667–1680.
Lee J-E, Wang C, Xu S, Cho Y-W, Wang L, Feng X, Baldridge A, Sartorelli V, Zhuang L, Peng

W, et al. 2013. H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *Elife* **2**: e01503.
Lee J, Kim D-H, Lee S, Yang Q-H, Lee DK, Lee S-K, Roeder RG, Lee JW. 2009. A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proc Natl Acad Sci U S A* **106**: 8513–8.
Lee S-M, Lee J, Noh K-M, Choi W-Y, Jeon S, Oh GT, Kim-Ha J, Jin Y, Cho S-W, Kim Y-J. 2017. Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. *Proc Natl Acad Sci* **114**: E1885–E1894.
Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell*.
Lesch BJ, Dokshin GA, Young RA, McCarrey JR, Page DC. 2013. A set of genes critical to development is epigenetically poised in mouse germ cells from fetal stages through completion of meiosis. *Proc Natl Acad Sci* **110**: 16061–16066.
Lesch BJ, Page DC. 2014. Poised chromatin in the mammalian germ line. *Development* **141**: 3619–3626.
Lesch BJ, Silber SJ, McCarrey JR, Page DC. 2016. Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat Genet* **48**: 888–894.
Levin HL, Moran J V. 2011. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* **12**: 615–627.
Lewis EB. 1978. A gene complex controlling segementation in Drosophila. *Nature* **276**: 565–570.
Lewis SH, Salmela H, Obbard DJ. 2016. Duplication and diversification of dipteran argonaute genes, and the evolutionary divergence of Piwi and Aubergine. *Genome Biol Evol* **8**: 507–518.
Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**: 589–595.
Li J, Moazed D, Gygi SP. 2002. Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. *J Biol Chem* **277**: 49383–8.
Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-)* **326**: 289–293.
Liu J, Wu X, Zhang H, Pfeifer GP, Lu Q. 2017. Dynamics of RNA Polymerase II Pausing and Bivalent Histone H3 Methylation during Neuronal Differentiation in Brain Development. *Cell Rep* **20**: 1307–1318.
Liu L, Cheung TH, Charville GW, Hurgo BMC, Leavitt T, Shih J, Brunet A, Rando TA. 2013. Chromatin Modifications as Determinants of Muscle Stem Cell Quiescence and Chronological Aging. *Cell Rep* **4**: 189–204.
Lu X, Zhao BS, He C. 2015. TET family proteins: Oxidation activity, interacting molecules, and functions in diseases. *Chem Rev* **115**: 2225–2239.
Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**: 251–260.
Lyko F, Ramsahoye BH, Jaenisch R. 2000. DNA methylation in Drosophila melanogaster. *Nature* **408**: 538–540.
MacRae EK. 1967. The fine structure of sensory receptor processes in the auricular epithelium of the planarian, Dugesia tigrina. *Zeitschrift f??r Zellforsch und Mikroskopische Anat*.
Malik S, Roeder RG. 2010. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat Rev Genet* **11**: 761–772.
Mangel M, Bonsall MB, Aboobaker A. 2016. Feedback control in planarian stem cell systems. *BMC Syst Biol* **10**: 17.
Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. 2017. Efficient and accurate detection of splice junctions from RNAseq with Portcullis. *bioRxiv*.
Marheineke K, Krude T. 1998. Nucleosome assembly activity and intracellular localization of human CAF-1 changes during the cell division cycle. *J Biol Chem* **273**: 15279–86.
Marks H, Kalkan T, Menafra R, Denissov S, Jones K, Hofemeister H, Nichols J, Kranz A, Francis Stewart A, Smith A, et al. 2012. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**: 590–604.
Marsit CJ. 2015. Influence of environmental exposure on human epigenetic regulation. *J Exp Biol*

218: 71–79.
Martens JHA, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T. 2005. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J* **24**: 800–812.
Martin Gonzalez J, Morgani SM, Bone RA, Bonderup K, Abelchian S, Brakebusch C, Brickman JM. 2016. Embryonic Stem Cell Culture Conditions Support Distinct States Associated with Different Developmental Stages and Potency. *Stem Cell Reports* **7**: 177–191.
May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, et al. 2012. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**: 89–93.
McKanna JA. 1968. Fine structure of the protonephridial system in planaria. *Zeitschrift f□r Zellforsch und Mikroskopische Anat* **92**: 509–523.
Mellén M, Ayata P, Dewell S, Kriaucionis S, Heintz N. 2012. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**: 1417–1430.
Menafra R, Stunnenberg HG. 2014. MBD2 and MBD3: elusive functions and mechanisms. *Front Genet* **5**: 428.
Mendizabal I, Yi S V. 2016. Whole-genome bisulfite sequencing maps from multiple human tissues reveal novel CpG islands associated with tissue-specific regulation. *Hum Mol Genet* **25**: 69–82.
Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**: 598–606.
Mihaylova Y, Abnave P, Kao D, Hughes S, Lai A, Jaber-Hijazi F, Kosaka N, Aboobaker AA. 2018. Conservation of epigenetic regulation by the MLL3/4 tumour suppressor in planarian pluripotent stem cells. *Nat Commun* **9**: 3633.
Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF, Shilatifard A. 2001. COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc Natl Acad Sci U S A* **98**: 12902–7.
Milne TA, Briggs SD, Brock HW, Martin ME, Gibbs D, Allis CD, Hess JL. 2002. MLL targets SET domain methyltransferase activity to Hox gene promoters. *Mol Cell* **10**: 1107–1117.
Mitra R, Fain-Thornton J, Craig NL. 2008. piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J* **27**: 1097–1109.
Mohan M, Herz H-M, Smith ER, Zhang Y, Jackson J, Washburn MP, Florens L, Eissenberg JC, Shilatifard A. 2011. The COMPASS Family of H3K4 Methylases in Drosophila. *Mol Cell Biol* **31**: 4310–4318.
Molinaro AM, Pearson BJ. 2016. In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. *Genome Biol* **17**: 87.
Morgan TH. 1898. Experimental studies of the regeneration of Planaria maculata. *Arch für Entwickelungsmechanik der Org* **7**: 364–397.
Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M, et al. 2011. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**: 298–303.
Morita M, Best JB. 1984. Electron microscopic studies of planarian regeneration. IV. Cell division of neoblasts inDugesia dorotocephala. *J Exp Zool* **229**: 425–436.
Morita M, Best JB, Noel J. 1969. Electron microscopic studies of planarian regeneration. *J Ultrastruct Res* **27**: 7–23.
Müller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, Miller EL, O'Connor MB, Kingston RE, Simon JA. 2002. Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell* **111**: 197–208.
Murakawa Y, Yoshihara M, Kawaji H, Nishikawa M, Zayed H, Suzuki H, Hayashizaki Y. 2016. Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. *Trends Genet* **32**: 76–88.
Nagaich AK, Walker DA, Wolford R, Hager GL. 2004. Rapid periodic binding and displacement of the glucocorticoid receptor during chromatin remodeling. *Mol Cell* **14**: 163–174.
Narasimhan K, Lambert SA, Yang AWH, Riddell J, Mnaimneh S, Zheng H, Albu M, Najafabadi

HS, Reece-Hoyes JS, Fuxman Bass JI, et al. 2015. Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. *Elife* **4**.
Newmark PA, Alvarado AS. 2002. Not your father's planarian: A classic model enters the era of functional genomics. *Nat Rev Genet*.
Newmark PA, Sánchez Alvarado A. 2000. Bromodeoxyuridine Specifically Labels the Regenerative Stem Cells of Planarians. *Dev Biol* **220**: 142–153.
Nitarska J, Smith JG, Sherlock WT, Hillege MMG, Nott A, Barshop WD, Vashisht AA, Wohlschlegel JA, Mitter R, Riccio A. 2016. A Functional Switch of NuRD Chromatin Remodeling Complex Subunits Regulates Mouse Cortical Development. *Cell Rep* **17**: 1683–1698.
Novo CL, Javierre B-M, Cairns J, Segonds-Pichon A, Wingett SW, Freire-Pritchett P, Furlan-Magaril M, Schoenfelder S, Fraser P, Rugg-Gunn PJ. 2018. Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. *Cell Rep* **22**: 2615–2627.
O'Shaughnessy-Kirwan A, Signolet J, Costello I, Gharbi S, Hendrich B. 2015. Constraint of gene expression by the chromatin remodelling protein CHD4 facilitates lineage specification. *Development* **142**: 2586–2597.
Ogas J, Kaufmann S, Henderson J, Somerville C. 1999. PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis. *Proc Natl Acad Sci U S A* **96**: 13839–13844.
Ohki I, Shimotake N, Fujita N, Jee JG, Ikegami T, Nakao M, Shirakawa M. 2001. Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* **105**: 487–497.
Önal P, Grün D, Adamidi C, Rybak A, Solana J, Mastrobuoni G, Wang Y, Rahn HP, Chen W, Kempa S, et al. 2012. Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. *EMBO J* **31**: 2755–2769.
Orii H, Ito H, Watanabe K. 2002. Anatomy of the Planarian Dugesia japonica I. The Muscular System Revealed by Antisera against Myosin Heavy Chains. *Zoolog Sci* **19**: 1123–1131.
Orii H, Sakurai T, Watanabe K. 2005. Distribution of the stem cells (neoblasts) in the planarian Dugesia japonica. *Dev Genes Evol* **215**: 143–157.
Orlando DA, Chen MW, Brown VE, Solanki S, Choi YJ, Olson ER, Fritz CC, Bradner JE, Guenther MG. 2014. Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep* **9**: 1163–1170.
Owlarn S, Bartscherer K. 2016. Go ahead, grow a head! A planarian's guide to anterior regeneration. *Regeneration* **3**: 139–155.
Palakodeti D, Smielewska M, Lu Y-C, Yeo GW, Graveley BR. 2008. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA* **14**: 1174–1186.
Pandey RR, Tokuzawa Y, Yang Z, Hayashi E, Ichisaka T, Kajita S, Asano Y, Kunieda T, Sachidanandam R, Chuma S, et al. 2013. Tudor domain containing 12 (TDRD12) is essential for secondary PIWI interacting RNA biogenesis in mice. *Proc Natl Acad Sci* **110**: 16492–16497.
Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JCH, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. 2011. The genetic landscape of the childhood cancer medulloblastoma. *Science (80-)* **331**: 435–439.
Pasini D, Bracken AP, Hansen JB, Capillo M, Helin K. 2007. The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol Cell Biol* **27**: 3769–79.
Pauler FM, Sloane MA, Huang R, Regha K, Koerner M V., Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP. 2009. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res* **19**: 221–233.
Pearson BJ, Alvarado AS. 2010. A planarian p53 homolog regulates proliferation and self-renewal in adult stem cell lineages. *Development* **137**: 213–221.
Pearson BJ, Eisenhoffer GT, Gurley KA, Rink JC, Miller DE, Alvarado AS. 2009. Formaldehyde-based whole-mount in situ hybridization method for planarians. *Dev Dyn* **238**: 443–450.

Pereira B, Le Borgne M, Chartier NT, Billaud M, Almeida R. 2013. MEX-3 proteins: recent insights on novel post-transcriptional regulators. *Trends Biochem Sci* **38**: 477–479.
Petersen CP, Reddien PW. 2009. A wound-induced Wnt expression program controls planarian regeneration polarity. *Proc Natl Acad Sci* **106**: 17061–17066.
Petersen CP, Reddien PW. 2008. Smed-betacatenin-1 is required for anteroposterior blastema polarity in planarian regeneration. *Science* **319**: 327–30.
Phalke S, Nickel O, Walluscheck D, Hortig F, Onorati MC, Reuter G. 2009. Retrotransposon silencing and telomere integrity in somatic cells of Drosophila depends on the cytosine-5 methyltransferase DNMT2. *Nat Genet* **41**: 696–702.
Piast M, Kustrzeba-Wójcicka I, Matusiewicz M, Banaś T. 2005. Molecular evolution of enolase. *Acta Biochim Pol* **52**: 507–13.
Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, Obermayer B, Theis FJ, Kocks C, Rajewsky N. 2018. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science (80-)* **360**: eaaq1723.
Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. 2018. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**: 858-871.e8.
Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, Carneiro MO, Carter SL, Cibulskis K, Erlich RL, et al. 2012. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**: 106–110.
Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov V V., et al. 2007. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science (80-)* **317**: 86–94.
Quillien A, Abdalla M, Yu J, Ou J, Zhu LJ, Lawson ND. 2017. Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq. *Cell Rep* **20**: 709–720.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
Raddatz G, Guzzardo PM, Olova N, Fantappie MR, Rampp M, Schaefer M, Reik W, Hannon GJ, Lyko F. 2013. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc Natl Acad Sci* **110**: 8627–8631.
Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc Regulates Transcriptional Pause Release. *Cell* **141**: 432–445.
Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–W165.
Rangan P, Malone CD, Navarro C, Newbold SP, Hayes PS, Sachidanandam R, Hannon GJ, Lehmann R. 2011. piRNA Production Requires Heterochromatin Formation in Drosophila. *Curr Biol* **21**: 1373–1379.
Rechavi O, Lev I. 2017. Principles of Transgenerational Small RNA Inheritance in Caenorhabditis elegans. *Curr Biol* **27**: R720–R730.
Reddien PW. 2013. Specialized progenitors and regeneration. *Development* **140**: 951–957.
Reddien PW, Bermange AL, Murfitt KJ, Jennings JR, Sánchez Alvarado A. 2005a. Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. *Dev Cell* **8**: 635–649.
Reddien PW, Oviedo NJ, Jennings JR, Jenkin JC, Sánchez Alvarado A. 2005b. SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* **310**: 1327–30.
Reuter H, März M, Vogg MC, Eccles D, Grífol-Boldú L, Wehner D, Owlarn S, Adell T, Weidinger G, Bartscherer K. 2015. β-Catenin-Dependent Control Of Positional Information Along The AP body axis in planarians involves a teashirt family member. *Cell Rep* **10**: 253–265.
Reynolds N, Latos P, Hynes-Allen A, Loos R, Leaford D, O'Shaughnessy A, Mosaku O, Signolet J, Brennecke P, Kalkan T, et al. 2012a. NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. *Cell Stem Cell* **10**: 583–594.
Reynolds N, Salmon-Divon M, Dvinge H, Hynes-Allen A, Balasooriya G, Leaford D, Behrens A, Bertone P, Hendrich B. 2012b. NuRD-mediated deacetylation of H3K27 facilitates

recruitment of Polycomb Repressive Complex 2 to direct gene repression. *EMBO J* **31**: 593–605.
Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res* 1299–1309.
Rink JC. 2013. Stem cell systems and regeneration in planaria. *Dev Genes Evol* **223**: 67–84.
Rink JC, Gurley KA, Elliott SA, Sánchez Alvarado A. 2009. Planarian Hh signaling regulates regeneration polarity and links Hh pathway evolution to cilia. *Science* **326**: 1406–10.
Rink JC, Vu HT-K, Alvarado AS. 2011. The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. *Development*.
Robb SMC, Alvarado AS. 2014. Histone Modifications and Regeneration in the Planarian Schmidtea mediterranea. In *Current Topics in Developmental Biology*, Vol. 108 of, pp. 71–93.
Robb SMC, Alvarado AS. 2002. Identification of immunological reagents for use in the study of freshwater planarians by means of whole-mount immunofluorescence and confocal microscopy. *genesis* **32**: 293–298.
Robb SMC, Gotting K, Ross E, Sánchez Alvarado A. 2015. SmedGD 2.0: The Schmidtea mediterranea genome database. *Genesis*.
Robb SMC, Ross E, Alvarado AS. 2007. SmedGD: the Schmidtea mediterranea genome database. *Nucleic Acids Res* **36**: D599–D606.
Robb SMC, Ross E, Sánchez Alvarado A. 2008. SmedGD: the Schmidtea mediterranea genome database. *Nucleic Acids Res* **36**: D599-606.
Roberts-Galbraith RH, Brubacher JL, Newmark PA. 2016. A functional genomics screen in planarians reveals regulators of whole-brain regeneration. *Elife* **5**.
Romero BT, Evans DJ, Aboobaker AA. 2012. FACS analysis of the planarian stem cell compartment as a tool to understand regenerative mechanisms. *Methods Mol Biol* **916**: 167–179.
Roquis D, Lepesant JMJ, Picard MAL, Freitag M, Parrinello H, Groth M, Emans R, Cosseau C, Grunau C. 2015. The epigenome of schistosoma mansoni provides insight about how cercariae poise transcription until infection. *PLoS Negl Trop Dis* **9**.
Rosado Fantappié M, Rodrigues Pereira Gimba E, Rumjanek FD. 2001. Lack of DNA methylation in Schistosoma mansoni. *Exp Parasitol* **98**: 162–166.
Rošić S, Amouroux R, Requena CE, Gomes A, Emperle M, Beltran T, Rane JK, Linnett S, Selkirk ME, Schiffer PH, et al. 2018. Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nat Genet* **50**: 452–459.
Ross KG, Omuro KC, Taylor MR, Munday RK, Hubert A, King RS, Zayas RM. 2015. Novel monoclonal antibodies to study tissue regeneration in planarians. *BMC Dev Biol* **15**.
Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33**: 1165–1172.
Rozanski A, Moon H, Brandl H, Martín-Durán JM, Grohme MA, Hüttner K, Bartscherer K, Henry I, Rink JC. 2019. PlanMine 3.0—improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic Acids Res* **47**: D812–D820.
Rozhkov N V., Hammell M, Hannon GJ. 2013. Multiple roles for Piwi in silencing Drosophila transposons. *Genes Dev* **27**: 400–412.
Ruiz-Trillo I, Riutort M, Timothy D, Littlewood J, Herniou EA, Baguñà J. 1999. Acoel flatworms: Earliest extant bilaterian metazoans, not members of platyhelminthes. *Science (80-)* **283**: 1919–1923.
Sachs M, Onodera C, Blaschke K, Ebata K, Song J, Ramalho-Santos M. 2013. Bivalent Chromatin Marks Developmental Regulatory Genes in the Mouse Embryonic Germline InVivo. *Cell Rep* **3**: 1777–1784.
Sagai T. 2005. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**: 797–803.
Sahu S, Dattani A, Aboobaker AA. 2017. Secrets from immortal worms: What can we learn about

biological ageing from the planarian model system? *Semin Cell Dev Biol* **70**: 108–121.
Salvetti A. 2005. DjPum, a homologue of Drosophila Pumilio, is essential to planarian stem cell maintenance. *Development* **132**: 1863–1874.
Salvetti A, Rossi L, Deri P, Batistoni R. 2000. An MCM2-related gene is expressed in proliferating cells of intact and regenerating planarians. *Dev Dyn* **218**: 603–14.
Sanchez Alvarado A, Newmark PA. 1999. Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc Natl Acad Sci* **96**: 5049–5054.
Schaefer M, Hagemann S, Hanna K, Lyko F. 2009. Azacytidine inhibits RNA methylation at DNMT2 target sites in human cancer cell lines. *Cancer Res* **69**: 8127–8132.
Schaefer M, Lyko F. 2010. Lack of evidence for DNA methylation of Invader4 retroelements in Drosophila and implications for Dnmt2-mediated epigenetic regulation. *Nat Genet* **42**: 920–921.
Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett SW, et al. 2015a. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res* **25**: 582–597.
Schoenfelder S, Sugar R, Dimond A, Javierre BM, Armstrong H, Mifsud B, Dimitrova E, Matheson L, Tavares-Cadete F, Furlan-Magaril M, et al. 2015b. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet*.
Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, van Lohuizen M, Tanay A, Cavalli G. 2009. Functional Anatomy of Polycomb and Trithorax Chromatin Landscapes in Drosophila Embryos ed. R. Kingston. *PLoS Biol* **7**: e1000013.
Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**: 212–216.
Schwaiger M, Schönauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, Schinko JB, Renfer E, Fredman D, Technau U. 2014. Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res* **24**: 639–650.
Scimone ML, Cote LE, Reddien PW. 2017. Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature* **551**: 623–628.
Scimone ML, Cote LE, Rogers T, Reddien PW. 2016. Two FGFRL-Wnt circuits organize the planarian anteroposterior axis. *Elife* **5**: e12845.
Scimone ML, Kravarik KM, Lapan SW, Reddien PW. 2014. Neoblast specialization in regeneration of the planarian schmidtea mediterranea. *Stem Cell Reports* **3**: 339–352.
Scimone ML, Meisel J, Reddien PW. 2010. The Mi-2-like Smed-CHD4 gene is required for stem cell differentiation in the planarian Schmidtea mediterranea. *Development* **137**: 1231–1241.
Sebé-Pedrós A, Ballaré C, Parra-Acero H, Chiva C, Tena JJ, Sabidó E, Gómez-Skarmeta JL, Di Croce L, Ruiz-Trillo I. 2016. The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. *Cell* **165**: 1224–1237.
Shannon M. 2003. Differential Expansion of Zinc-Finger Transcription Factor Loci in Homologous Human and Mouse Gene Clusters. *Genome Res* **13**: 1097–1110.
Shaver S, Casas-Mollano JA, Cerny RL, Cerutti H. 2010. Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. *Epigenetics* **5**: 301–312.
Shibata N, Kashima M, Ishiko T, Nishimura O, Rouhana L, Misaki K, Yonemura S, Saito K, Siomi H, Siomi MC, et al. 2016. Inheritance of a Nuclear PIWI from Pluripotent Stem Cells by Somatic Descendants Ensures Differentiation by Silencing Transposons in Planarian. *Dev Cell* **37**: 226–237.
Shibata N, Rouhana L, Agata K. 2010. Cellular and molecular dissection of pluripotent adult somatic stem cells in planarians. *Dev Growth Differ* **52**: 27–41.
Shimbo T, Du Y, Grimm SA, Dhasarathy A, Mav D, Shah RR, Shi H, Wade PA. 2013. MBD3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS Genet* **9**: e1004028.
Shoji M, Tanaka T, Hosokawa M, Reuter M, Stark A, Kato Y, Kondoh G, Okawa K, Chujo T, Suzuki T, et al. 2009. The TDRD9-MIWI2 Complex Is Essential for piRNA-Mediated Retrotransposon Silencing in the Mouse Male Germline. *Dev Cell* **17**: 775–787.
Shpiz S, Olovnikov I, Sergeeva A, Lavrov S, Abramov Y, Savitsky M, Kalmykova A. 2011.

Mechanism of the piRNA-mediated silencing of Drosophila telomeric retrotransposons. *Nucleic Acids Res* **39**: 8703–8711.
Sienski G, Batki J, Senti K-A, Dönertas D, Tirian L, Meixner K, Brennecke J. 2015. Silencio/CG9754 connects the Piwi-piRNA complex to the cellular heterochromatin machinery. *Genes Dev* **29**: 2258–71.
Sienski G, Dönertas D, Brennecke J. 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* **151**: 964–980.
Sijen T, Plasterk RHA. 2003. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature* **426**: 310–314.
Sikorski TW, Buratowski S. 2009. The basal initiation machinery: beyond the general transcription factors. *Curr Opin Cell Biol* **21**: 344–351.
Simon JA, Kingston RE. 2013. Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. *Mol Cell* **49**: 808–824.
Siomi MC, Mannen T, Siomi H. 2010. How does the Royal Family of Tudor rule the PIWI-interacting RNA pathway? *Genes Dev* **24**: 636–646.
Smith AG, Heath JK, Donaldson DD, Wong GG, Moreau J, Stahl M, Rogers D. 1988. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature* **336**: 688–690.
Smith S, Stillman B. 1989. Purification and characterization of CAF-I, a human cell factor required for chromatin assembly during DNA replication in vitro. *Cell* **58**: 15–25.
Solana J. 2013. Closing the circle of germline and stem cells: the Primordial Stem Cell hypothesis. *Evodevo* **4**: 2.
Solana J, Gamberi C, Mihaylova Y, Grosswendt S, Chen C, Lasko P, Rajewsky N, Aboobaker AA. 2013. The CCR4-NOT Complex Mediates Deadenylation and Degradation of Stem Cell mRNAs and Promotes Planarian Stem Cell Differentiation ed. P.A. Newmark. *PLoS Genet* **9**: e1004003.
Solana J, Irimia M, Ayoub S, Orejuela MR, Zywitza V, Jens M, Tapial J, Ray D, Morris Q, Hughes TR, et al. 2016. Conserved functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative splicing in planarians. *Elife* **5**.
Solana J, Kao D, Mihaylova Y, Jaber-Hijazi F, Malla S, Wilson R, Aboobaker A. 2012. Defining the molecular profile of planarian pluripotent stem cells using a combinatorial RNAseq, RNA interference and irradiation approach. *Genome Biol* **13**: R19.
Solana J, Lasko P, Romero R. 2009. Spoltud-1 is a chromatoid body component required for planarian long-term stem cell self-renewal. *Dev Biol* **328**: 410–421.
Spangrude GJ, Heimfeld S, Weissman IL. 1988. Purification and characterization of mouse hematopoietic stem cells. *Science (80-)* **241**: 58–62.
Spitz F, Gonzalez F, Duboule D. 2003. A Global Control Region Defines a Chromosomal Regulatory Landscape Containing the HoxD Cluster. *Cell* **113**: 405–417.
Steele RE, David CN, Technau U. 2011. A genomic view of 500 million years of cnidarian evolution. *Trends Genet* **27**: 7–13.
Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, et al. 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**: 888–903.
Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, Koseki H, Brockdorff N, Fisher AG, Pombo A. 2007. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat Cell Biol* **9**: 1428–1435.
Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**: 41–45.
Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow S, Iakovenko N, Hausdorf B, Petersen M, et al. 2014. Platyzoan Paraphyly Based on Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia. *Mol Biol Evol* **31**: 1833–1849.
Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–1976.
Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes

under a mutation model. *Proc Natl Acad Sci* **87**: 4692–4696.
Swapna LS, Molinaro AM, Lindsay-Mosher N, Pearson BJ, Parkinson J. 2018. Comparative transcriptomic analyses and single-cell RNA sequencing of the freshwater planarian Schmidtea mediterranea identify major cell types and pathway conservation. *Genome Biol* **19**.
Tadepally HD, Burger G, Aubry M. 2008. Evolution of C2H2-zinc finger genes and subfamilies in mammals: Species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol Biol* **8**: 176.
Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**: 930–5.
Takahashi K, Yamanaka S. 2006. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**: 663–676.
Tan TCJ, Rahman R, Jaber-Hijazi F, Felix DA, Chen C, Louis EJ, Aboobaker A. 2012. Telomere maintenance and telomerase activity are differentially regulated in asexual and sexual worms. *Proc Natl Acad Sci* **109**: 4209–4214.
Tanay A, O'Donnell AH, Damelin M, Bestor TH. 2007. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proc Natl Acad Sci U S A* **104**: 5521–5526.
Tejada-Romero B, Carter J-M, Mihaylova Y, Neumann B, Aboobaker AA. 2015. JNK signalling is necessary for a Wnt- and stem cell-dependent regeneration programme. *Development* **142**: 2413–2424.
Tharp ME, Bortvin A. 2016. DjPiwiB: A Rich Nuclear Inheritance for Descendants of Planarian Stem Cells. *Dev Cell* **37**: 204–206.
Theunissen O, Rudt F, Guddat U, Mentzel H, Pieler T. 1992. RNA and DNA binding zinc fingers in Xenopus TFIIIA. *Cell* **71**: 679–690.
Thi-Kim Vu H, Rink JC, McKinney SA, McClain M, Lakshmanaperumal N, Alexander R, Sánchez Alvarado A. 2015. Stem cells and fluid flow drive cyst formation in an invertebrate excretory organ. *Elife* **4**.
Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.
Till JE, McCulloch EA. 1961. A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. *Radiat Res* **14**: 213.
Treisman JE, Follette PJ, O'Farrell PH, Rubin GM. 1995. Cell proliferation and DNA replication defects in a Drosophila MCM2 mutant. *Genes Dev* **9**: 1709–1715.
Tu KC, Pearson BJ, Sánchez Alvarado A. 2012. TORC1 is required to balance cell proliferation and cell death in planarians. *Dev Biol* **365**: 458–469.
Tuorto F, Herbst F, Alerasool N, Bender S, Popp O, Federico G, Reitter S, Liebers R, Stoecklin G, Gröne H-J, et al. 2015. The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. *Embo J* **34**: 2350–62.
Tweedie S, Charlton J, Clark V, Bird A. 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol* **17**: 1469–1475.
Tweedie S, Ng HH, Barlow AL, Turner BM, Hendrich B, Bird A. 1999. Vestiges of a DNA methylation system in Drosophila melanogaster? *Nat Genet* **23**: 389–390.
Unhavaithaya Y, Shin TH, Miliaras N, Lee J, Oyama T, Mello CC. 2002. MEP-1 and a homolog of the NURD complex component Mi-2 act together to maintain germline-soma distinctions in C. elegans. *Cell* **111**: 991–1002.
van Wolfswinkel JC. 2014. Piwi and potency: PIWI proteins in animal stem cells and regeneration. *Integr Comp Biol* **54**: 700–713.
Van Wolfswinkel JC, Wagner DE, Reddien PW. 2014. Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell Stem Cell* **15**: 326–339.
Vásquez-Doorman C, Petersen CP. 2016. The NuRD complex component p66 suppresses photoreceptor neuron regeneration in planarians. *Regeneration* **3**: 168–178.
Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J, Schier AF. 2010. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464**: 922–926.

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* **7**.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854–858.
Voigt P, LeRoy G, Drury WJ, Zee BM, Son J, Beck DB, Young NL, Garcia BA, Reinberg D. 2012. Asymmetrically modified nucleosomes. *Cell* **151**: 181–193.
Voigt P, Tee WW, Reinberg D. 2013. A double take on bivalent promoters. *Genes Dev* **27**: 1318–1338.
von Zelewsky T, Palladino F, Brunschwig K, Tobler H, Hajnal A, Müller F. 2000. The C. elegans Mi-2 chromatin-remodelling proteins function in vulval cell fate determination. *Development* **127**: 5277–5284.
Vos JC, Plasterk RH. 1994. Tc1 transposase of Caenorhabditis elegans is an endonuclease with a bipartite DNA binding domain. *EMBO J* **13**: 6125–32.
Voss TC, Schiltz RL, Sung M-H, Yen PM, Stamatoyannopoulos JA, Biddie SC, Johnson TA, Miranda TB, John S, Hager GL. 2011. Dynamic Exchange at Regulatory Elements during Chromatin Remodeling Underlies Assisted Loading Mechanism. *Cell* **146**: 544–554.
Wagner DE, Ho JJ, Reddien PW. 2012. Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. *Cell Stem Cell* **10**: 299–311.
Wagner DE, Wang IE, Reddien PW. 2011. Clonogenic Neoblasts Are Pluripotent Adult Stem Cells That Underlie Planarian Regeneration. *Science (80-)* **332**: 811–816.
Wagner EJ, Carpenter PB. 2012. Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol* **13**: 115–126.
Wahle E, Winkler GS. 2013. RNA decay machines: Deadenylation by the Ccr4–Not and Pan2–Pan3 complexes. *Biochim Biophys Acta - Gene Regul Mech* **1829**: 561–570.
Wang IE, Lapan SW, Scimone ML, Clandinin TR, Reddien PW. 2016. Hedgehog signaling regulates gene expression in planarian glia. *Elife* **5**.
Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci* **104**: 18613–18618.
Wang Y, Zhang H, Chen Y, Sun Y, Yang F, Yu W, Liang J, Sun L, Yang X, Shi L, et al. 2009. LSD1 Is a Subunit of the NuRD Complex and Targets the Metastasis Programs in Breast Cancer. *Cell* **138**: 660–672.
Wasik K, Gurtowski J, Zhou X, Ramos OM, Delás MJ, Battistoni G, El Demerdash O, Falciatori I, Vizoso DB, Smith AD, et al. 2015. Genome and transcriptome of the regeneration-competent flatworm, Macrostomum lignano. *Proc Natl Acad Sci U S A* **112**: 12462–7.
Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H. 2006. Identification and characterization of two novel classes of small RNAs in the mouse germline: Retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev* **20**: 1732–1743.
Webster MW, Stowell JA, Passmore LA. 2019. RNA-binding proteins distinguish between similar sequence motifs to promote targeted deadenylation by Ccr4-Not. *Elife* **8**: 1–56.
Wegner M. 2010. All purpose Sox: The many roles of Sox proteins in gene expression. *Int J Biochem Cell Biol* **42**: 381–390.
Wegner M, Riethmacher D. 2001. Chronicles of a switch hunt: gcm genes in development. *Trends Genet* **17**: 286–290.
Weick E-M, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* **141**: 3458–3471.
Weiner A, Lara-Astiaso D, Krupalnik V, Gafni O, David E, Winter DR, Hanna JH, Amit I. 2016. Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. *Nat Biotechnol* **34**: 953–961.
Wenemoser D, Reddien PW. 2010. Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Dev Biol* **344**: 979–991.
Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA. 2012. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation.

Nature **482**: 221–225.
Williams RL, Hilton DJ, Pease S, Willson T a, Stewart CL, Gearing DP, Wagner EF, Metcalf D, Nicola N a, Gough NM. 1988. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature* **336**: 684–687.
Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD - Taxonomically broad transcription factor predictions: New content and functionality. *Nucleic Acids Res* **36**.
Witchley JN, Mayer M, Wagner DE, Owen JH, Reddien PW. 2013. Muscle Cells Provide Instructions for Planarian Regeneration. *Cell Rep* **4**: 633–641.
Wittlieb J, Khalturin K, Lohmann JU, Anton-Erxleben F, Bosch TCG. 2006. Transgenic Hydra allow in vivo tracking of individual stem cells during morphogenesis. *Proc Natl Acad Sci* **103**: 6208–6211.
Wu M, Wang PF, Lee JS, Martin-Brown S, Florens L, Washburn M, Shilatifard A. 2008. Molecular Regulation of H3K4 Trimethylation by Wdr82, a Component of Human Set1/COMPASS. *Mol Cell Biol* **28**: 7337–7344.
Wu SF, Zhang H, Cairns BR. 2011. Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. *Genome Res* **21**: 578–589.
Wudarski J, Simanov D, Ustyantsev K, de Mulder K, Grelling M, Grudniewska M, Beltman F, Glazenburg L, Demircan T, Wunderer J, et al. 2017. Efficient transgenesis and annotated genome sequence of the regenerative flatworm model Macrostomum lignano. *Nat Commun* **8**: 2120.
Wurtzel O, Cote LE, Poirier A, Satija R, Regev A, Reddien PW. 2015. A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians. *Dev Cell* **35**: 632–645.
Wurtzel O, Oderberg IM, Reddien PW. 2017. Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell* **40**: 491-504.e5.
Yagi M, Kishigami S, Tanaka A, Semi K, Mizutani E, Wakayama S, Wakayama T, Yamamoto T, Yamada Y. 2017a. Derivation of ground-state female ES cells maintaining gamete-derived DNA methylation. *Nature* **548**: 224–227.
Yagi M, Yamanaka S, Yamada Y. 2017b. Epigenetic foundations of pluripotent stem cells that recapitulate in vivo pluripotency. *Lab Investig* **97**: 1133–1141.
Yang C, Stiller JW. 2014. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. *Proc Natl Acad Sci* **111**: 5920–5925.
Yazawa S, Umesono Y, Hayashi T, Tarui H, Agata K. 2009. Planarian Hedgehog/Patched establishes anterior-posterior polarity by regulating Wnt signaling. *Proc Natl Acad Sci* **106**: 22329–22334.
Yildirim O, Li R, Hung JH, Chen PB, Dong X, Ee LS, Weng Z, Rando OJ, Fazzio TG. 2011. Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**: 1498–1510.
Ying Q-L, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A. 2008. The ground state of embryonic stem cell self-renewal. *Nature* **453**: 519–523.
Yu Y, Gu J, Jin Y, Luo Y, Preall JB, Ma J, Czech B, Hannon GJ. 2015. Panoramix enforces piRNA-dependent cotranscriptional silencing. *Science* **350**: 339–42.
Yuan Y-W, Wessler SR. 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci* **108**: 7884–7889.
Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**: 2227–2241.
Zauri M, Berridge G, Thézénas M-L, Pugh KM, Goldin R, Kessler BM, Kriaucionis S. 2015. CDA directs metabolism of epigenetic nucleosides revealing a therapeutic window in cancer. *Nature* **524**: 114–118.
Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science (80-)* **328**: 916–919.
Zemach A, Zilberman D. 2010. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Curr Biol* **20**: R780-5.
Zeng A, Li H, Guo L, Gao X, McKinney S, Wang Y, Yu Z, Park J, Semerad C, Ross E, et al. 2018. Prospectively Isolated Tetraspanin + Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* **173**: 1593-1608.e20.

Zhang X, Novera W, Zhang Y, Deng L-W. 2017. MLL5 (KMT2E): structure, function, and clinical relevance. *Cell Mol Life Sci* **74**: 2333–2344.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.
Zhang Y, Ng HH, Erdjument-Bromage H, Tempst P, Bird A, Reinberg D. 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev* **13**: 1924–1935.
Zhou L, Mitra R, Atkinson PW, Hickman AB, Dyda F, Craig NL. 2004. Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* **432**: 995–1001.
Zhou W, Zhu P, Wang J, Pascual G, Ohgi KA, Lozach J, Glass CK, Rosenfeld MG. 2008. Histone H2A Monoubiquitination Represses Transcription by Inhibiting RNA Polymerase II Transcriptional Elongation. *Mol Cell* **29**: 69–80.
Zhu SJ, Hallows SE, Currie KW, Xu C, Pearson BJ. 2015. A mex3 homolog is required for differentiation during planarian stem cell lineage development. *Elife* **4**: 1–23.
Zhu SJ, Pearson BJ. 2013. The Retinoblastoma pathway regulates stem cell proliferation in freshwater planarians. *Dev Biol* **373**: 442–452.
Zou X, Ma W, Solov'Yov IA, Chipot C, Schulten K. 2012. Recognition of methylated DNA through methyl-CpG binding domain proteins. *Nucleic Acids Res* **40**: 2747–2758.
Zuryn S, Ahier A, Portoso M, White ER, Morin M-C, Margueron R, Jarriault S. 2014. Sequential histone-modifying activities determine the robustness of transdifferentiation. *Science (80-)* **345**: 826–829.