Identification of enhancer-like elements defines regulatory networks active in planarian adult stem cells

Jakke Neiro[1*], Divya Sridhar[1], Anish Dattani[2] and Aziz Aboobaker[1*]

1. Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom
2. Living Systems Institute, University of Exeter EX4 4QD, UK
* Corresponding authors

Abstract

Planarians have become an established model system to study regeneration and stem cells, but the regulatory elements in the genome remain almost entirely undescribed. Here, by integrating epigenetic and expression data we use multiple sources of evidence to identify enhancer elements active in the adult stem cell populations that drive regeneration. We have used ChIP-seq data to identify regions with histone modifications consistent with enhancer identity and activity, and ATAC-seq data to identify accessible chromatin. Overlapping these signals allowed for the identification of a set of high confidence candidate enhancers predicted to be active in planarian adult stem cells. These enhancers are enriched for conserved transcription factor (TF) binding sites for TFs and TF families expressed in planarian adult stem cells. Foot-printing analyses provided further evidence that these potential TF binding sites are occupied in adult stem cells. We integrated these analyses to build testable hypotheses for the regulatory function of transcription factors in stem cells, both with respect to how pluripotency might be regulated, and to how lineage differentiation programs are controlled. Our work identifies active enhancers regulating adult stem cells and regenerative mechanisms.

1. Introduction

The molecular and evolutionary mechanisms of regeneration remain underexplained compared to animal development. This can be attributed to the greater technical difficulty historically associated with studying the molecular mechanisms of adult biology compared with development. However, an ever-broadening repertoire of model organisms for regeneration, progress in understanding the variety of cellular and molecular mechanisms used across taxa, and advances in experimental tools are serving to close this gap. One area of slower progress has been our knowledge of the regulatory logic of regenerative mechanisms, with relatively few studies in regenerative models providing precise insight into the epigenetic regulation of regeneration. While transcription factors have been assigned functions, and in some cases a list of likely targets, the regulatory logic is defined in only a few cases (Pascual-Carreras et al. 2020, Wang et al. 2020). This is particularly true in the highly regenerative planarian model system, where only a few studies credibly address epigenetic and regulatory mechanisms (Duncan et al. 2015, Dattani et al. 2018, Mihaylova et al. 2018, Pascual-Carreras et al. 2020).

Enhancers are distal-acting elements that regulate transcription initiation when transcription factors (TFs) attach to transcription factor binding sites, containing conserved sequences. In a regenerative context, active enhancers have been identified at the tissue-level in zebrafish, the African killifish *Nothobranchius furzeri,* and the fruit fly *Drosophila melanogaster*, and at the whole-body level in the acoel worm *Hofstenia miamia* and the cnidarian *Hydra vulgaris* (Harris et al. 2016, Kang et al. 2016, Goldman et al. 2017, Gehrke et al. 2019, Murad et al. 2019, Yang et al. 2019, Wang et al. 2020). Despite their importance as a model for understanding regeneration, there are only a few putative enhancers identified or implicated in planarians (Pascual-Carreras et al. 2020), partly due to the lack of transgenic approaches allowing for direct functional testing of potential regulatory elements. Hence, other analytical methods must be used to discover enhancer regions *de novo* across the genome. A set of putative planarian TFs have been identified based on available transcriptomic data, but genome-wide TF interaction with potential enhancers has not been elucidated (Swapna et al. 2018; Pascual-Carreras et al. 2020). Ultimately, TF interactions with enhancers can be used to construct gene regulatory networks (GRNs) (Duren et al. 2017, Lowe et al. 2019, Miraldi et al. 2019, Duren et al. 2020, Janssens et al. 2022).

In planarians, regeneration is based on a population of somatic pluripotent stem cells called neoblasts that can differentiate into all cell types of the adult body plan and are the only cycling cells in the adult worm (Aboobaker 2011, Rink 2013, Zhu and Pearson 2016). Nonetheless, how neoblasts dynamically retain pluripotency, specify fate, and differentiate remains unknown from a regulatory network perspective, and a number of different models have been hypothesized (Adler and Sánchez Alvarado 2015, Molinaro and Pearson 2018, Raz et al. 2021). Fluorescence activated cell sorting (FACS) separates cells into three compartments: S/G2/M cell-cycle stage

neoblasts (X1), G1 stage neoblasts and post-mitotic progeny (X2), and differentiated cells (Xins) (Hayashi et al. 2006, Labbé et al. 2012, Önal et al. 2012, Romero et al. 2012, Solana et al. 2012). The X1 cells have been used to catalog genes expressed in neoblasts, and these studies have revealed that neoblasts are heterogeneous and can be subdivided into multiple classes based on their gene expression profiles (van Wolfswinkel et al 2014; Fincher et al., 2018; Plass et al., 2018). Some classes of fate-specified neoblasts are termed specialized neoblasts, which include precursors to eyes, protonephridia, epidermis, intestine, pharynx, neurons, and muscle (van Wolfswinkel et al 2014; Raz et al. 2021). Fate-specific transcription factors (FSTFs) are expressed in S/G2/M phases and direct fate specification into different cell types (Raz et al. 2021). Evidence from current experimental data supports a model where specialized neoblasts can divide asymmetrically giving rise to one fate-specified, post-mitotic daughter and a naïve neoblast that may specify a different fate (Raz et al 2021). Genome-wide identification of active enhancers in the X1 compartment would shed light on the GRNs regulating the dynamic behavior of neoblasts in planarian regeneration (Labbé et al. 2012, Önal et al. 2012, Solana et al. 2012).

Identifying enhancers is challenging and especially so in non-model organisms, but various genome-wide high-throughput sequencing techniques have revealed signatures indicative of enhancers (Shlyueva et al. 2014, Tomoyasu and Halfon 2020). Chromatin accessibility has proven to be a universal attribute of active enhancers and other regulatory regions in eukaryotes (Thomas et al. 2011, West et al. 2014, Zhu et al. 2015, Daugherty et al. 2017, Klemm et al. 2019). Assay for transposase-accessible chromatin using sequencing (ATAC-seq) has become the standard method of mapping open chromatin regions in various taxa (Buenrostro et al. 2013, Li et al. 2019). However, chromatin accessibility is not unique and specific to enhancers, and thus other complementary methods are used to discriminate enhancers from other open chromatin regions.

Enhancers are known to be flanked by nucleosomes with specific histone modifications (Shlyueva et al. 2014). In both mammals and *D. melanogaster*, acetylated lysine 27 of histone H3 (H3K27ac) is known to mark active enhancers together with mono-methylation of lysine 4 on histone H3 (H3K4me1), whereas H3K4me1 alone marks predetermined or poised enhancers (Heintzman et al. 2007, Creyghton et al. 2010, Ernst et al. 2011, Rada-Iglesias et al. 2011, Bonn et al. 2012, Arnold et al. 2013, Calo and Wysocka 2013). However, poised enhancers are not always activated, most enhancers are activated without a prior poised state, and some poised enhancers are later actively repressed, indicating that the poised enhancer state is not necessarily indicative of a pre-activation state (Bonn et al. 2012, Rada-Iglesias et al. 2012, Koenecke et al. 2017). Nonetheless, the bivalent epigenetic signature of H3K27ac and H3K4me1 seems to be a conserved indicator of active enhancers in metazoans and has been successfully used for enhancer detection in non-model organisms (Gaiti et al. 2017, Jänes et al. 2018). In planarian neoblasts, the same histone modifications mark the active, suppressed, and bivalent state of promoters as in vertebrates

(Dattani et al. 2018). Taken together, these data suggest that epigenetic marks may have a conserved association with regulatory elements across bilaterians (Schwaiger et al. 2014, Sebé-Pedrós et al. 2016, Gaiti et al. 2017, Dattani et al. 2018).

Here, we significantly improve upon the annotation of the planarian genome (Grohme et al. 2018), take a computational approach to identify all TFs in this annotation, identify putative enhancers in the planarian genome supported by multiple lines of evidence, and then construct hypothetical GRNs active in neoblasts. We find that multiple enhancers have evidence of FSTF mediated regulation, supporting the view that fate specification occurs in the S/G2/M phases. The FSTFs of different cell types seem to cross-regulate each other, revealing a dynamic GRN governing neoblast fate specification. We identified enhancers linked to a number of unstudied transcription factors implicating them in potentially central roles in neoblast GRNs. Enhancers linked to well-known planarian positional genes suggest regulatory mechanisms for some of the known links between these genes implicated by phenotypic studies. Overall, this work provides a foundation for future work on the regulatory logic of planarian stem cell biology.

2. Results

2.1 Annotation with a full range of transcriptome samples identifies more than 3000 new protein coding genes in the *Schmidtea mediterranea* genome

We refined and extended the current annotation of the *Schmidtea mediterranea* genome (SMESG.1 genome and SMESG high confidence annotation at Planmine (Brandl et al. 2016)). We performed a genome-directed annotation based on the genome sequence and 183 independent RNA-seq datasets, including data both from whole worms and cell compartments (Fig. 1 A and B) (This new annotation has also been used in García-Castro et al. (2021)). By including a diverse and large set of RNA-seq data, we sought to characterise transcripts undetected in individual studies and annotation attempts (Hoff and Stanke 2013, 2019). Furthermore, we calculated proportional expression values for each cell population defined by FACS using approaches established previously (Fig. 1 A) (Dattani et al. 2018).

In total, our expression-driven annotation process identified 91,068 transcripts at 28,097 genomic loci (Fig. 1 B). The annotation process validated all gene models in the SMESG high confidence annotation, as all loci, transcripts and exons were also found in our new annotation. In total, 50,213 transcripts were identified as putative new isoforms of previously identified loci (Fig. 1 C). Furthermore, 7,412 new loci with 10,636 transcripts were found (Fig. 1 B and C). The protein coding potential of these new transcripts was assessed by defining putative open reading frames (ORFs) and scanning for protein structures (Fig. 1 B). In total, 3,121 new loci with 4,752 transcripts were predicted to be coding, while 4,291 loci with 5,884 transcripts were predicted to be non-coding (Fig. 1 B and C).

The newly described isoforms across the genome were slightly shorter than known transcripts (median length 1656 bp vs. 1618 bp), while the new coding and non-coding transcripts were much shorter (median length 583 vs. 388 bp) (Fig. 1 D). However, the mean Transcripts Per Kilobase Million (TPM) value measured across the RNA-seq samples for isoforms and new coding and non-coding transcripts did not differ much from the levels of known transcripts, and the mean TPM values for new coding and non-coding transcripts were slightly higher (Fig. S1 A). This suggests that the main advantage of our approach was to discover shorter transcripts (and encoded proteins) not found by previous annotation approaches (Fig. 1 D and S1 A). We sorted all annotated genes to FACS enrichment groups (Fig. 1 H), using previously described methods (Dattani et al. 2018). The number of genes with enriched expression was highest in the Xins compartment and lowest in X1 cells (Fig. 1 E). As expected, the shared enrichment between X1 and Xins is less common than between X1 and X2 and between X2 and Xins cells, congruent with the fact that X1 cells are all neoblasts, X2 is an amalgam of G1 neoblasts and differentiating post-mitotic progeny, and Xins cells are their collective differentiation products (Fig. 1 E). By inspecting the distribution of proportional expression values, we also see that the distribution is shifted towards the Xins compartment (Fig. 1 F and H).

The majority of the new coding loci were unique to *S. mediterranea* (70 %), but a fair share had a homolog in the *Dugesia japonica* genome (30 %) (E-value cutoff 10^{-10}, Fig. 1 G) (An et al. 2018). Only 27 non-coding loci (0.6 %) had a potential homolog in the human genome, but 666 loci (16 %) had a homolog in the *D. japonica* genome (Fig. 1 G). The gene ontology (GO) analysis of new coding loci revealed that the new genes were enriched for a function in RNA biosynthesis, metabolic processes, dephosphorylation, deacetylation, transmembrane transport, and mitochondrial functions (Fig. S1 C, D and E). These data suggest that many of our newly annotated genes are conserved amongst the group of planarians used intensively for regeneration and stem cell research.

Some genes in the new annotation displayed high levels of alternative splicing, including homologs of lectin, ankyrins, and dystonin (Fig. 1 H, S1 B, Table 1). In mammals, ankyrin-3 (ankyrin-G) is a structural protein localised to the axon initial segment (AIS) and the nodes of Ranvier, and alternative splicing is known to underlie its functional diversity (Fig. 1 H and S1 F) (Hopitzan et al., 2005; Lopez et al., 2017; Nelson and Jenkins, 2017). A planarian homolog of ankyrin-3 had 111 isoforms, including isoforms with one long exon, supporting the finding that giant ankyrin-based cytoskeleton of the AIS may have been present in the last common ancestor of bilaterians (Fig. 1 H and S1 F) (Jegla et al., 2016).

Taken together, these initial example analyses of our new annotation, particularly the discovery of many hundreds of new loci and thousands of putative alternative isoforms suggest it will have an important utility for the research community studying all aspects of planarian biology.

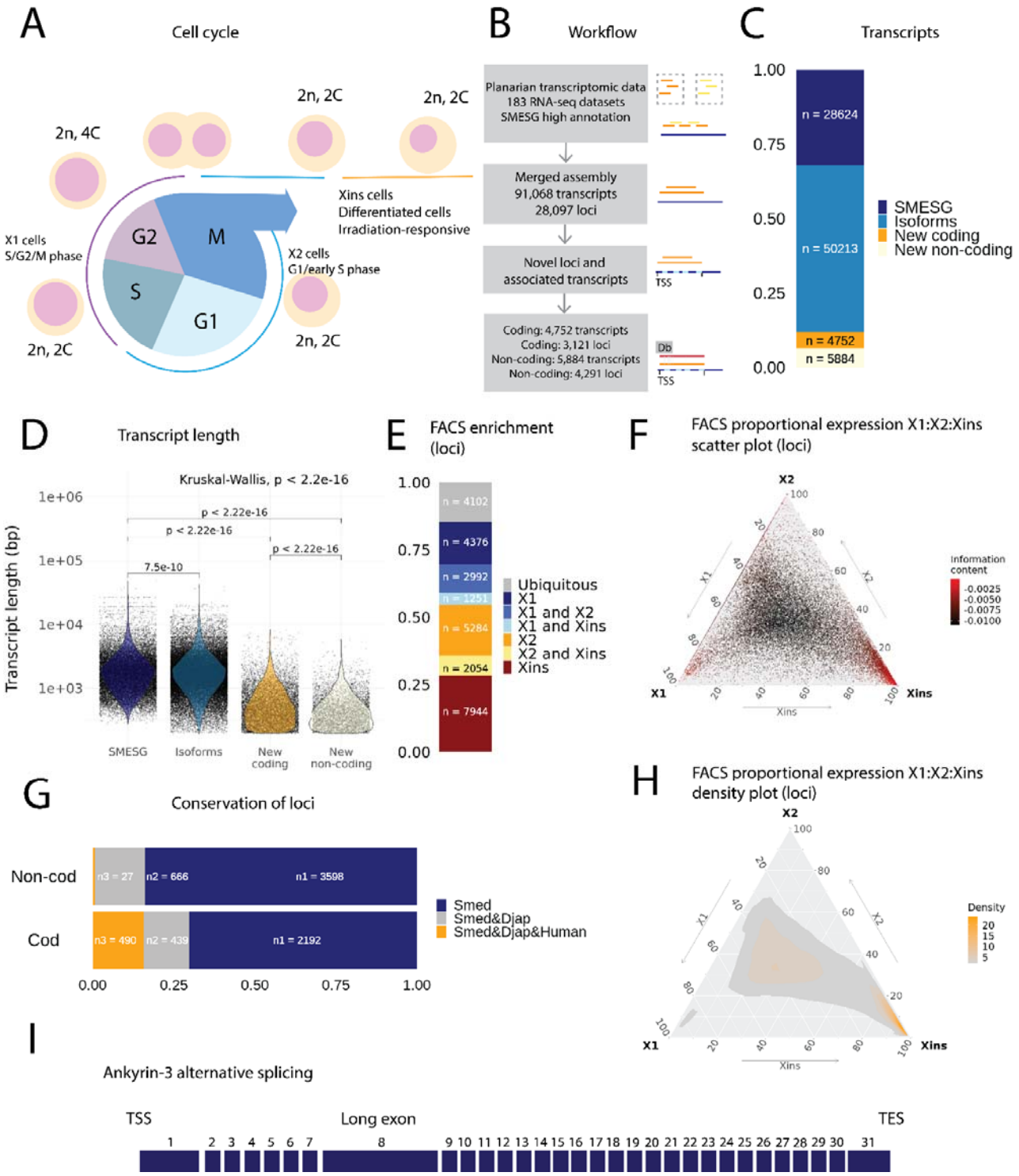
A
Cell cycle
2n, 4C
2n, 2C
2n, 2C
2n, 2C
2n, 2C
X1 cells
S/G2/M phase
G2
M
S
G1
Xins cells
Differentiated cells
Irradiation-responsive
X2 cells
G1/early S phase
B
Workflow
Planarian transcriptomic data
183 RNA-seq datasets
SMESG high annotation
Merged assembly
91,068 transcripts
28,097 loci
Novel loci and
associated transcripts
Coding: 4,752 transcripts
Coding: 3,121 loci
Non-coding: 5,884 transcripts
Non-coding: 4,291 loci
TSS
Db
TSS
C
Transcripts
1.00
0.75
0.50
0.25
0.00
n = 28624
n = 50213
n = 4752
n = 5884
SMESG
Isoforms
New coding
New non-coding
D
Transcript length
Kruskal-Wallis, p < 2.2e-16
Transcript length (bp)
1e+06
1e+05
1e+04
1e+03
p < 2.22e-16
p < 2.22e-16
p < 2.22e-16
7.5e-10
SMESG
Isoforms
New coding
New non-coding
E
FACS enrichment
(loci)
1.00
0.75
0.50
0.25
0.00
n = 4102
n = 4376
n = 2992
n = 1251
n = 5284
n = 2054
n = 7944
Ubiquitous
X1
X1 and X2
X1 and Xins
X2
X2 and Xins
Xins
F
FACS proportional expression X1:X2:Xins
scatter plot (loci)
X2
X1
Xins
Information
content
-0.0025
-0.0050
-0.0075
-0.0100
G
Conservation of loci
Non-cod
Cod
n3 = 27
n2 = 666
n1 = 3598
n3 = 490
n2 = 439
n1 = 2192
Smed
Smed&Djap
Smed&Djap&Human
0.00
0.25
0.50
0.75
1.00
H
FACS proportional expression X1:X2:Xins
density plot (loci)
X2
X1
Xins
Density
20
15
10
5
I
Ankyrin-3 alternative splicing
TSS
Long exon
TES
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Fig. 1. A. Diagram of the planarian neoblast cell cycle illustrating how the FACS cell compartments X1, X2, and Xins relate to different cell-cycle phases. **B.** Overview of methodology for annotating the *S. mediterranea* genome and identifying novel coding and non-coding transcripts. In total 183 RNA-seq data sets were aligned to the SMESG.1 genome (Planmine). Novel transcripts were assembled by using the SMESG-high confidence annotation as a reference (Planmine). **C.** The proportion of transcript identities in the new annotation. “SMESG” are transcripts present in the known SMESG-high confidence annotation, “Isoforms” are transcripts that are isoforms or splicing variants of known transcripts in the SMESG-high annotation, and “New coding” and “New non-coding” are previously unknown transcripts deemed to have coding and non-coding potential, respectively. **D.** The transcript lengths of known transcripts (SMESG), isoforms, new coding transcripts, and new non-coding transcripts. **E.** FACS categorization of 28,003 total loci into enrichment groups (Dattani et al., 2018). The enrichment groups are X1 (X1 proportional expression ≥ 50%), X2 (X2 proportional expression ≥ 50%), Xins (Xins proportional expression ≥ 50%), X1 and X2 (X1 + X2 proportional expression ≥ 75%, neither enriched in X1 nor X2), X1 and Xins (X1 + Xins proportional expression ≥ 75%, neither enriched in X1 nor Xins), X2 and Xins (X2 + Xins proportional expression ≥ 75%, neither enriched in X2 nor Xins), and ubiquitous (loci not categorized into the enrichment groups above and with roughly equal proportion in all three groups). **F.** Ternary plot of proportional expression values in the X1, X2, and Xins cell compartments. Each dot represents one single loci. The information content metric represents the enrichment of a locus to one of the three compartments. **G.** Conservation of new coding and non-coding loci to loci in *Dugesia japonica* and humans. “Smed” (n1) are loci exclusive to *S. mediterranea*, “Smed&Djap” (n2) are loci with a homolog in *D. japonica* but not in humans, and “Smed&Djap&Human” (n3) are loci with homologs both in *D. japonica* and humans. **H.** Ternary plot of the density distribution of proportional expression values in the X1, X2, and Xins cell compartments. **I.** Schematic of the alternative splicing of ankyrin-3.

Table 1.

Genes with more than 100 transcript isoforms.

Number of isoforms	Gene	Description
571	TMEM25	Transmembrane protein 25
291	CLEC18B	C-type lectin domain family 18 member B
206	ANK2	Ankyrin 2
143	GABPB2	GA binding protein transcription factor subunit
138	PACRG	Parkin coregulated
112	ANK3	Ankyrin 3
106	DST	Dystonin
104	ANK3	Ankyrin 3

2.2 Comprehensive annotation of planarian transcription factors highlights a diversity of unknown zinc fingers

We screened for TFs in the new annotation using the same approach as Swapna et al. (2018) (Fig. 2 A). We validated the TF potential through a systematic protein domain annotation, assessed for homology to known TFs, and manually reviewed the list of TFs to assign them to planarian TFs present in the literature and databases (Fig. 2 A). Altogether, we found putative 551 TFs in *S. mediterranea*, of which we found 248 to be described in the planarian literature in some way. The naming of planarian TFs in the literature was mostly consistent, but some inconsistencies were also found (Supplementary File 1). We classified the TFs into four structural categories: basic domains, zinc domains, helix-turn-helix domains, and other domains (Stegmaier et al. 2004). Most basic domains had been described by Cowles et al. (2013), but significantly we could identify new homologs of *Atf*, *Batf*, *Creb*, *Htf*, *Mad*, *Matf*, *MyoD*, *Npas*, and *Pdp* family members, each with broad established roles in metazoan biology (Fig. 2 B).

In contrast, we identified multiple uncharacterized zinc finger domain TFs (ZNFs), many of which have not received much consideration yet in planarian regeneration research. While some of these unstudied ZNFs could be assigned to well-known ZNF families such as GATA, KLF, EGR, and PRDM, many could not be assigned to well-described families. The nomenclature of ZNFs was based on the naming of human proteins to which they have the highest identity, and hence many appear in the ZNF and ZNP categories (Fig. 2 C). Interestingly, we also find ZNFs related to SCAN-domain containing zinc fingers (ZSCAN) and Pogo transposable elements with KRAB domain (POGK) previously only described in vertebrates (Emerson and Thomas 2011, Gao et al. 2020). However, these planarian ZNFs do not contain the SCAN or KRAB domains, and the similarity arises mainly from the DNA-binding domains, respectively. While members of the pogo transposable element superfamily are found throughout the metazoans, the KRAB sub-family is specific to vertebrates (Gao et al. 2020). For this reason, these ZNFs have been provisionally named Zscan-like and POGK-like (Fig. 2 C)

For helix-turn-helix domains, new homologs were found to several described TF families, including *Nkx* and *Six*, and a few transcription factors belonging to families not previously described in planarians were also discovered, such as *Lmx* and *Shox* (Fig. 2 D). Some TFs were also newly annotated for other domain families (Fig. 2 E).

We proceeded to allocate proportional expression values to the TFs with respect to the X1, X2, and Xins cell compartments (Dattani et al. 2018). The distribution of the proportional expression values was even, although a slight skew towards the X1 compartment was evident (Fig. 2 F and G). Most TFs assigned to the X1 and Xins compartments and the least to the X2 compartment (Fig. 2 F and G). Very few TFs had expression in X1s and Xins to the exclusion of X2 expression

as would be expected by overall lineage relationships. There was no compartment-specific enrichment for different TF domains (χ^2 test, $p = 0.2$).

We then moved to assigning predicted target binding motifs to the annotated set of planarian transcription factors. Most studies in non-model organisms have tended to use motifs directly from only a single model organism or solely *de novo* motif enrichment without reference to the TFs actually present in the studied organism (Gaiti et al. 2017, Gehrke et al. 2019, Murad et al. 2019). However, we used the same approach as Siebert et al. (2019) and searched the JASPAR database for TFs with the highest similarity to predict motifs for planarian TFs (Fig. 2 A). In total, we found 166 motifs that were assigned to 263 TFs with a normal distribution of motif information value (Fig. 2 H, Supplementary file 1). The most informative motifs were found for *zfp-7*, *pbx*, *prep*, *gli-3*, and *rfx7*, while the least informative motifs were for *hesl-2*, *pou2/3*, *irx*, *tead1* and *lmx1* (Fig. 2 I).

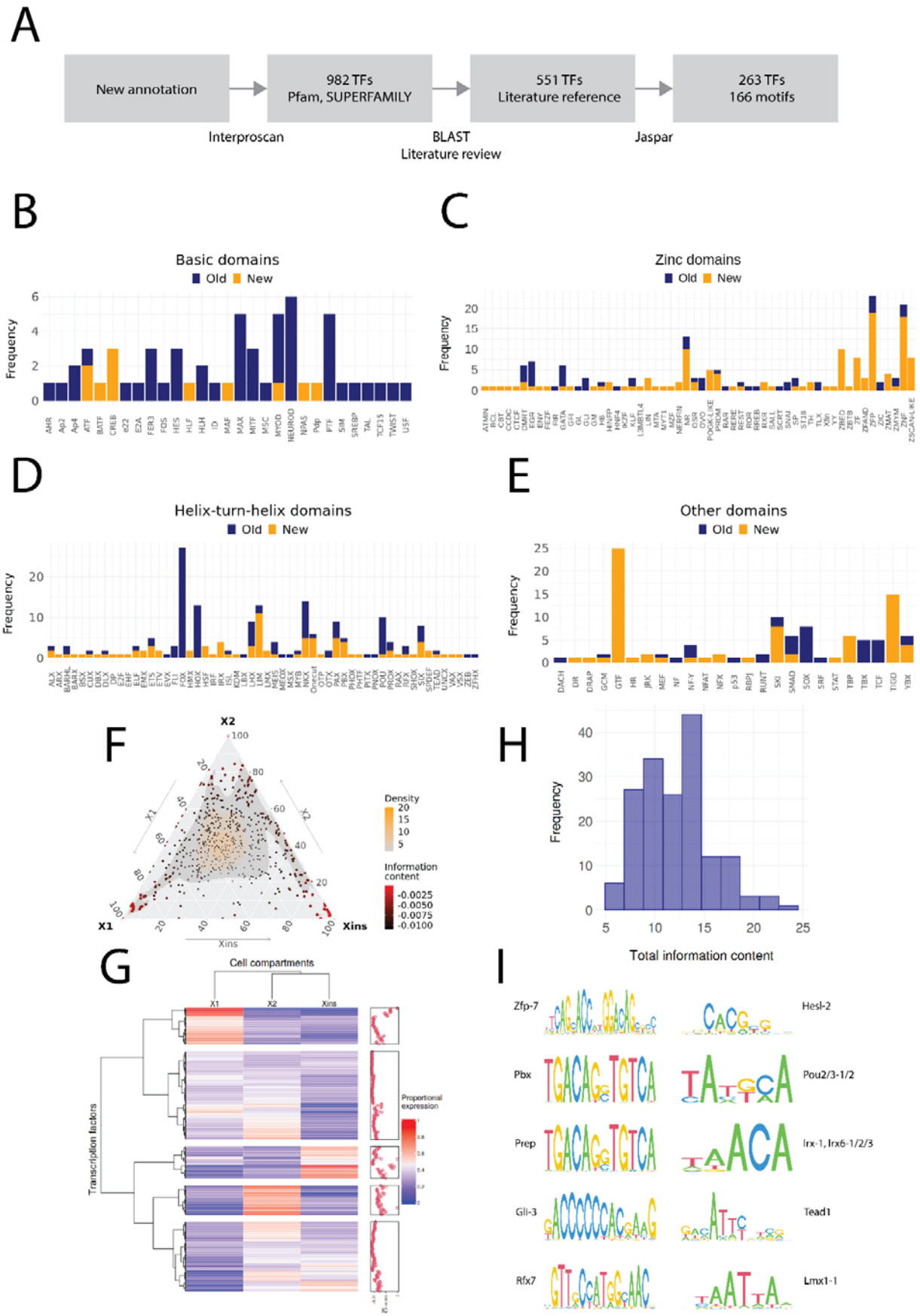
A
New annotation
982 TFs
Pfam, SUPERFAMILY
551 TFs
Literature reference
263 TFs
166 motifs
Interproscan
BLAST
Literature review
Jaspar
B
Basic domains
Old
New
Frequency
C
Zinc domains
Old
New
Frequency
D
Helix-turn-helix domains
Old
New
Frequency
E
Other domains
Old
New
Frequency
F
X2
X1
Xins
Density
Information content
G
Cell compartments
X1
X2
Xins
Transcription factors
Proportional expression
H
Frequency
Total information content
I
Zfp-7
Hesl-2
Pbx
Pou2/3-1/2
Prep
Irx-1, Irx6-1/2/3
Gli-3
Tead1
Rfx7
Lmx1-1

Fig. 2 A. Diagram of the TF annotation process. The coding transcripts identified in the new genome annotation were screened for conserved TF structures by using Interproscan and the Pfam and SUPERFAMILY databases. Homologs were then searched for the TFs with BLAST, and the planarian literature was reviewed to assign screened TFs to previously described. The manually reviewed TF proteins were then used to predict motifs. **B, C, D** and **E.** Identified TF families with basic domains (**B**), (**C**), (**D**), and (**E**). “Old” refers to TFs present in the literature, while “New” are TFs without a reference in the literature or new additional homologs. **F.** Ternary plot of proportional expression values in the X1, X2, and Xins cell compartments. Each dot represents one single TF. The information content metric represents the enrichment of a locus to one of the three compartments. **G.** Heatmap and hierarchical clustering of proportional FACS expression profiles for all TFs. **H.** Histogram of the total information content of the identified motifs. **I.** Example sequence logos for some of the best characterised motifs. The motifs of pbx (MA0782.1) and prep (MA0783.1) have almost identical PWMs, but they are listed as separate motif entries in JASPAR.

2.3 Histone modifications and chromosome accessibility mark enhancer-like regions

To identify potential enhancer regions, we analyzed previously generated ChIP-seq data with respect to the enhancer-associated histone modification H3K4me1 in X1 cells (Mihaylova et al., 2018). Furthermore, we sequenced the epigenome of X1 cells with respect to the histone modification H3K27ac to identify genomic regions indicative of an active enhancer state. H3K27ac was enriched at the promoter region, suggesting that the overall epigenomic mechanism of the histone modification is conserved in planarians (Fig. S2 A) (Gaiti et al. 2017). We identified 37,345 H3K27ac peaks and 13,868 H3K4me1 peaks that were generally less than 200 bp wide (Fig. 3 A and B). At H3K4me1 peaks, the H3K27ac signal was strongest at the peak center at almost all peaks, while at H3K27ac peaks the H3K4me1 displayed a bimodal peak at the peak center of most peaks (Fig. 3 C and D). This pattern of H3K4me1 flanking H3K27ac peaks at enhancers has been implicated in mammals, and our data also suggest this seems to be evolutionarily conserved within metazoans (Gorkin et al. 2012, Spicuglia et al. 2012, Pundhir et al. 2016).

As our ChIP-seq data followed well-established enhancer-like patterns, we used the ChIP peaks to select enhancer-like regions. We calculated a mean peak value at all peaks with respect to the H3K27ac and H3K4me1 signal (Fig. 3 E) and we selected all H3K27ac peaks that were at most 500 bp from H3K4me1 peaks and determined these 5,529 peaks to be an initial set of active enhancer-like regions in cycling adult stem cells (Fig. 3 F). The H3K27ac and H3K4me1 signals correlated at these enhancer-like regions more than at the other peaks in the genome (Fig. 3 F).

Furthermore, we calculated the mean peak value with respect to the change in H3K4me1 signal upon RNAi-mediated *Smed-lpt* knockdown (corresponding to the N-terminus of mammalian mll3/4 or kmt2c/d, see Mihaylova et al. 2018). In mammals, Mll3 and Mll4 are two paralogous methyltransferases of the COMPASS family (SET1/MLL) that regulate enhancer activity by mono-methylating H3K4 (Wang et al. 2021). In addition, the mll3/4 methyltransferase complex associates with the histone acetyltransferase p300/CBP that mediates H3K27 acetylation at enhancers and thus gives rise to the active enhancer landscape (Wang et al. 2021). The enhancer-like regions were clearly more responsive to the knockdown of *Smed-lpt* as compared to random points within the genome, and the response was most evident at the center of predicted enhancer-like regions (Fig. 3 G, H and K). Thus, the H3K4me1 reduction after *Smed-lpt* knockdown provides further support for the identification of active enhancers in planarian stem cells.

We optimized and performed ATAC-seq on X1 cells to measure high-resolution chromatin accessibility in conjunction with the histone modifications (Buenrostro et al. 2013), as a potential further source of evidence. Enhancer-like regions identified by ChIP-seq analysis had a more accessible chromatin configuration than random points in the genome, and the peaks of open regions were positioned at the center of the enhancer-like regions implicated by ChIP-seq data

(Fig. 3 I, J and L). Thus ATAC-seq provides independent evidence for these regions being active enhancers in planarian stem cells.

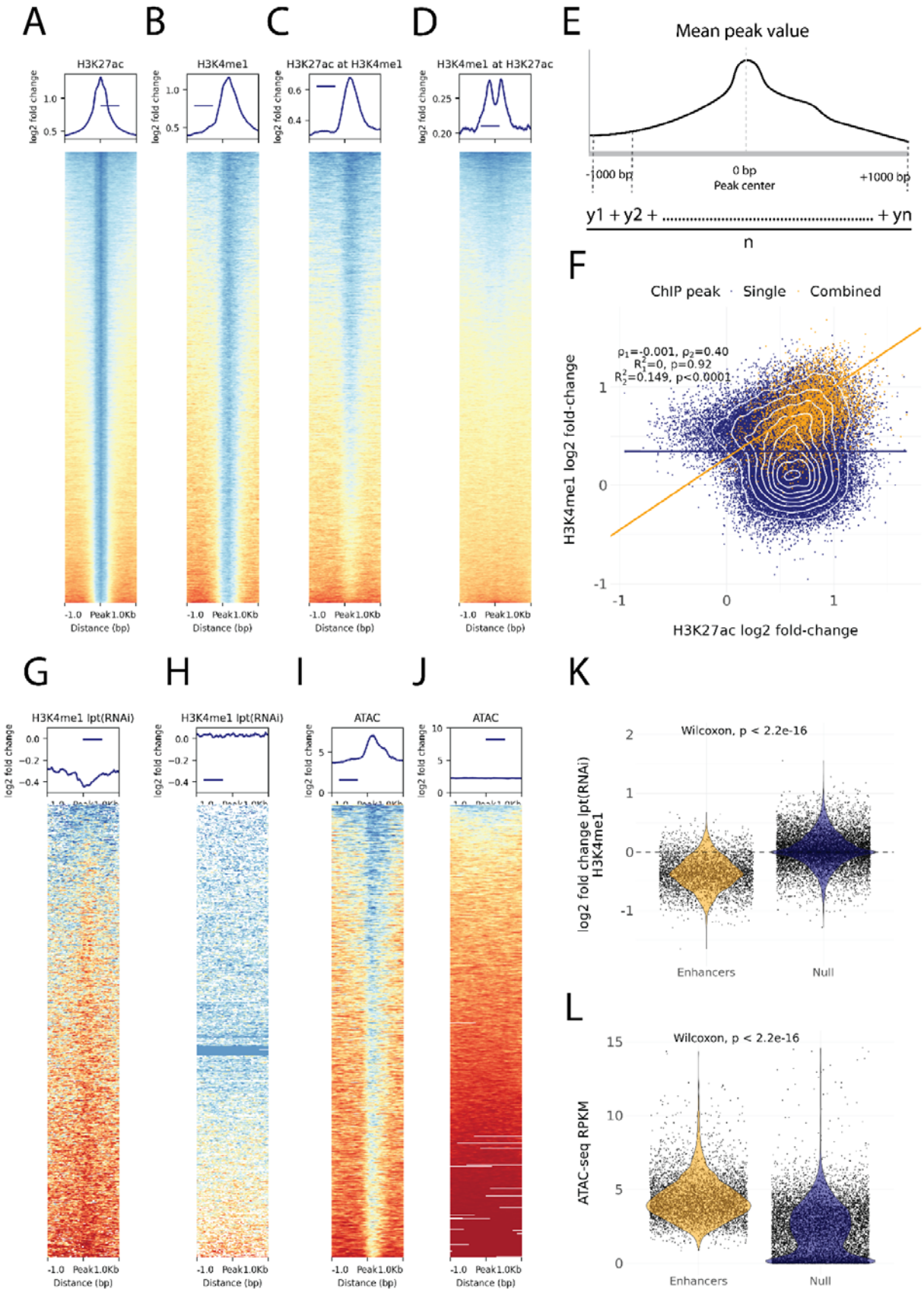
A
H3K27ac
B
H3K4me1
C
H3K27ac at H3K4me1
D
H3K4me1 at H3K27ac
log2 fold change
-1.0 Peak 1.0Kb
Distance (bp)
E
Mean peak value
-1000 bp
0 bp
Peak center
+1000 bp
y1 + y2 + ... + yn
n
F
ChIP peak
Single
Combined
ρ1=-0.001, ρ2=0.40
R21=0, p=0.92
R22=0.149, p<0.0001
H3K4me1 log2 fold-change
H3K27ac log2 fold-change
G
H3K4me1 Ipt(RNAi)
H
H3K4me1 Ipt(RNAi)
I
ATAC
J
ATAC
K
Wilcoxon, p < 2.2e-16
log2 fold change Ipt(RNAi) H3K4me1
Enhancers
Null
L
Wilcoxon, p < 2.2e-16
ATAC-seq RPKM
Enhancers
Null

Fig. 3 A. H3K27ac ChIP-seq signal around independently called H3K27ac peaks. **B.** H3K4me1 ChIP-seq signal around independently called H3K4me1 peaks. **C.** H3K27ac signal around the same H3K4me1 peaks as in B. The signal centers around the center region. **D.** The H3K4me1 signal around the same H3K27ac peaks as in A. The signal displays a bimodal shape around the center region. The signal of the y axis and the heatmap in A, B, C and D is given as the log2 fold-change relative to the input and the peak width is 2,000 bp **E.** A visual diagram of the calculation of the mean peak value. Mean peak value is defined as the mean of signal values taken at an interval of 10 bp from 1000 bp upstream to 1000 bp downstream of the peak center. **F.** Scatter plot of H3K27ac and H3K4me1 log2 fold-change mean peak values at the ChIP-seq peaks. H3K4me1 peaks that are at most 500 bp from a H3K27ac peak are defined as combined peaks and putative enhancer-like regions, while the non-overlapping H3K4me1 and H3K27ac peaks are referred to as single peaks. **G.** The change in the H3K4me1 signal upon *lpt* (mll3/4) RNAi at the combined peaks or putative enhancer-like regions. The change is strongest at the peak center. The log2 fold-change is calculated as the RNAi value relative to the wildtype value. **H.** The H3K4me1 *lpt(RNAi)* signal at 10,000 random regions in the genome. **I.** The ATAC-seq signal at the combined peaks or putative enhancer-like regions. The signal is given as reads per kilo base per million mapped reads (RPKM) **J.** The ATAC-seq signal at 10,000 random regions in the genome. **K.** Violin plot of H3K4me1 *lpt(RNAi)* signal at putative enhancer-like regions and random regions in the genome. The H3K4me1 signal is reduced at enhancers. **L.** Violin plot of the ATAC-seq signal at putative enhancer-like regions and random regions in the genome. The ATAC-seq signal is higher at enhancers.

2.4 Gene regulatory networks involving fate-specific transcription factors in neoblasts

Having defined enhancer-like regions (Supplemental File 2), integration with both TF and potential target gene expression data allowed us to begin constructing preliminary gene regulatory networks (GRNs) to demonstrate the utility of our neoblast enhancer predictions. Firstly, we assigned enhancer-like regions to their closest gene, assuming a high probability that these will be the putative target genes of enhancer-like regions. The distance from the enhancer-like regions to the target gene transcription start site (TSS) varied from being proximal to the promoter region to being as far as 89 kb away (Fig. 4 A). We found that TFs themselves were significantly enriched within the set of all predicted target genes (Fig. 4 B). Furthermore, these TFs were in turn enriched for the 43 fate-specific transcription factors (FSTFs) previously shown to be expressed in S/G2/M neoblasts (Fig. 4 B) (Raz et al. 2021). In addition to transcription factor activity, the putative target genes linked to enhancer-like regions were enriched for RNA metabolic and biosynthetic processes and other biosynthetic processes (Fig. 4 C). These data suggest that our enhancer predictions represent real enhancers involved in regulating key aspects of neoblast biology.

To establish more persuasive and direct regulatory links between TFs and enhancer-like regions we used ATAC-seq foot printing to detect bound motifs in the planarian stem cell genome (Bentsen et al. 2020) (Fig. 4 D), Supplemental File 3). ATAC-seq footprints are short inaccessible or less accessible regions within an otherwise accessible region, indicative of DNA binding by a TF or another DNA-binding protein (Bentsen et al. 2020) (Fig. 4 D). As for the raw ATAC-seq signal, footprint scores were higher in enhancer-like regions than in random regions of the genome (Fig. S2 C and D). Overall foot printing analysis found 22,489 putatively bound TF motifs in the enhancer-like regions but no bound motifs in random regions, providing further support that these regions are bona fide enhancers. The TFs with the most binding sites predicted by footprint analysis in the genome overall were *irx-1*/*irx6-1/2/3*, *egr4-1/2, smad1/9*, *sp5*, and *nf-ya1/2*, while the TFs with the least bound motifs were *mef2-1*, *mef2-2*, *phox*, *pou6-2*, and *lmx1-2* (Fig. 2 E, Supplementary file 4). Overall, these data provide a resource of studying GRNs and specific regulatory interactions in planarians neoblasts.

To test the utility of these data for building and testing GRNs, we focused on the previously defined FSTFs, established the regulatory links between them and constructed a putative GRN based on our datasets (Fig. 4 F, Supplemental File 5). In total, we could include 35/43 FSTFs into a GRN prediction and found evidence for multiple cross-regulatory links that may serve to allow stem cells to decide between fates, e.g. neural FSTFs regulate enhancers of muscle FSTFs and vice versa (Fig. 4 F). The interactions in this GRN allow specific hypotheses to be formed and tested in the future. As examples, we present in detail the genomic region of the epidermal FSTF *p53*, the muscle FSTF *myoD*, and the neural FSTFs *nkx6-1* and *soxB1-1* (Fig. 5 A, B, C and D). The

epidermal FSTF *p53* has multiple TFs predicted to bind at the promoter region, including a muscle FSTF (*myoD*) (Scimone et al. 2014, 2017), and TFs relating to position (*zic-1*) (Vásquez-Doorman and Petersen 2014) and cell migration (*zeb-1* and *snail-1/2*) (Abnave et al. 2017) (Fig. 5 A). An enhancer-like region within an intron of the upstream gene *ythdc23* was predicted to be associated with the downstream muscle FSTF *myoD,* as the TSS of *myoD* is closer of the two and *ythdc23* is lowly expressed in X1 and X2 cells. The enhancer-like region includes bound TF motifs such as the neoblast-enriched zinc finger *fir-1*, the muscle segment homeobox gene *msx*, the neural *pou-p1*, *sp6-9*, *lhx2-1*, and position-related *foxA(P)* (Fig. 5 B). A distal enhancer-like region was found downstream of *nkx6-1,* including bound TF motifs such as the neoblast-enriched *egr-1*, the neural *ovo-1* and *fli-1*, the muscle-related *twist*, the position-related *smad4-1*, and the pigmentation-related *ets-1* (Fig. 5 C). A distal enhancer-like region was found upstream of *soxB1-1*, including bound TF motifs such as the position-related *smad1* and *foxA(P)* (Molina et al. 2007, Reddien et al. 2007, Pascual-Carreras et al. 2021), neural *da* (Cowles et al. 2013), intestinal *hnf4* (Wagner et al. 2011, Scimone et al. 2014, van Wolfswinkel et al. 2014), and *zeb-1* and *snail-1/2* relating to cell migration (Abnave et al. 2017) (Fig. 5 D). In the future, these regulatory links can be verified and studied by experiments using functional genomics involving the RNAi knockdown of individual genes and subsequent RNA-seq and ATAC-seq analysis.

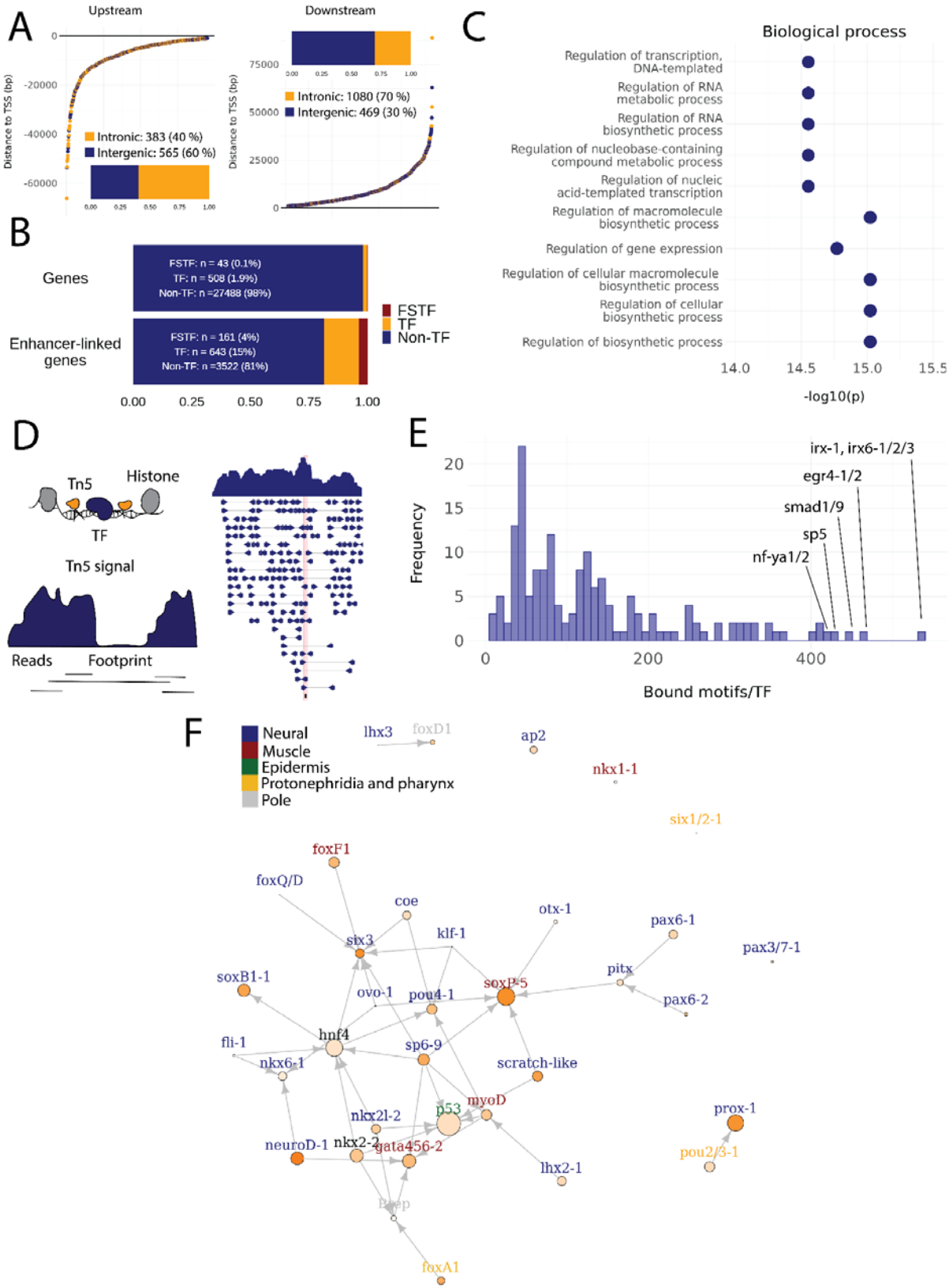
A
Upstream
Downstream
Distance to TSS (bp)
Intronic: 383 (40 %)
Intergenic: 565 (60 %)
Intronic: 1080 (70 %)
Intergenic: 469 (30 %)
B
Genes
FSTF: n = 43 (0.1%)
TF: n = 508 (1.9%)
Non-TF: n =27488 (98%)
Enhancer-linked genes
FSTF: n = 161 (4%)
TF: n = 643 (15%)
Non-TF: n =3522 (81%)
FSTF
TF
Non-TF
C
Biological process
Regulation of transcription, DNA-templated
Regulation of RNA metabolic process
Regulation of RNA biosynthetic process
Regulation of nucleobase-containing compound metabolic process
Regulation of nucleic acid-templated transcription
Regulation of macromolecule biosynthetic process
Regulation of gene expression
Regulation of cellular macromolecule biosynthetic process
Regulation of cellular biosynthetic process
Regulation of biosynthetic process
-log10(p)
D
Tn5
Histone
TF
Tn5 signal
Reads
Footprint
E
Frequency
irx-1, irx6-1/2/3
egr4-1/2
smad1/9
sp5
nf-ya1/2
Bound motifs/TF
F
Neural
Muscle
Epidermis
Protonephridia and pharynx
Pole
lhx3
foxD1
ap2
nkx1-1
six1/2-1
foxF1
foxQ/D
coe
otx-1
pax6-1
pax3/7-1
six3
klf-1
pitx
soxB1-1
ovo-1
pou4-1
soxP-5
pax6-2
fli-1
hnf4
sp6-9
nkx6-1
scratch-like
myoD
nkx2l-2
p53
prox-1
neuroD-1
nkx2-2
gata456-2
pou2/3-1
lhx2-1
Pitx
foxA1

Fig. 4 A. Distance of enhancer-like regions to the transcription start site (TSS) of nearest genes. The nearest genes are assumed to be the putative target genes of the enhancer-like regions. **B.** The proportion of FSTFs, TFs and non-TF genes of all planarian genes and of the target genes linked to enhancer-like regions. FSTFs and TFs are enriched in the set of target genes (χ^2=2660, $p < 0.001$). **C.** Gene ontology (GO) analysis of biological processes of enhancer-linked target genes compared to all genes in the planarian genome. **D.** Schematic overview of ATAC-seq foot printing (Bentsen et al. 2020). Both the tagmenting enzyme Tn5 and TF bind to accessible regions of the genome. Footprints are short and sharp inaccessible segments in otherwise accessible regions. The genomic track figure represents an example of ATAC-seq reads and footprint score overlapping with a *sp5* motif in the genome. **E.** Histogram of the total number of bound motifs in the genome for each TF. The TFs with most bound motifs in the genome are marked. **F.** Putative GRN of FSTFs in neoblasts. The epidermal FSTFs are labelled in green, the pole-associated FSTFs are labelled in grey, the muscle FSTFs are labelled in red, the intestinal FSTFs are labelled in black, and the neural FSTFs are labelled in blue. The size of the nodes reflects the absolute expression in X1 cells (transcripts per million reads or TPM), and the color of the node reflects the proportional expression in X1 cells (the more orange, the higher proportional expression). The arrows represent a regulatory link, but don't discriminate between positive and negative interactions.

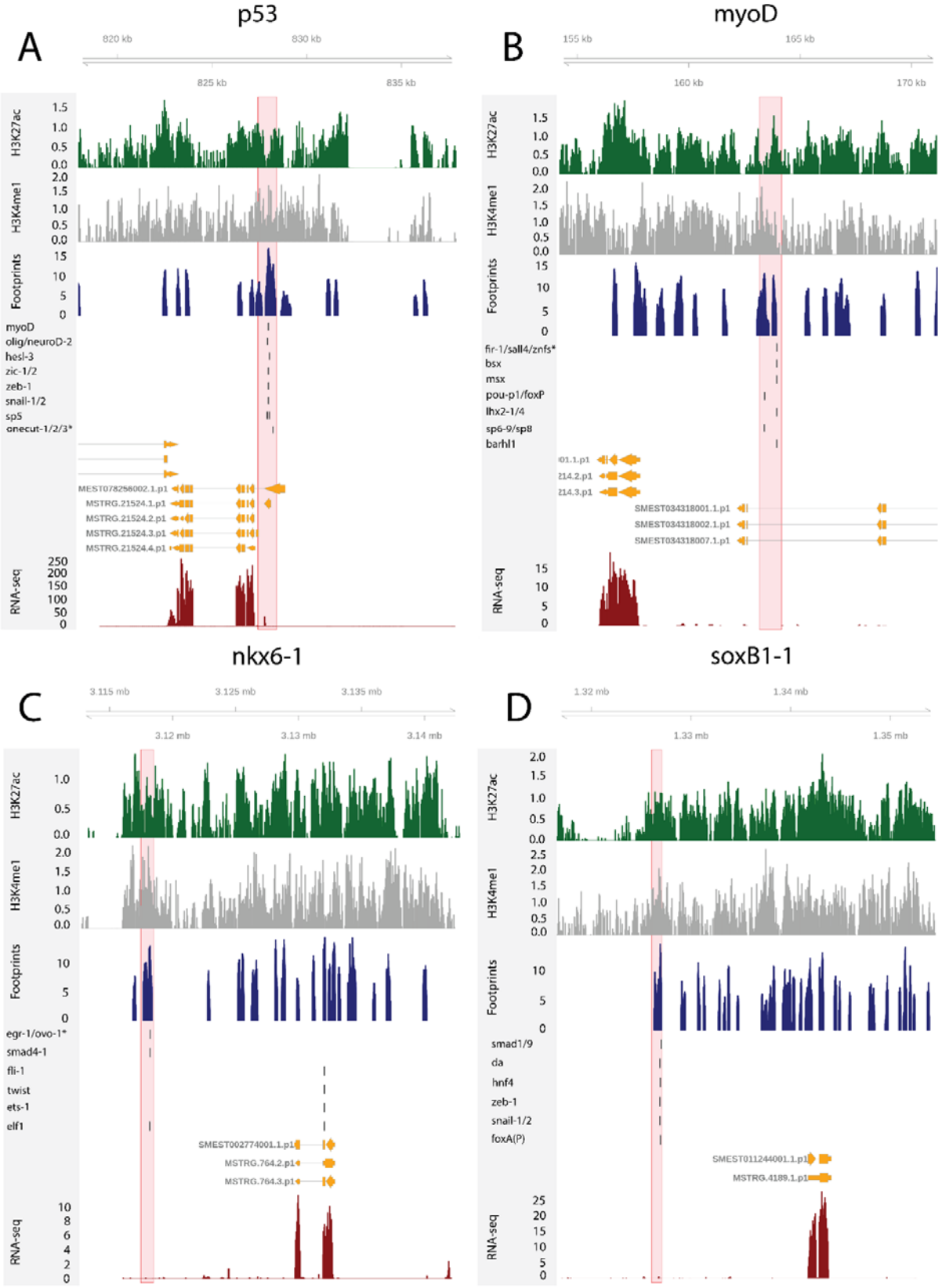
A
p53
820 kb
830 kb
825 kb
835 kb
H3K27ac
H3K4me1
Footprints
myoD
olig/neuroD-2
hesl-3
zic-1/2
zeb-1
snail-1/2
sp5
onecut-1/2/3*
MSTRG.21524.1.p1
MSTRG.21524.2.p1
MSTRG.21524.3.p1
MSTRG.21524.4.p1
RNA-seq
B
myoD
155 kb
165 kb
160 kb
170 kb
fir-1/sall4/znfs*
bsx
msx
pou-p1/foxP
lhx2-1/4
sp6-9/sp8
barhl1
SMEST034318001.1.p1
SMEST034318002.1.p1
SMEST034318007.1.p1
C
nkx6-1
3.115 mb
3.125 mb
3.135 mb
3.12 mb
3.13 mb
3.14 mb
egr-1/ovo-1*
smad4-1
fli-1
twist
ets-1
elf1
SMEST002774001.1.p1
MSTRG.764.2.p1
MSTRG.764.3.p1
D
soxB1-1
1.32 mb
1.34 mb
1.33 mb
1.35 mb
smad1/9
da
hnf4
zeb-1
snail-1/2
foxA(P)
SMEST011244001.1.p1
MSTRG.4189.1.p1

Fig. 5 The genomic tracks of four FSTFs (**A**, **B**, **C** and **D**). The first track is the H3K27ac signal (log2 scale compared to the input sample), the second track is H3K4me1 signal (log2 scale compared to the input sample), the third track is the ATAC-seq footprint score (TOBIAS footprint score), the fourth track represents bound motifs in an enhancer-like region (as predicted by TOBIAS), the fifth track represents the transcripts of the gene in the new annotation, and the sixth track represents the gene expression level of the gene in X1 cells (RNA-seq FPKM or Fragments Per Kilobase of transcript per Million mapped reads). **A.** Genomic track of the epidermal FSTF *p53*. The promoter region marked in red (width 1000 bp) has multiple bound motifs. **B.** Genomic track of the muscle FSTF *myoD*. An enhancer-like region marked in red (width 1000 bp) is located in the intron of the gene YTH domain containing 2 (*ythdc23*). The enhancer-like region is closer to the *myoD* promoter than the *ythdc23* promoter and *ythdc23* is not expressed almost at all in neoblasts, and hence the enhancer-like region is likely linked to *myoD* rather than *ythdc2*. **C.** The neural FSTF *nkx6-1* has a distal enhancer-like region marked in red (width 1000 bp) downstream of the gene body. **D.** The neural FSTF *soxB1-1* has a distal enhancer-like region marked in red (width 1000 bp) upstream of the gene body.

2.5 Multiple enhancers are linked to unstudied transcription factors

In addition to the relatively well-known planarian FSTFs (Raz et al., 2021), we investigated GRNs relating to less well studied TFs with multiple bound motifs in associated enhancer-like regions, as these may also have a potentially central role in the neoblast regulation. We defined the number of enhancer-like regions, number of bound motifs, and number of unbound motifs for each TF (Fig. 6 A, Supplementary File 6). In addition, we selected TFs expressed in the X1 compartment (proportional expression of X1 > 1/3) and constructed the putative GRN summarizing all the 502 regulatory links of these 109 TFs (Fig. 6 B). In the X1 compartment (X1% > 1/3), the least well studied TFs with the most bound motifs in enhancer-like regions were *znf596*, *tbx-20*, *hesl-1*, *atoh8-1*, and *ikzf1* (Fig. 6 A and B, Supplementary table 2). Outside the X1 compartment, *msx*, *zf-6*, *vsx*, and *pdp-1* had the most bound motifs (Fig. 6 A).

The TF with most regulatory interactions, *znf596,* has been characterized to be expressed in neoblasts and more specifically in the neoblasts committed to the neural fate, but otherwise its function is unknown (Fig. 6 A, B, and C, Supplementary File 6) (Fincher et al. 2018). We found that numerous FSTFs were binding to putative intronic enhancers within *znf596*, including neural FSTFs *sp6-9*, *neuroD-1*, and *ovo-1*, the intestinal FSTF *gata456-1* and/or the muscle FSTF *gata456-1*, and the position control genes (PCGs) *smad1/9* (BMP signaling), *isl-1*, and *prep* (Fig. 6 C). Furthermore, numerous other previously uncharacterized zinc fingers were linked to *znf596*. In single-cell transcriptomics data, the gene is clearly expressed in a subset of neoblasts, calling for further mechanistic studies of its regulatory function. The binding motif of *znf596* could not be predicted, so its target genes cannot .be implicated within the current data sets and approaches

We identified the same five planarian *tbx* genes that have been previously reported, namely *tbx-1/10*, *tbx-2/3a*, *tbx-2/3b*, *tbx-2/3c*, and *tbx-20* (Tewari et al. 2019). In mammals, the *tbx* genes have multiple roles in development, and *tbx3* is involved in regulating pluripotency of embryonic stem cells (Baldini et al. 2017, Khan et al. 2020). Nonetheless, the function of planarian *tbx* genes has not been clarified. Here, we found several bound motifs of *tbx* genes in the genome (*tbx-1/10* 91 motifs, *tbx-2/3a/c* 81 motifs, *tbx-20* 62 motifs) and multiple bound motifs at enhancer like regions linked to *tbx* genes (*tbx-20* 142 motifs, *tbx-2/3c* 34 motifs, *tbx-2/3a* 11 motifs) (Fig. 6 A, B, D, Supplementary table Enhancers). *tbx-20* had a clear intronic enhancer-like region and a distal enhancer-like region, both containing motifs for TFs implicated as positional control genes (PCGs) (hox1, hox4a, prep, smad1, smad9, zic-1, sp5) (Molinaro et al. 2007, Reddien et al. 2007, Felix and Aboobaker 2010, Vásquez-Doorman and Petersen 2014, Tewari et al. 2019), involved in neoblast migration TFs (snail-1 and snail-2) (Abnave et al. 2017), and various other TFs (Fig. 6 D).

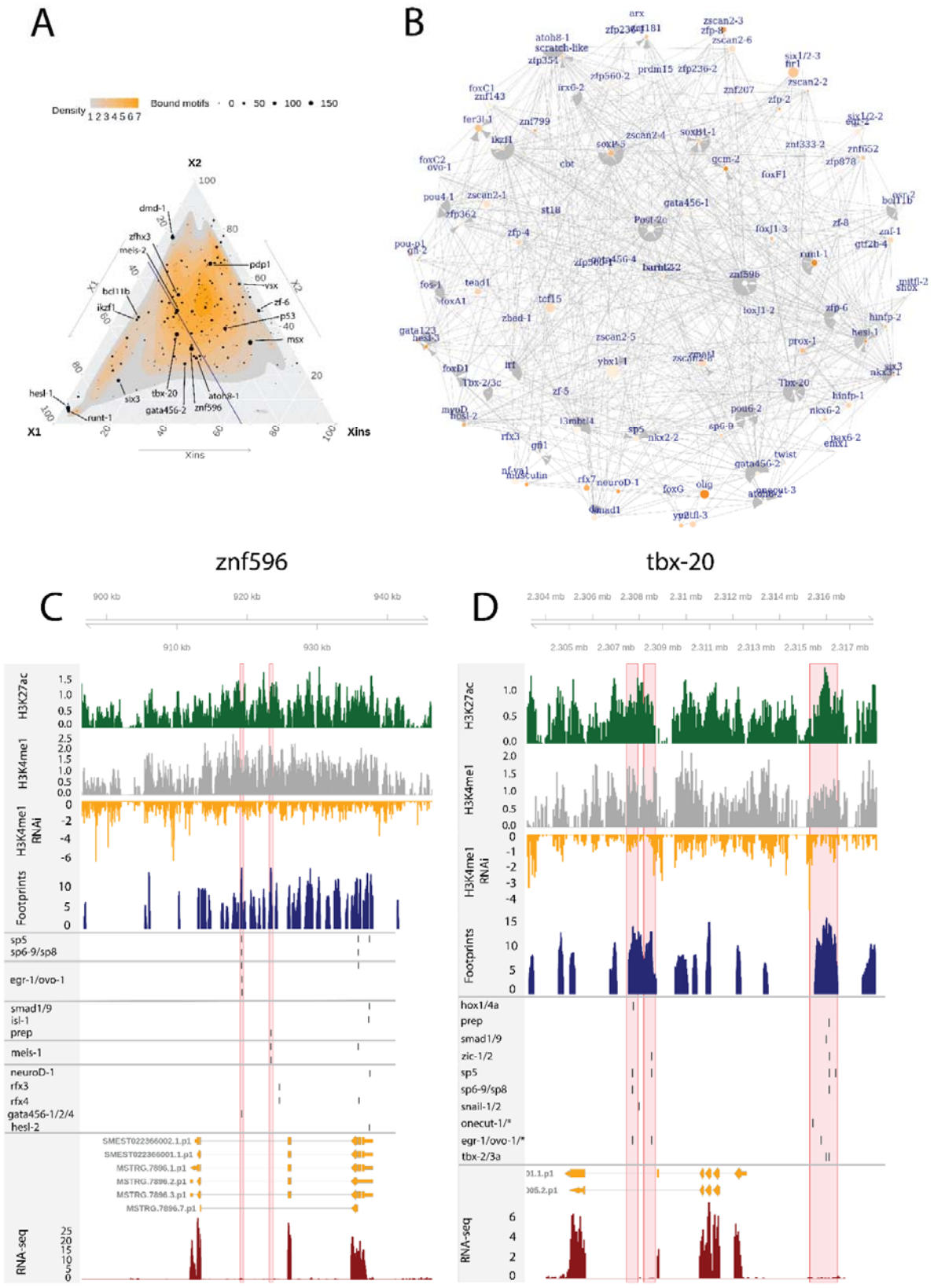
A
Density
Bound motifs
X1
X2
Xins
B
znf596
tbx-20
C
D
H3K27ac
H3K4me1
H3K4me1 RNAi
Footprints
RNA-seq

Fig. 6 A. Ternary plot of the proportional expression of all TFs in the cell compartments (X1, X2, and Xins), and the number of bound motifs linked to the TFs. The size of the dot represents the number of bound motifs. The TFs with most bound motifs in enhancer-like regions are named and marked. **B.** The putative GRN of all TFs in the X1 compartment. The GRN includes all TFs with a X1 proportional expression higher than 1/3. The size of the nodes reflects the absolute expression in X1 cells (transcripts per million reads or TPM), and the color of the node reflects the proportional expression in X1 cells (the more orange, the higher proportional expression). The genomic tracks of *znf596* (**C**) and *tbx-20* (**D**). The first track is the H3K27ac signal (log2 scale compared to the input sample), the second track is H3K4me1 signal (log2 scale compared to the input sample), the third track is the log2 fold-change upon lpt(RNAi), the fourth track is ATAC-seq footprint score (TOBIAS footprint score), the fifth track represents bound motifs in an enhancer-like region (as predicted by TOBIAS), the sixth track represents the transcripts of the gene in the new annotation, and the seventh track represents the gene expression level of the gene in X1 cells (RNA-seq FPKM or Fragments Per Kilobase of transcript per Million mapped reads). **C.** Genomic track of *znf596*. Two intronic enhancer-like regions are marked in red (width 500 bp). **D.** Genomic track of *tbx-20*. Two intronic (width 500 bp) and one distal enhancer-like region (width 1200 bp) are marked in red. The distal enhancer-like region displays eRNA activity.

2.6 Elements of the planarian positional GRN are active in neoblasts

Finally, we studied enhancer-like regions and putative regulatory links associated with the well-known planarian positional control genes (PCGs). Planarians are a primary model system to understand how positional information guides and directs stem cell function during regeneration (Reddien, 2018), and therefore some functional genomics data exists with regard to PCGs, enabling the comparison with our findings (Tewari et al., 2019). A constitutive positional information system is established by the regional expression pattern of PCGs that pattern the anterior-posterior (AP), dorsal-ventral (DV), and medial-lateral (ML) axes (Reddien, 2018).

The AP axis is patterned by the Wnt signalling pathway: high Wnt activity specifies posterior identity, while low activity specifies anterior identity (Reddien, 2018). Upon β-catenin (knockdown, posteriorly expressed *sp5* and the Hox genes *post-2c*, *post-2d*, *lox5a*, and *hox4b* are rapidly down-regulated (Tewari et al. 2019). In multiple vertebrate species, *sp5* is known to be a direct target of Wnt signalling (Weidinger et al. 2005, Fujimura et al. 2007), and this regulatory link seems to be conserved in planarians (Tewari et al. 2019). Here, we found a bound *sp5* footprint in the vicinity of the *post-2c* and *lox5a* promoters, suggesting and further supporting that *sp5* regulates the expression of these posterior PCGs (Fig. 7 A). *sp5* had 431 bound motifs and was the fourth most bound motif in X1 cells overall, further suggesting that *sp5* mediates the global positional information provided by Wnt signalling in planarians (Supplementary table 2).

Through ATAC-seq and ChIPmentation techniques, Pascual-Carreras et al. (2020) screened for cis-regulatory elements in planarian tissues in *notum* and *wnt1* (RNAi) animals. Upon *wnt1* knockdown, posterior Hox genes *hox4b*, *post-2c*, *post-2b*, *lox5a*, *lox5b* and *wnt11-1*, *wnt11-2*, *fzd4* and *sp5* were down-regulated, replicating the results of β-catenin knockdown (Pascual-Carreras et al. 2020). In addition, two *foxG* binding sites were found in the first intron of *wnt1,* and *foxG* knockdown was found to phenocopy *wnt1* knockdown, supporting that *foxG* is an upstream regulator of *wnt1* (Pascual-Carreras et al. 2020). Here, we found one enhancer-like region in the first intron of *wnt1* with a high level of H3K27ac, H3K4me1 and ATAC-seq foot printing scores for Fox family transcription factors (Fig. 7 B). This motif implicated by foot printing analysis is the same as one of the motifs described by previous work (Pascual-Carreras et al. 2020). We did not see evidence of binding at the second motif, and neither was this motif within one of our predicted enhancers (Fig. 7 B).

Although the planarian Hox genes are expressed in a regionalized manner along the AP axis, the knockdown of the genes apart from *post-2d* do not result in homeostatic or regeneration-associated phenotypic changes (Currie et al. 2016, Arnold et al. 2021). Instead, the five Hox genes *hox1*, *hox3a*, *hox3b*, *lox5b*, and *post2b* have been shown to be involved in asexual reproduction by regulating fission at potential cryptic segment boundaries (Arnold et al. 2021). Here, we did not find

any bound motifs implicated by foot printing associated with *hox1*, *hox3a*, and *hox3b*, and only a few at the promoters of *lox5b* and *post2b*, suggesting that this component of the GRN of asexual reproduction is not active in neoblasts.

Prep, *zic-1*, *isl-1* and *foxD1* are PCGs that are expressed in the anterior pole of both intact and regenerating planarians (Felix and Aboobaker 2010, Vásquez-Doorman and Petersen 2014, Vogg et al. 2014). Interestingly, we found that these TFs are bound to several motifs in enhancer-like regions of X1 cells (Supplementary table Enhancers, *prep* 132 motifs, *zic-1/2* 330 motifs, *isl-1* 42 motifs, *foxD1/2/3* 42 motifs) and a few motifs were found to bind to the enhancer-like regions linked to *prep*, *zic-1*, *isl-1* and *foxD1* (Fig. 7 C, Supplementary table Enhancers). *Prep* has enhancer-like regions that have bound motifs of smad factors (Bmp signalling components) and *nkx2-2/3/4* (Fig. 7 C). Knockdown of *nkx2-2* (also known as *DTH-1* or *nkx2.2*) causes blastemal defects both at the anterior and posterior ends, while knockdown of *prep* leads to defects in the anterior, suggesting that *nkx2-2* might work upstream of *prep* at the anterior end (Felix and Aboobaker 2010, Forsthoefel et al. 2012).

Lastly, we studied the Hedgehog signaling receptor *patched-1* (*ptch-1*) and found an intronic enhancer-like region containing multiple bound motifs, including the Hedgehog pathway TFs *gli-1/2* as would be expected for, the Wnt effector *sp5*, the Hox genes *lox5a*/*b*, BMP signaling components (*smad4-1*, *smad1*/*9*), the anterior pole TFs *prep* and *zic-1*, the cell migration factors *snail-1/2*, and the muscle TFs *myoD* and *twist* (Fig. 7 D). Hedgehog signaling in planarians is known to have pleiotropic functions, including interaction with the Wnt signaling pathway and glial cell function (Rink et al. 2009, Yazawa et al. 2009, Wang et al. 2016). Taken together, we found that elements of positional control gene regulatory network are active in planarian S/G2/M neoblasts and an integral part of the overall GRN determining neoblast behavior (Fig. 7 E).

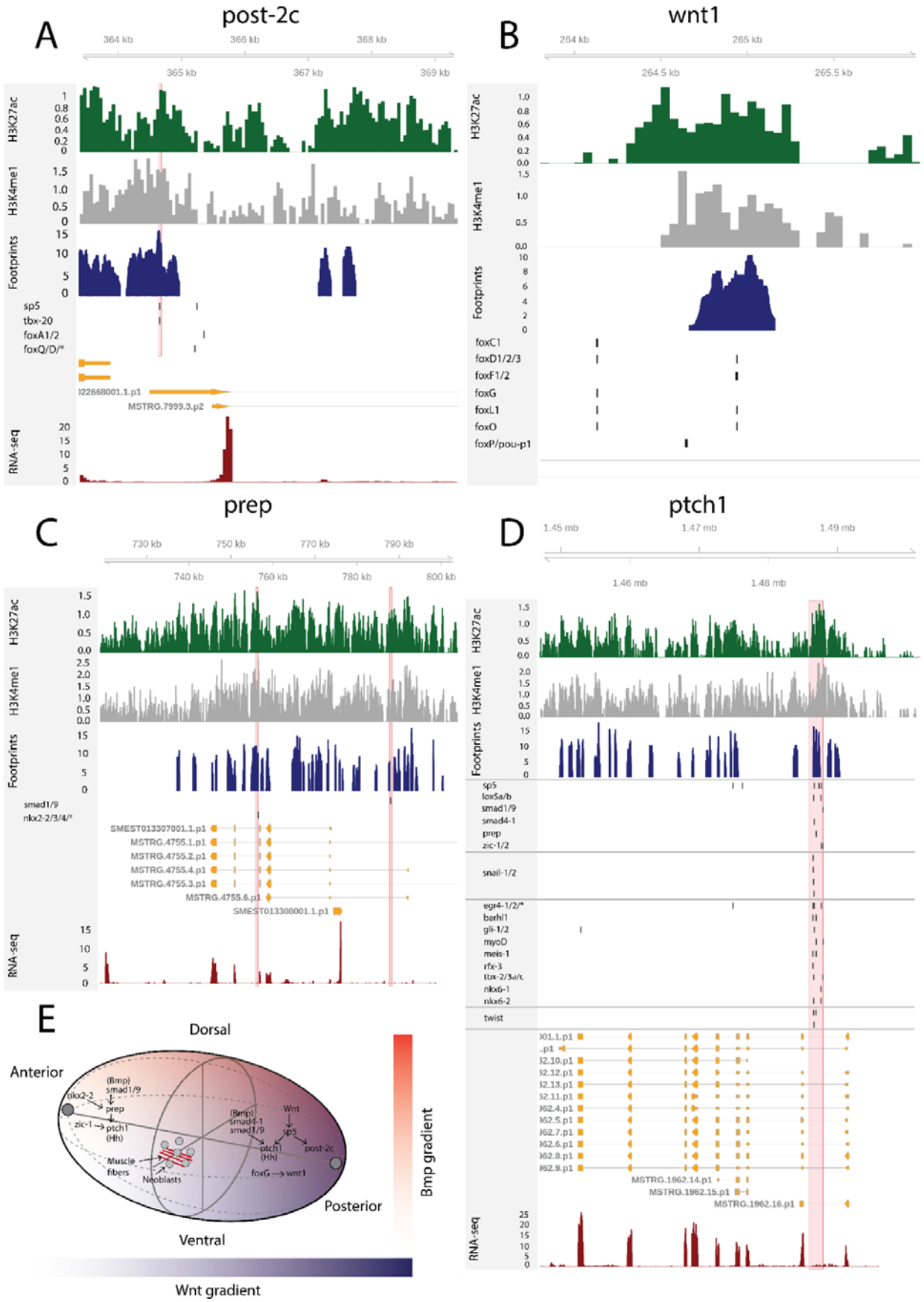
A
post-2c
364 kb
366 kb
368 kb
365 kb
367 kb
369 kb
H3K27ac
H3K4me1
Footprints
sp5
tbx-20
foxA1/2
foxQ/D/*
MSTRG.7999.3.p2
RNA-seq
B
wnt1
264 kb
265 kb
264.5 kb
265.5 kb
H3K27ac
H3K4me1
Footprints
foxC1
foxD1/2/3
foxF1/2
foxG
foxL1
foxO
foxP/pou-p1
C
prep
730 kb
750 kb
770 kb
790 kb
740 kb
760 kb
780 kb
800 kb
H3K27ac
H3K4me1
Footprints
smad1/9
nkx2-2/3/4/*
SMEST013307001.1.p1
MSTRG.4755.1.p1
MSTRG.4755.2.p1
MSTRG.4755.4.p1
MSTRG.4755.3.p1
MSTRG.4755.6.p1
SMEST013308001.1.p1
RNA-seq
D
ptch1
1.45 mb
1.47 mb
1.49 mb
1.46 mb
1.48 mb
H3K27ac
H3K4me1
Footprints
sp5
lox5a/b
smad1/9
smad4-1
prep
zic-1/2
snail-1/2
egr4-1/2/*
barhl1
gli-1/2
myoD
meis-1
rfx-3
tbx-2/3a/c
nkx6-1
nkx6-2
twist
MSTRG.1962.14.p1
MSTRG.1962.15.p1
MSTRG.1962.16.p1
RNA-seq
E
Dorsal
Anterior
Posterior
Ventral
nkx2-2
(Bmp)
smad1/9
prep
zic-1
ptch1
(Hh)
Muscle
fibers
Neoblasts
(Bmp)
smad4-1
smad1/9
Wnt
sp5
ptch1
(Hh)
post-2c
foxG → wnt1
Bmp gradient
Wnt gradient

Fig. 7. The genomic tracks of *post-2c* (**A**), *wnt1* (**B**), *prep* (**C**), and *ptch1* (**D**). The first track is the H3K27ac signal (log2 scale compared to the input sample), the second track is H3K4me1 signal (log2 scale compared to the input sample), the third track is ATAC-seq footprint score (TOBIAS footprint score), the fourth track represents bound motifs in an enhancer-like region (as predicted by TOBIAS), the fifth track represents the transcripts of the gene in the new annotation, and the sixth track represents the gene expression level of the gene in X1 cells (RNA-seq FPKM or Fragments Per Kilobase of transcript per Million mapped reads). **A.** Genomic track of *post-2c*. One upstream cis-regulatory regions is marked in red (width 50 bp). The foxQ/D* footprint represents motif MA0851.1 corresponding to TFs foxQ/D, foxJ1-1, foxJ1-2, foxJ1-3, foxJ1-4, foxJ1-5, foxN2/3-1, foxN2/3-2, and foxN2/3-3. **B.** Genomic track of *wnt1*. One intronic enhancer-like region with one potential Fox family TF binding motif with evidence of binding from foot printing analysis, other Fix TF motifs lie outside the enhancer predicted by ChIP-seq data (Pascual-Carreras et al. 2020). **C.** Genomic track of *prep*. Two intronic enhancer-like regions are marked in red (width 500 bp). **D.** Genomic track of *ptch1*. One intronic enhancer-like region (width 2000 bp) is marked in red. **E.** Schematic of PCG regulatory links discussed.

3. Discussion

In this study, we improved the current genome annotation of *Schmidtea mediterranea* by integrating available RNA-seq data and identified new coding and non-coding transcripts. We reviewed the literature on both computationally and experimentally derived planarian transcription factors and defined a high confidence set of transcription factors and where possible binding motifs. We developed and performed ATAC-seq on the proliferating stem cell compartment and determined genomic regions of open chromatin. We analyzed genome-wide profiles of the histone modifications H3K27ac and H3K4me1, along with chromatin accessibility data and used these epigenetic data to delineate putative active enhancers in planarian stem cells. Lastly, we identified potential target genes and binding transcription factors of the enhancers and constructed hypothetical gene regulatory networks active in the stem cells.

We found more than 50,000 new isoforms to known transcripts with our annotation approach, and this has already helped to improve the clustering of single-cell RNA-seq data (García-Castro et al. 2020). This demonstrated that the principle of integrating available RNA-seq data into a comprehensive expression-driven annotation can significantly improve genome annotation. We were also able to annotate more than 7000 new loci, with over 3000 predicted to be coding, increasing the number of protein coding genes by more than 10%. These new proteins tend to be shorter, are more likely to be proteins specific to planarians, but are expressed at similar levels to the rest of the coding transcriptome. This, together with the extensive annotation of alternate splice forms gives a more complete picture of the genome of the model planarian.

Altogether, we annotated 551 TFs that were distributed evenly with respect to enrichment across all three FACS cell compartments in planarians (X1, X2, and Xins). In planarians, most studied functionally have been involved in regulating differentiation. Here we define many neoblast-enriched TFs that have not been formally studied, with potential roles in maintaining neoblast pluripotency and potentially early lineage commitment. A future systematic functional screen of these may help to uncover the GRN network that maintains pluripotency analogous to that in mammals and vertebrates (Takahashi and Yamanaka 2006, Takahashi et al. 2007). In particular, we found a large number of uncharacterized planarian zinc finger TFs, and this large and diverse family of TFs is still relatively poorly studied, perhaps because they evolve relatively rapidly in metazoans in general and often homology cannot be confidently assigned across phyla (Albà 2017, Cassandri et al. 2017, Najafabadi 2017). Overall, for the main goals of this study, our analysis and identification of TFs allowed us to confidently assign likely binding motifs to just under half of the annotated planarian TFs (Figure 2 E, Supplementary File 1). In the future as the number of planarians studies using ATAC data increases, and by looking at the actual sequence motifs implicated by foot printing analyses it should be possible to refine motifs for planarian TFs and

even define motifs for some TFs to which motifs could not be assigned, for example in the diverse zinc finger TF group.

Based on combining epigenomic experiments and data types, we could identify putative intergenic and intronic enhancers in the planarian the genome of proliferating stem cells. The combined use of ChIP-seq data, RNAi of a histone methyltransferase combines with ChIP-seq, ATAC-seq data and foot printing analyses together provided strong evidence for the identification of bona-fide planarians stem cell enhancers. In the future genome wide ChIP-seq against the transcriptional cofactor p300 could serve as a complementary high-throughput approach to further substantiate these enhancers (Visel et al. 2009; Schwaiger et al. 2014), if available antibodies recognize the planarian ortholog of this protein (Fraguas et al. 2021, Stelman et al. 2021).

Verification of the function and targets of planarian enhancers is currently not possible using traditional approaches as no transgenic reporter technologies enabling enhancer-reporter constructs to be assayed are available. Therefore we are bound to relying on less direct evidence from genome wide sequencing technologies. Here, we assigned enhancers to target genes based on distance and the expression of the target gene, similar to established protocols used by others (Duren et al. 2017). We find cases where regulatory interactions suggested by previous expression base studies of neoblast fate control and RNAi based studies of gene function are supported by the GRNs interactions uncovered by our analyses. We find evidence that regulators of one differentiation lineage bind the enhancers of genes that regulate another, presumably acting as repressors. Experiments based on our data combining RNAi against TFs with RNA-seq and ATAC-seq approaches will allow these GRNS to be studied further to help us understand how stem cells drive regeneration and homeostasis in planarians. In the future both promoter-capture HiC and the co-accessibility of putative enhancers and promoters in scATAC-seq data would offer further computational possibilities to study these promoter-enhancer interactions (Schoenfelder et al. 2015, Pliner et al. 2018).

Taken together our definition of enhancers in stem cells genome wide creates a foundation for constructing detailed GRNs to help understand regenerative mechanisms in planarians.

4. Materials and methods

4.1 Reference assembly and annotations

The sexual genome SMESG.1 genome, the SMESG high confidence annotation (SMESG-high release 1), and the SMESG annotation filtered for repeats (SMESG-repeat release 2) were downloaded from PlanMine (Brandl et al. 2016, Grohme et al. 2018). Available RNA-seq data sets were aligned to the genome with HISAT2 (version 2.1.0) using default parameter settings and providing known splice sites from the SMESG-high annotation (Kim et al. 2015). Transcripts were assembled and merged from the alignments with StringTie using the SMESG-high annotation as a reference (Pertea et al. 2015, 2016).

The new expression-driven annotation (Garcia-Castron et al. 2021, https://jakke-neiro.github.io/Oxplatys/) was compared to the SMESG-high annotation with gffcompare (Pertea et al. 2015, 2016). Transcripts labelled with the class code "=" were classified as full matches corresponding to the transcripts in the SMESG.1 genome, while transcripts labelled with the class codes "c", "k", "m", "n", "j", "e", and "o" were classified as isoforms. Lastly, transcripts labelled with class codes "u" (unknown or intergenic), "i" (fully contained within a reference intron), and "x" (exonic overlap on the opposite strand) were selected as potential candidates for new coding and non-coding transcripts (Pertea et al. 2016, Wu et al. 2016, Azlan et al. 2019). Transdecoder was used to identify open reading frames (ORFs) in these new transcripts, and transcripts longer than 100 amino acids were classified as putative coding transcripts (Haas et al. 2013, Wu et al. 2016, Azlan et al. 2019). Subsequently, InteProScan and BlastX against UniProt were used to look for protein-like structures in the translated ORFs of the remaining transcripts, and transcripts without any hits were retained as putative non-coding transcripts Jones et al. 2014, Wu et al. 2016).

4.2 Expression values, homology and gene ontology

To obtain FACS-specific expression values, the same FACS RNA-seq datasets used by Dattani et al. (2018) were used. The RNA-seq datasets were pseudo-aligned with Salmon using selective alignment, k-mer size 31 and 50 bootstrap iterations (Patro et al. 2017). The abundance estimates were converted to the Kallisto compatible format with wasabi, and the transcript counts were normalized with Sleuth (Pimentel et al. 2017). The TPM values of individual transcripts were summed to calculate TPM values for each gene. The mean TPM value across all samples in a FACS cell compartment was calculated for each gene to get the final absolute TPM value. The proportional expression values ($Xk\%$) were calculated by dividing the compartment specific TPM value by the sum of TPM values of all compartments:

$$Xk\% = \frac{Xk(TPM)}{X1(TPM) + X2(TPM) + Xins(TPM)}, Xk \in \{X1, X2, Xins\}$$

Transcripts were categorized into FACS enrichment categories as follows: X1 when X1% > 0.5, X2 when X2% > 0.5, Xins when Xins% > 0.5, X1 and X2 when X1%+X2% > 0.75 and neither enriched in X1 nor X2, X1 and Xins when X1%+Xins% > 0.75 and neither enriched in X1 nor Xins, X2 and Xins when X2%+Xins% > 0.75 and neither enriched in X2 nor Xins, and ubiquitous when not enriched in any of the above categories (Dattani et al. 2018). Transcript lengths and expression values were compared with Kruskall-Wallis test. Ternary plots and kernel density estimation of the proportional expression values were generated with the package ggtern (Hamilton and Ferry 2018). Information content (IC) was defined to represent as a singular scalar the divergence of a combination of proportional expression values from the even distribution point (X1% = X2% = Xins%):

$$IC = 0.01 * p_1 * \log(p_1) + 0.01 * p_2 * \log(p_2) + 0.01 * p_3 * \log(p_3)$$

where $p_1 = X1\%$, $p_2 = X2\%$, $p_3 = Xins\%$, and $\min(\{p_1, p_2, p_3\}) = 0.01$.

The homology of the transcripts was investigated by using Blastn (Altschul et al. 1990). The non-coding and coding transcripts were aligned to the transcriptomes of humans (GRCh38.p13) and *Dugesia japonica* (assembled in García-Castro et al. (2021)). The threshold for non-coding transcripts was e-value = 10^{-5} and for coding transcripts e-value = 10^{-10}. Gene Ontology (GO) enrichment analysis of new coding transcripts with respect to known coding transcripts was performed using topGO with Fisher's exact test (Alexa and Rahnenfuhrer 2020). The data for proportional expression and homology assigned by BLAST of all transcripts is available at https://jakke-neiro.github.io/Oxplatys/)

4.3 Identification and characterization of transcription factors

We conducted a systematic domain annotation of all transcripts by using the InterProScan resource (Jones et al. 2014). Transcripts with a InterProScan description "transcription factor", a Pfam hit to the Pfam families listed in the transcription factor database DNA-binding domain (DBD) v2.0, or a SUPERFAMILY hit listed in the SUPERFAMILY families in DBD were classified as TFs (Wilson et al. 2008). We used Blastn to align the potential planarian TFs to TFs in humans and fruit flies, and potential TFs without hits were filtered out. The planarian literature was systematically reviewed (Supplementary File 1) and if a TF was mentioned, the GenBank accession number or primer information was retrieved to establish the exact sequence in the literature. This sequence from the literature was aligned to the potential TFs (default parameters for GenBank accession entries and word size 10 for primers) to correctly assign the TFs in the literature to our transcripts (Supplementary File 1). If a planarian TF had previously been described in the literature under a certain name, one name was chosen as the primary name, while all other alternative synonyms used for the same TF were listed as secondary names (Supplementary File 1). If a potential TF had not been described in the planarian literature, the name of the best human or fruit fly Blast hit

was used (Supplementary Table 1). The TFs were categorized into four main groups (Basic domain, Zinc fingers, Helix-turn-helix, Alpha-helical, Immunoglobulin-fold, Beta-hairpin, Beta-sheet, Beta-barrel, Others) according to the TRANSFAC database (Stegmaier et al. 2004) (Supplementary File 1).

Ternary plots, kernel density estimation and the information content of the proportional expression values of TFs were calculated as for other transcripts (4.2 Expression values, homology, and gene ontology). Hierarchical clustering of FACS proportions was performed using Euclidean distance and Ward's method with the package hclust (van den Boogaart and Tolosana-Delgado 2008). The frequencies of different TF domains among the cellular enrichment classes X1, X2 and Xins (for which FACS proportional expression is higher or equal to 50%) were compared with the chi-square test using Yates correction.

4.4 Motifs of transcription factors

The motifs of the TFs (Supplementary File 1) were predicted by using the JASPAR profile inference tool (Fornes et al. 2020). The protein sequences were retrieved with Transdecoder (4.1 Reference assembly and annotations). The information content for each motif was determined by calculating the mean of Shannon's entropy at each position in the position weight matrix (PWM). The motifs were visualized with seqLogo (Bembom 2019).

4.5 ATAC-seq library preparation

A standard protocol was used for preparing the ATAC-seq library (Buenrostro et al. 2013). The X1 cell compartments were isolated by FACS (Romero et al. 2012), and two replicates with 120,000 – 250,000 cells were collected from each compartment. The cells were washed and centrifuged (1 XPBS, 1200 RPM). Lysis buffer (10 mM Tri-Cl (pH7.5), 10 mM NaCl, 3 mM MgCl2, 0.1% NP-40) was added, the cells were centrifuged (500 RPM, 10 min, 4°C), and the nuclei pellet was collected. Then, the transposase mix (25 µl 2X TD Buffer, 2.5 µl Tn5 Transposase, 22.5 µl of nuclease-free water) was added, the cells were resuspended and incubated (37°C, 60 min). Finally, DNA was isolated using the Zymogen Clean & Concentrator Kit, and eluted in EB buffer.

Subsequently, the eluted DNA was used for PCR amplification and library purification. DNA was amplified using a standard reaction (10 µl of purified transposed DNA, 10 µl of nuclease-free water, 15 µl Nextera PCR Master Mix, 5 µl of PCR primer cocktail, 5 µl Index Primer 1, 5 µl Index Primer 2, 72°C 3 mins, 98°C 30 secs, 14 cycles of 98°C 10 secs, 63°C 30 secs, 72°C for 1 min). Finally, the libraries were cleaned with the AMPure bead purification kit. Samples were paired-end sequenced on the Illumina NextSeq.

4.6 ChIP-seq library preparation and sequencing

The H3K27ac ChiP-seq library was prepared and sequenced using established protocols (Mihaylova et al. 2018, Dattani et al. 2018). 600,000−700,000 planarian X1 cells were isolated for each experimental replicate (actual sample and input control). The H3K27ac Abcam ab4729 antibody was used for immunoprecipitation. BAM files of mapped reads are available https://jakke-neiro.github.io/Oxplatys/.

4.7 ChIP-seq data analysis

The ChIP-seq data with respect to H3K4me1 and H3K4me1 after *lpt(RNAi)* was reanalyzed (Mihaylova et al. 2018, PRJNA338116) alongside the H3K27ac data prepared here. The reads were quality-checked with FastQC and trimmed with Trimmomatic (Andrews et al. 2010, Bolger et al. 2014). The reads were aligned to the SMESG.1 genome with Bowtie2 (Langmead et al. 2009). Only uniquely mapped reads and reads with a quality score greater than 10 were retained with Samtools (Li et al. 2009). Peaks were called with MACS2 with default parameters for H3K4me1 and the broad peak parameter for H3K27ac (Zhang et al. 2008). Coverage tracks were generated with deepTools bamCoverage and the mean coverage track was determined with wiggletools (Zerbino et al. 2014, Ramírez et al. 2016). Heatmaps of ChIP-seq profiles were generated with deepTools computeMatrix and plotHeatmap 1,000 bp upstream and downstream of the peak center (Ramírez et al. 2016). The mean peak value was calculated as the mean of the ChIP-seq signal 1000 bp upstream and downstream of the peak center. The H3K27ac peaks that were at most 500 bp from the nearest H3K4me1 peak were selected as putative enhancers. Random genomic regions were generated with bedtools random. The difference of mean peak value at enhancers and random regions was tested with Wilcoxon's test. BAM files of these analyses are available at https://jakke-neiro.github.io/Oxplatys/.

4.8 ATAC-seq data analysis and motif footprinting

The quality of paired-end reads was assessed with FastQC and adapter sequences were removed with Trimmomatic (Andrews et al. 2010, Bolger et al. 2014). The reads were aligned to the SMESG.1 genome with Bowtie2 (Langmead et al. 2009). Only uniquely mapped reads and reads with a quality score greater than 10 were retained with samtools (Li et al. 2009). Duplicates were removed with Samtools (Li et al. 2009). Insert sizes were calculated with Picard CollectInsertSizeMetrics (http://broadinstitute.github.io/picard/).

For coverage track generation and peak calling, samples were downsampled to a standard of ~20M reads using Picard DownsampleSam (http://broadinstitute.github.io/picard/). Deeptools2 bamCoverage was used to generate coverage tracks for each sample in bigwig and bedgraph formats by normalizing with respect to reads million mapped reads (RPKM) (Ramírez et al. 2016). TOBIAS was used to look for footprints in the ATAC-seq by using the motifs predicted by Jaspar

(4.4 Motifs of transcription factors) (Bentsen et al. 2020). BAM files of mapped reads are available at https://jakke-neiro.github.io/Oxplatys/.

4.9 Enhancer targets and gene regulatory networks

The putative enhancers were assigned to the nearest TSS with ChIPSeeker (Yu et al. 2015). The footprints in enhancers were used to create links between TFs and target genes for network construction using a Python script (available at https://jakke-neiro.github.io/Oxplatys/). Networks were visualized with the R package Igraph (https://github.com/igraph/rigraph) while the genomic regions were visualized with Gviz (Hahne and Ivanek 2016).

References

Abnave, P., Aboukhatwa, E., Kosaka, N., Thompson, J., Hill, M. A., & Aboobaker, A. A. (2017). Epithelial-mesenchymal transition transcription factors control pluripotent adult stem cell migration in vivo in planarians. Development, 144(19), 3440-3453.

Aboobaker AA. Planarian stem cells: a simple paradigm for regeneration. Trends in cell biology. 2011 May 1;21(5):304-11.

Aerts S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. InCurrent topics in developmental biology 2012 Jan 1 (Vol. 98, pp. 121-145). Academic Press.

Agata, K., Tasaki, J., Nakajima, E., & Umesono, Y. (2014). Recent identification of an ERK signal gradient governing planarian regeneration. Zoology, 117(3), 161-162.

Albà, M. M. (2017). Zinc-finger domains in metazoans: evolution gone wild. Genome biology, 18(1), 1-3.

Alexa A, Rahnenfuhrer J (2020). topGO: Enrichment Analysis for Gene Ontology. R package version 2.40.0.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

An, Y., Kawaguchi, A., Zhao, C., Toyoda, A., Sharifi-Zarchi, A., Mousavi, S. A., ... & Agata, K. (2018). Draft genome of Dugesia japonica provides insights into conserved regulatory elements of the brain restriction gene nou-darake in planarians. Zoological letters, 4(1), 1-12.

Andersson R, Andersen PR, Valen E, Core LJ, Bornholdt J, Boyd M, Jensen TH, Sandelin A. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nature communications. 2014 Nov 12;5(1):1-0.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar;507(7493):455-61.

Andrews S, Krueger F, Seconds-Pichon A, Biggins F, Wingett S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. 2010.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013 Mar 1;339(6123):1074-7.

Arnold, C., Migueles-Lozano, A., Mann, F. G., Lange, J. J., Seidel, C., & Alvarado, A. S. (2021). Hox genes regulate asexual reproductive behavior and tissue segmentation in adult animals. bioRxiv.

Azlan A, Obeidat SM, Yunus MA, Azzam G. Systematic identification and characterization of Aedes aegypti long noncoding RNAs (lncRNAs). Scientific reports. 2019 Aug 21;9(1):1-9.

Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic acids research. 2015 Jul 1;43(W1):W39-49.

Baldini, A., Fulcoli, F. G., & Illingworth, E. (2017). Tbx1: transcriptional and developmental functions. Current topics in developmental biology, 122, 223-243.

Bembom O (2019). seqLogo: Sequence logos for DNA sequence alignments. R package version 1.52.0.

Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., ... & Looso, M. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nature communications, 11(1), 1-11.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology. 2006;24:11:1429-35.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proceedings of the National Academy of Sciences. 2002 Jan 22;99(2):757-62.

Blassberg RA, Felix DA, Tejada-Romero B, Aboobaker AA. PBX/extradenticle is required to re-establish axial structures and polarity during planarian regeneration. Development. 2013 Feb 15;140(4):730-9.

Blumberg A, Zhao Y, Huang YF, Dukler N, Rice EJ, Krumholz K, Danko CG, Siepel A. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. BioRxiv. 2019 Jan 1:690644.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20.

Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczyński B, Riddell A, Furlong EE. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nature genetics. 2012 Feb;44(2):148.

Brandl H, Moon H, Vila-Farré M, Liu SY, Henry I, Rink JC. PlanMine–a mineable resource of planarian biology and biodiversity. Nucleic acids research. 2016 Jan 4;44(D1):D764-73.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology, 34(5), 525-527.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods. 2013 Dec;10(12):1213.

Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why?. Molecular cell. 2013 Mar 7;49(5):825-37.

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., ... & Raschellà, G. (2017). Zinc-finger proteins in health and disease. Cell death discovery, 3(1), 1-12.

Chen CC, Wang IE, Reddien PW. pbx is required for pole and eye regeneration in planarians. Development. 2013 Feb 15;140(4):719-29.

Cheng LC, Tu KC, Seidel CW, Robb SM, Guo F, Alvarado AS. Cellular, ultrastructural and molecular analyses of epidermal cell development in the planarian Schmidtea mediterranea. Developmental biology. 2018 Jan 15;433(2):357-73.

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics, 33(18), 2938-2940.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics. 2014 Dec;46(12):1311.

Cowles MW, Brown DD, Nisperos SV, Stanley BN, Pearson BJ, Zayas RM. Genome-wide analysis of the bHLH gene family in planarians identifies factors required for adult neurogenesis and neuronal regeneration. Development. 2013 Dec 1;140(23):4691-702.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010 Dec 14;107(50):21931-6.

Currie, K. W., Brown, D. D., Zhu, S., Xu, C., Voisin, V., Bader, G. D., & Pearson, B. J. (2016). HOX gene complement and expression in the planarian Schmidtea mediterranea. EvoDevo, 7(1), 1-11.

Dattani A, Kao D, Mihaylova Y, Abnave P, Hughes S, Lai A, Sahu S, Aboobaker AA. Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells. Genome research. 2018 Oct 1;28(10):1543-54.

Dattani A, Sridhar D, Aboobaker AA. Planarian flatworms as a new model system for understanding the epigenetic regulation of stem cell pluripotency and differentiation. InSeminars in cell & developmental biology 2019 Mar 1 (Vol. 87, pp. 79-94). Academic Press.

Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. Genome research. 2017;27:12:2096-107.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS biology. 2010 May;8(5).

Duncan, E. M., Chitsazan, A. D., Seidel, C. W., & Alvarado, A. S. (2015). Set1 and MLL1/2 target distinct sets of functionally different genomic loci in vivo. Cell reports, 13(12), 2741-2755.

Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proceedings of the National Academy of Sciences. 2017 Jun 20;114(25):E4914-23.

Duren, Z., Chen, X., Xin, J., Wang, Y., & Wong, W. H. (2020). Time course regulatory analysis based on paired expression and chromatin accessibility data. Genome research, 30(4), 622-634.

Emerson, R. O., & Thomas, J. H. (2011). Gypsy and the birth of the SCAN domain. Journal of virology, 85(22), 12043-12052.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May;473(7345):43.

Felix, D. A., & Aboobaker, A. A. (2010). The TALE class homeobox gene Smed-prep defines the anterior compartment for head regeneration. PLoS genetics, 6(4), e1000915.

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian Schmidtea mediterranea. Science, 360(6391).

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., ... & Mathelier, A. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic acids research, 48(D1), D87-D92.

Forsthoefel, D. J., James, N. P., Escobar, D. J., Stary, J. M., Vieira, A. P., Waters, F. A., & Newmark, P. A. (2012). An RNAi screen reveals intestinal regulators of branching morphogenesis, differentiation, and stem cell proliferation in planarians. Developmental cell, 23(4), 691-704.

Forsthoefel DJ, Cejda NI, Khan UW, Newmark PA. Cell-type diversity and regionalized gene expression in the planarian intestine. Elife. 2020 Apr 2;9:e52613.

Fraguas, S., Cárcel, S., Vivancos, C., Molina, M. D., Ginés, J., Mazariegos, J., ... & Cebrià, F. (2021). CREB-binding protein (CBP) gene family regulates planarian survival and stem cell differentiation. Developmental Biology, 476, 53-67.

Fujimura, N., Vacik, T., Machon, O., Vlcek, C., Scalabrin, S., Speth, M., ... & Kozmik, Z. (2007). Wnt-mediated down-regulation of Sp1 target genes by a transcriptional repressor Sp5. Journal of Biological Chemistry, 282(2), 1225-1237.

Gaiti F, Jindrich K, Fernandez-Valverde SL, Roper KE, Degnan BM, Tanurdžić M. Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. Elife. 2017 Apr 11;6:e22194.

Gaiti, F., Calcino, A. D., Tanurdžić, M., & Degnan, B. M. (2017). Origin and evolution of the metazoan non-coding regulatory genome. Developmental biology, 427(2), 193-202.

Gao, B., Wang, Y., Diaby, M., Zong, W., Shen, D., Wang, S., ... & Song, C. (2020). Evolution of pogo, a separate superfamily of IS630-Tc1-mariner transposons, revealing recurrent domestication events in vertebrates. Mobile DNA, 11(1), 1-15.

Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research. 2020;48:D1:D58-D64; doi: 10.1093/nar/gkz980

García-Castro H, Kenny NJ, Álvarez-Campos P, Mason V, Schönauer A, Sleight VA, Neiro J, Aboobaker A, Permanyer J, Iglesias M, Irimia M. ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. Genome biology, 22(1), 1-34.

Gaspar J. ATAC-seq Guidelines. Harvard FAS Informatics. 2019.

Gehrke AR, Neverett E, Luo YJ, Brandt A, Ricci L, Hulett RE, Gompers A, Ruby JG, Rokhsar DS, Reddien PW, Srivastava M. Acoel genome reveals the regulatory landscape of whole-body regeneration. Science. 2019;363:6432:eaau6173.

Goldman JA, Kuzu G, Lee N, Karasik J, Gemberling M, Foglia MJ, Karra R, Dickson AL, Sun F, Tolstorukov MY, Poss KD. Resolving heart regeneration by replacement histone profiling. Developmental cell. 2017;40:4:392-404.

Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, Beer MA, Pavan WJ, McCallion AS. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. Genome research. 2012 Nov 1;22(11):2290-301.

Grant CE, Johnson J, Bailey TL, Noble WS. MCAST: scanning for cis-regulatory motif clusters. Bioinformatics. 2016 Apr 15;32(8):1217-9.

Grohme MA, Schloissnig S, Rozanski A, Pippel M, Young GR, Winkler S, Brandl H, Henry I, Dahl A, Powell S, Hiller M. The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. Nature. 2018 Feb;554(7690):56-61.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols. 2013 Aug;8(8):1494.

Hahne F, Ivanek R (2016). “Statistical Genomics: Methods and Protocols.” In Mathé E, Davis S (eds.), chapter Visualizing Genomic Data Using Gviz and Bioconductor, 335–351. Springer New York, New York, NY. ISBN 978-1-4939-3578-9, doi: 10.1007/978-1-4939-3578-9_16, http://dx.doi.org/10.1007/978-1-4939-3578-9_16.

Hayashi T, Asami M, Higuchi S, Shibata N, Agata K. Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. Development, growth & differentiation. 2006 Aug;48(6):371-80.

Hamilton NE, Ferry M. ggtern: Ternary diagrams using ggplot2. Journal of Statistical Software. 2018 Dec 1;87(1):1-7.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics. 2007 Mar;39(3):311-8.

Hetzel J, Duttke SH, Benner C, Chory J. Nascent RNA sequencing reveals distinct features in plant transcription. Proceedings of the National Academy of Sciences. 2016 Oct 25;113(43):12316-21.

Hoang, V. T., Nyswaner, K., Torres-Ayuso, P., & Brognard, J. (2020). The protein kinase MAP3K19 phosphorylates MAP2Ks and thereby activates ERK and JNK kinases and increases viability of KRAS-mutant lung cancer cells. Journal of Biological Chemistry, 295(25), 8470-8479.

Hoff KJ, Stanke M. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic acids research. 2013 Jul 1;41(W1):W123-8.

Hoff KJ, Stanke M. Predicting genes in single genomes with augustus. Current protocols in bioinformatics. 2019 Mar;65(1):e57.

Hopitzan, A. A., Baines, A. J., Ludosky, M. A., Recouvreur, M., & Kordeli, E. (2005). Ankyrin-G in skeletal muscle: tissue-specific alternative splicing contributes to the complexity of the sarcolemmal cytoskeleton. Experimental cell research, 309(1), 86-98.

Ibrahim MM, Karabacak A, Glahs A, Kolundzic E, Hirsekorn A, Carda A, Tursun B, Zinzen RP, Lacadie SA, Ohler U. Determinants of promoter and enhancer transcription directionality in metazoans. Nature communications. 2018 Oct 26;9(1):1-5.

Janssens, J., Aibar, S., Taskiran, I. I., Ismail, J. N., Gomez, A. E., Aughey, G., ... & Aerts, S. (2022). Decoding gene regulation in the fly brain. Nature, 1-7.

Jegla, T., Nguyen, M. M., Feng, C., Goetschius, D. J., Luna, E., van Rossum, D. B., ... & Rolls, M. M. (2016). Bilaterian giant ankyrins have a common evolutionary origin and play a conserved role in patterning the axon initial segment. PLoS genetics, 12(12), e1006457.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236-40.

Jänes J, Dong Y, Schoof M, Serizay J, Appert A, Cerrato C, Woodbury C, Chen R, Gemma C, Huang N, Kissiov D. Chromatin accessibility dynamics across C. elegans development and ageing. Elife. 2018 Oct 26;7:e37344.

Kang J, Hu J, Karra R, Dickson AL, Tornini VA, Nachtrab G, Gemberling M, Goldman JA, Black BL, Poss KD. Modulation of tissue repair by regeneration enhancer elements. Nature. 2016;532:7598:201-6.

Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, Ninomiya-Fukuda N. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. Genome research. 2014 Apr 1;24(4):708-17.

Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G, Baranasic D. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research. 2018 Jan 4;46(D1):D260-6.

Khan, S. F., Damerell, V., Omar, R., Du Toit, M., Khan, M., Maranyane, H. M., ... & Prince, S. (2020). The roles and regulation of TBX3 in development and disease. Gene, 726, 144223.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010 May;465(7295):182-7.

Kim TK, Hemberg M, Gray JM. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. Cold Spring Harbor perspectives in biology. 2015 Jan 1;7(1):a018622.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature methods. 2015 Apr;12(4):357-60.

Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics. 2019 Apr;20(4):207-20.

Koenecke N, Johnston J, He Q, Meier S, Zeitlinger J. Drosophila poised enhancers are generated during tissue patterning with the help of repression. Genome research. 2017 Jan 1;27(1):64-74.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, Kolpakov FA. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic acids research. 2018 Jan 4;46(D1):D252-9.

Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DD, Ross EJ, Voisin V, Bader GD, Blencowe BJ, Pearson BJ. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. Stem cells. 2012 Aug;30(8):1734-45.

Lai AG, Aboobaker AA. EvoRegen in animals: time to uncover deep conservation or convergence of adult stem cell evolution and regenerative processes. Developmental biology. 2018 Jan 15;433(2):118-31.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009 Mar;10(3):R25.

Lavoie, H., Gagnon, J., & Therrien, M. (2020). ERK signalling: A master regulator of cell behaviour, life and fate. Nature Reviews Molecular Cell Biology, 21(10), 607-632.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nature Reviews Genetics. 2016 Apr;17(4):207.

Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome biology. 2019 Dec 1;20(1):45.

Lopez, A. Y., Wang, X., Xu, M., Maheshwari, A., Curry, D., Lam, S., ... & Cooper, E. C. (2017). Ankyrin-G isoform imbalance and interneuronopathy link epilepsy and bipolar disorder. *Molecular psychiatry*, *22*(10), 1464-1472.

Lowe, E. K., Cuomo, C., Voronov, D., & Arnone, M. I. (2019). Using ATAC-seq and RNA-seq to increase resolution in GRN connectivity. Methods in cell biology, 151, 115-126.

Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic acids research. 2007 Jul 1;35(suppl_2):W253-8.

McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC bioinformatics. 2010 Dec 1;11(1):165.

Meers MP, Adelman K, Duronio RJ, Strahl BD, McKay DJ, Matera AG. Transcription start site profiling uncovers divergent transcription and enhancer-associated RNAs in Drosophila melanogaster. BMC genomics. 2018 Dec;19(1):157.

Mihaylova Y, Abnave P, Kao D, Hughes S, Lai A, Jaber-Hijazi F, Kosaka N, Aboobaker AA. Conservation of epigenetic regulation by the MLL3/4 tumour suppressor in planarian pluripotent stem cells. Nature communications. 2018 Sep 7;9(1):1-7.

Molina, M. D., Saló, E., & Cebrià, F. (2007). The BMP pathway is essential for re-specification and maintenance of the dorsoventral axis in regenerating and intact planarians. Developmental biology, 311(1), 79-94.

Murad R, Macias-Munoz A, Wong A, Ma X, Mortazavi A. Integrative analysis of Hydra head regeneration reveals activation of distal enhancer-like elements. bioRxiv. 2019 Jan 1:544049.

Najafabadi, H. S., Garton, M., Weirauch, M. T., Mnaimneh, S., Yang, A., Kim, P. M., & Hughes, T. R. (2017). Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. Genome biology, 18(1), 1-15.

NCBI Resource Coordinators. Database resources of the national center for biotechnology information. Nucleic acids research. 2015 Jan 28;43(Database issue):D6.

Nelson, A. D., & Jenkins, P. M. (2017). Axonal membranes and their domains: assembly and function of the axon initial segment and node of Ranvier. Frontiers in cellular neuroscience, 11, 136.

Nettling M, Treutler H, Grau J, Keilwagen J, Posch S, Grosse I. DiffLogo: a comparative visualization of sequence motifs. BMC bioinformatics. 2015 Dec;16(1):387.

Önal P, Grün D, Adamidi C, Rybak A, Solana J, Mastrobuoni G, Wang Y, Rahn HP, Chen W, Kempa S, Ziebold U. Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. The EMBO journal. 2012 Jun 13;31(12):2755-69.

Owlarn, S., Klenner, F., Schmidt, D., Rabert, F., Tomasso, A., Reuter, H., ... & Bartscherer, K. (2017). Generic wound signals initiate regeneration in missing-tissue contexts. Nature communications, 8(1), 1-13.

Pachkov M, Balwierz PJ, Arnold P, Ozonov E, Van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic acids research. 2012 Nov 23;41(D1):D214-20.

Pascual-Carreras, E., Marin-Barba, M., Castillo-Lara, S., Coronel-Cordoba, P., Magri, M. S., Wheeler, G. N., ... & Adell, T. (2020). Genomic analyses reveal FoxG as an upstream regulator of wnt1 required for posterior identity specification in planarians. *bioRxiv*.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature methods. 2017 Apr;14(4):417.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015 Mar;33(3):290.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016 Sep;11(9):1650.

Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nature methods. 2017 Jul;14(7):687.

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., ... & Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science, 360(6391).

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., ... & Adey, A. C. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Molecular cell, 71(5), 858-871.

Pundhir S, Bagger FO, Lauridsen FB, Rapin N, Porse BT. Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality. Nucleic acids research. 2016 May 19;44(9):4037-51.

Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. Current protocols in bioinformatics, 47(1), 11-12.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011 Feb;470(7333):279-83.

Rahman S, Zorca CE, Traboulsi T, Noutahi E, Krause MR, Mader S, Zenklusen D. Single-cell profiling reveals that eRNA accumulation at enhancer–promoter loops is not required to sustain transcription. Nucleic acids research. 2017 Apr 7;45(6):3017-30.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research. 2016 Apr 13;44(W1):W160-5.

Raz, A. A., Wurtzel, O., & Reddien, P. W. (2021). Planarian stem cells specify fate yet retain potency during the cell cycle. Cell Stem Cell.

Reddien, P. W., Bermange, A. L., Kicza, A. M., & Sánchez Alvarado, A. (2007). BMP signaling regulates the dorsal planarian midline and is needed for asymmetric regeneration.

Reddien PW. The cellular and molecular basis for planarian regeneration. Cell. 2018 Oct 4;175(2):327-45.

Romero BT, Evans DJ, Aboobaker AA. FACS analysis of the planarian stem cell compartment as a tool to understand regenerative mechanisms. InProgenitor Cells 2012 (pp. 167-179). Humana Press, Totowa, NJ.

Sandelin A, Wasserman WW. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. Journal of molecular biology. 2004 Apr 23;338(2):207-15.

Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, Furlong EE. A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes & development. 2007 Feb 15;21(4):436-49.

Scimone, M. L., Kravarik, K. M., Lapan, S. W., & Reddien, P. W. (2014). Neoblast specialization in regeneration of the planarian Schmidtea mediterranea. Stem cell reports, 3(2), 339-352.

Scimone, M. L., Cote, L. E., & Reddien, P. W. (2017). Orthogonal muscle fibres have different instructive roles in planarian regeneration. Nature, 551(7682), 623-628.

Schmitz JF, Zimmer F, Bornberg-Bauer E. Mechanisms of transcription factor evolution in Metazoa. Nucleic acids research. 2016 Jul 27;44(13):6287-97.

Schoenfelder S, Sugar R, Dimond A, Javierre BM, Armstrong H, Mifsud B, Dimitrova E, Matheson L, Tavares-Cadete F, Furlan-Magaril M, Segonds-Pichon A. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nature genetics. 2015;47:10:1179.

Schwaiger M, Schönauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, Schinko JB, Renfer E, Fredman D, Technau U. Evolutionary conservation of the eumetazoan gene regulatory landscape. Genome research. 2014;24:4:639-50.

Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J. J., Sabidó, E., ... & Ruiz-Trillo, I. (2016). The dynamic regulatory genome of Capsaspora and the origin of animal multicellularity. Cell, 165(5), 1224-1237.

Shannon P, Richards M (2019). MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. R package version 1.28.0.

Siebert S, Farrell JA, Cazet JF, Abeykoon Y, Primack AS, Schnitzler CE, Juliano CE. Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. Science. 2019 Jul 26;365(6451):eaav9314.

Solana J, Kao D, Mihaylova Y, Jaber-Hijazi F, Malla S, Wilson R, Aboobaker A. Defining the molecular profile of planarian pluripotent stem cells using a combinatorial RNA-seq, RNA interference and irradiation approach. Genome biology. 2012 Mar;13(3):R19.

Spicuglia S, Vanhille L. Chromatin signatures of active enhancers. Nucleus. 2012 Mar 1;3(2):126-31.

Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nature reviews genetics. 2012 Sep;13(9):613-26.

Stegmaier P, Kel AE, Wingender E. Systematic DNA-binding domain classification of transcription factors. Genome Informatics. 2004;15(2):276-86.

Stelman, C. R., Smith, B. M., Chandra, B., & Roberts-Galbraith, R. H. (2021). CBP/p300 homologs CBP2 and CBP3 play distinct roles in planarian stem cell function. Developmental Biology, 473, 130-143.

Sun, Y., Liu, W. Z., Liu, T., Feng, X., Yang, N., & Zhou, H. F. (2015). Signaling pathway of MAPK/ERK in cell proliferation, differentiation, migration, senescence and apoptosis. Journal of Receptors and Signal Transduction, 35(6), 600-604.

Swapna LS, Molinaro AM, Lindsay-Mosher N, Pearson BJ, Parkinson J. Comparative transcriptomic analyses and single-cell RNA sequencing of the freshwater planarian Schmidtea mediterranea identify major cell types and pathway conservation. Genome biology. 2018 Dec;19(1):1-22.

Tang, Z., Wu, Y., Yang, Y., Yang, Y. C. T., Wang, Z., Yuan, J., ... & Li, K. (2017). Comprehensive analysis of long non-coding RNAs highlights their spatio-temporal expression patterns and evolutional conservation in Sus scrofa. Scientific Reports, 7(1), 1-12.

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell, 126(4), 663-676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell, 131(5), 861-872.

Tasaki, J., Shibata, N., Nishimura, O., Itomi, K., Tabata, Y., Son, F., ... & Umesono, Y. (2011). ERK signaling controls blastema cell differentiation during planarian regeneration. Development, 138(12), 2417-2427.

Tasaki, J., Shibata, N., Sakurai, T., Agata, K., & Umesono, Y. (2011). Role of c-Jun N-terminal kinase activation in blastema formation during planarian regeneration. Development, growth & differentiation, 53(3), 389-400.

Tewari, A. G., Owen, J. H., Petersen, C. P., Wagner, D. E., & Reddien, P. W. (2019). A small set of conserved genes, including sp5 and Hox, are activated by Wnt signaling in the posterior of planarians and acoels. PLoS genetics, 15(10), e1008401.

Thieffry A, Vigh ML, Bornholdt J, Ivanov M, Brodersen P, Sandelin A. Characterization of Arabidopsis thaliana Promoter Bidirectionality and Antisense RNAs by Inactivation of Nuclear RNA Decay Pathways. The Plant Cell. 2020 Jun 1;32(6):1845-67.

Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, Biggin MD. Dynamic reprogramming of chromatin accessibility during Drosophilaembryo development. Genome biology. 2011 May 1;12(5):R43.

Tomoyasu Y, Halfon MS. How to study enhancers in non-traditional insect models. Journal of Experimental Biology. 2020 Feb 1;223(Suppl 1).

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ. Assessing computational tools for the discovery of transcription factor binding sites. Nature biotechnology. 2005 Jan;23(1):137-44.

van den Boogaart KG, Tolosana-Delgado R. “Compositions”: a unified R package to analyze compositional data. Computers & Geosciences. 2008 Apr 1;34(4):320-38.

van Wolfswinkel, J. C., Wagner, D. E., & Reddien, P. W. (2014). Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. Cell stem cell, 15(3), 326-339.

Vásquez-Doorman, C., & Petersen, C. P. (2014). zic-1 Expression in Planarian neoblasts after injury controls anterior pole regeneration. PLoS genetics, 10(7), e1004452.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457:7231:854-8.

Vogg, M. C., Owlarn, S., Rico, Y. A. P., Xie, J., Suzuki, Y., Gentile, L., ... & Bartscherer, K. (2014). Stem cell-dependent formation of a functional anterior regeneration pole in planarians requires Zic and Forkhead transcription factors. Developmental Biology, 390(2), 136-148.

Wang, W., Hu, C. K., Zeng, A., Alegre, D., Hu, D., Gotting, K., ... & Alvarado, A. S. (2020). Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates. *Science*, *369*(6508).

Wang, L. H., Aberin, M. A. E., Wu, S., & Wang, S. P. (2021). The MLL3/4 H3K4 methyltransferase complex in establishing an active enhancer landscape. Biochemical Society Transactions.

Warman E, Forrest D, Wade JT, Grainger DC. Widespread divergent transcription from prokaryotic promoters. BioRxiv. 2020 Jan 1.

Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. Journal of molecular biology. 1998 Apr 24;278(1):167-81.

Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. Package 'corrplot'. Statistician. 2017 Oct 16;56:316-24.

Weidinger, G., Thorpe, C. J., Wuennenberg-Stapleton, K., Ngai, J., & Moon, R. T. (2005). The Sp1-related transcription factors sp5 and sp5-like act downstream of Wnt/β-catenin signaling in mesoderm and neuroectoderm patterning. Current biology, 15(6), 489-500.

West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ, Tolstorukov MY, Kingston RE. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. Nature communications. 2014 Aug 27;5(1):1-2.

Wilson, D., Charoensawan, V., Kummerfeld, S. K., & Teichmann, S. A. (2008). DBD—taxonomically broad transcription factor predictions: new content and functionality. Nucleic acids research, 36(suppl_1), D88-D92.

Wu Y, Cheng T, Liu C, Liu D, Zhang Q, Long R, Zhao P, Xia Q. Systematic identification and characterization of long non-coding RNAs in the silkworm, Bombyx mori. PloS one. 2016;11(1).

Yáñez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. Genome research. 2012 Oct 1;22(10):2018-30.

Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. Trends in Genetics. 2013 Jan 1;29(1):11-22.

Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome biology. 2020 Dec 1;21(1):22.

Yang K, Kang J. Tissue regeneration enhancer elements: a way to unlock endogenous healing power. Developmental Dynamics. 2019 Jan;248(1):34-42.

Yin S, Lu K, Tan T, Tang J, Wei J, Liu X, Hu X, Wan H, Huang W, Fan Y, Xie D. Transcriptomic and open chromatin atlas of high-resolution anatomical regions in the rhesus macaque brain. Nature Communications. 2020 Jan 24;11(1):1-3.

Young RS, Kumar Y, Bickmore WA, Taylor MS. Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. Genome biology. 2017 Dec;18(1):242.

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics, 31(14), 2382-2383.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017 Jan;8(1):28-36.

Zambelli F, Pesole G, Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Briefings in bioinformatics. 2013 Mar 1;14(2):225-37.

Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M. Whole-genome ChIP–chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes & development. 2007 Feb 15;21(4):385-90.

Zerbino, D. R., Johnson, N., Juettemann, T., Wilder, S. P., & Flicek, P. (2014). WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. Bioinformatics, 30(7), 1008-1009.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008 Nov;9(9):R137.

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green, M. R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC bioinformatics, 11(1), 237.

Zhu SJ, Pearson BJ. (Neo) blast from the past: new insights into planarian stem cell lineages. Current opinion in genetics & development. 2016 Oct 1;40:74-80.

Zhu B, Zhang W, Zhang T, Liu B, Jiang J. Genome-wide prediction and validation of intergenic enhancers in Arabidopsis using open chromatin signatures. The Plant Cell. 2015 Sep 1;27(9):2415-26.

Supplementary material

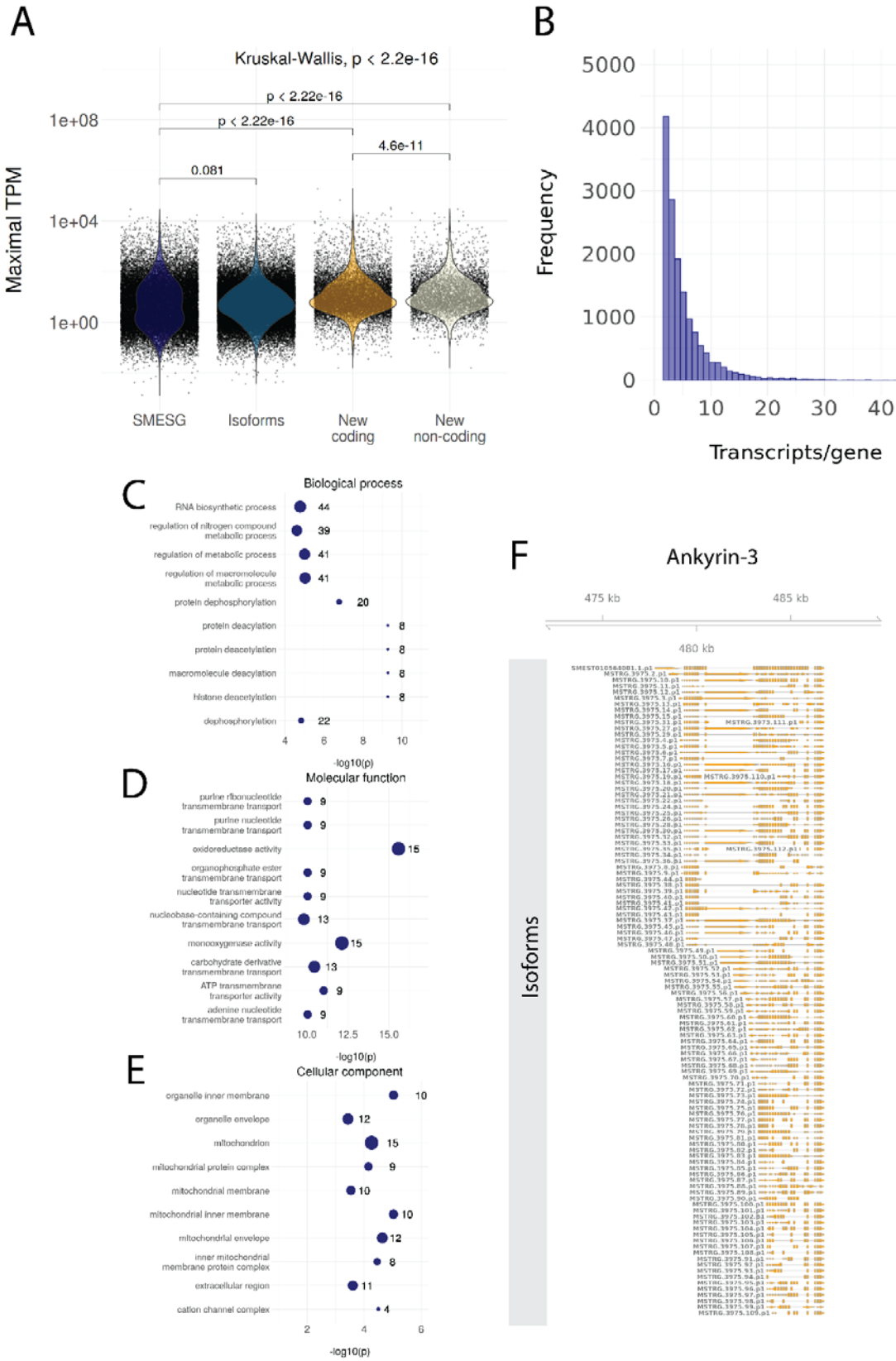

Fig. S1. A. Maximal TPM of known transcripts (SMESG), isoforms, new coding transcripts, and new non-coding transcripts. Maximal TPM is defined as the highest TPM value in the 183 RNA-seq samples used for reannotation. **B.** The number of transcripts per gene (distribution up to 50 transcripts/gene). **C.** The GO terms for biological processes enriched in the new coding transcripts. **D.** The GO terms for molecular functions enriched in the new coding transcripts. **E.** The GO terms for cellular components enriched in the new coding transcripts. **F.** All isoforms for ankyrin-3 (ankyrin-G).

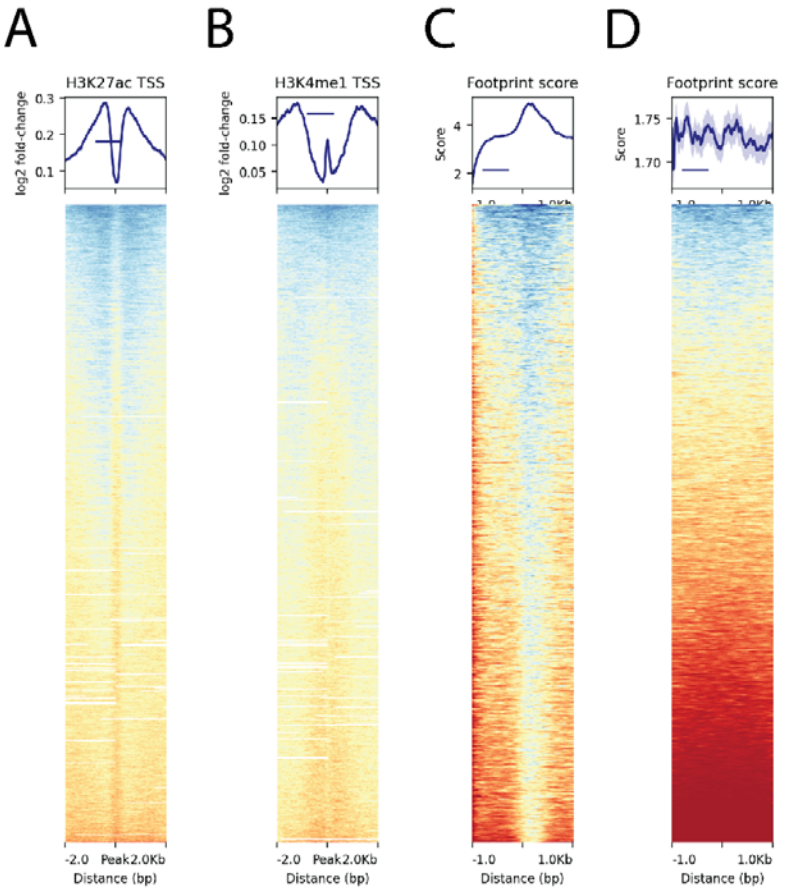
A
B
C
D
H3K27ac TSS
log2 fold-change
0.3
0.2
0.1
-2.0
Peak
2.0Kb
Distance (bp)
H3K4me1 TSS
log2 fold-change
0.15
0.10
0.05
-2.0
Peak
2.0Kb
Distance (bp)
Footprint score
Score
4
2
-1.0
1.0Kb
Distance (bp)
Footprint score
Score
1.75
1.70
-1.0
1.0Kb
Distance (bp)

Fig. S2 A. H3K27ac ChIP-seq signal around transcription start sites (TSS) **B.** H3K4me1 ChIP-seq signal around TSS. **C.** The ATAC-seq footprint score at the combined peaks or putative enhancer-like regions. **D.** The ATAC-seq footprint score at 10,000 random regions in the genome.