Epigenetic analyses of the planarian genome reveals conservation of bivalent promoters in animal stem cells.

Damian Kao, Yuliana Mihaylova, Samantha Hughes, Alvina Lai and A. Aziz Aboobaker

Department of Zoology, Tinbergen Building, South Parks Road, University of Oxford, Oxford OX1 3PS

Damian Kao, Damian.Kao@zoo.ox.ac.uk
Yuliana Mihaylova, Yuliana.Mihaylova@zoo.ox.ac.uk
Samantha Hughes, Samantha.Hughes@han.nl
Alvina Lai, Alvina.Lai@zoo.ox.ac.uk
Aziz Aboobaker, Aziz.Aboobaker@zoo.ox.ac.uk

Abstract

Background

Planarian flatworms have an indefinite capacity to regenerate due to a population of pluripotent adult stem cells (neoblasts). Previous studies have suggested that they have features in common with both pluripotent mammalian embryonic stem cells and germ line stem cells. However, little is known about the importance of epigenetic regulation in these cells, which is likely to be crucial for neoblast biology and regeneration. We set out to develop analytical and experimental tools for planarians to allow the study of epigenetic marks in neoblasts and allow direct comparison of this model system with other animals.

Results

We developed an optimized ChIP-seq protocol for planarian neoblasts that allowed us to generate genome wide profiles for H3K4me1, H3K4me3 and H3K27me3. These were found to correlate as expected with genome wide expression profiles from analyses of planarian RNA-seq data. We found that many genes that are silent in neoblasts and then switch in post-mitotic progeny during differentiation have both H3K4me3 and H3K27me3 at promoter regions and are therefore bivalent. Further analysis suggested that bivalency is present at hundreds of loci in the pluripotent neoblast population.

Conclusions

We confirm that epigenetic regulation is key to neoblast biology and that bivalent promoters are not confined to vertebrate lineages, but may be a conserved feature of animal stem cells. Our work further establishes planarian neoblasts as a powerful model system for understanding the epigenetic regulation of pluripotency and regeneration.

Background

The potential use of stem cells in regenerative medicine has driven research into exploring the molecular mechanisms that govern stem cell potency, maintenance and differentiation. Despite this, we clearly still need to better understand the fundamental regulatory processes underpinning stem cell function, preferably using *in vivo* model systems. Highly regenerative planarians can be considered as a relatively simple stem cell study system that offers a large pool of adult stem cells called neoblasts (NBs). These cells are the driving force behind the almost limitless capacity to regenerate [1-3]. With a simple bilaterian anatomy, the ability to study gene function using RNAi [4] and a growing list of well-defined markers [2,5-12], planarians make for a powerful model system for studying stem cell processes. There is growing evidence for the deep conservation of molecular regulation in stem cells across metazoans, particularly from work with planarians [10,13-19]. This suggests research on planarian NBs can also lead to new insights relevant to mammals.

Independent studies from multiple groups have characterized the transcriptome of NBs, using both bulk approaches [10,13,15,20,21] and single cell sampling and sequencing approaches [22-24]. The characterization of the transcriptome of NBs has relied on approaches that facilitate separation of NBs from the rest of the cells in the body or the targeted removal of NBs. This is followed by differential expression analysis to identify genes that have enriched NB expression. The most widely used of these approaches has involved the development of protocols for Fluorescence Activated Cell Sorting (FACS) that sort dissociated cells on the basis of nuclear to cytoplasmic ratio [25-27]. This approach results in three distinct cell populations, an immediately radiosensitive population of >2N cells representing dividing cells in S/-phaseG2/M NBs (called the X1 compartment), a population of ≤2N cells that is partially radiosensitive and, over a longer time frame, containing G1 NBs and also transient undifferentiated post-mitotic progeny (called the X2 compartment), and a radio-resistant population of post-mitotic differentiated cells (called the Xins compartment) (see Figure 1A for summary). By sequencing these cell populations in bulk [9,10,13,21,24,28], performing single cell multiplex PCR analysis [24] or single cell sequencing [22,23] many groups have contributed to characterizing the NB transcriptome. These studies have identified novel NB markers, described gene expression heterogeneity in the NB population and found other cell type specific markers.

Another approach has been to compare expression profiles of whole animals with and without NBs, where NBs have been removed by ionizing radiation to kill all cycling cells [20] or by RNAi of a gene [15] to rapidly deplete NBs. These two approaches, combined with subsequent functional analyses, have allowed the development of more detailed models of the dynamics of NB proliferation, self-renewal and differentiation during regeneration and homeostasis [7,9,29-36]. One important feature that these studies highlight are chromatin-modifying factors involved in epigenetic regulation, the expression of which is enriched in planarian NBs [10,13,15].

A regulatory mechanism of key importance that mammalian embryonic stem cells (ESCs) and germ line stem cells (GSCs) share is histone modification-based bivalent control of promoters [37-42]. Bivalent promoters are characterized by the simultaneous presence of the repressive mark H3K27me3 and the activation mark H3K4me3 around transcriptional start sites (TSS) [37,38]. This bivalent promoter configuration is commonly seen on genes 'poised' for activation upon stem cell commitment and differentiation. In ESCs this is thought to allow pluripotency and the capacity to sensitively respond to developmental signals to achieve rapid differentiation when required. Bivalent promoters may achieve this by suppressing the formation of active RNA polymerase II complexes on one hand (hence 'poised'), and on the other, not allow other less easily reversible suppressive regulatory mechanisms, like DNA methylation, to silence genes [40,41]. Bivalent promoters have been described in mammals [37-39,42] and in zebrafish [43]. However, genome wide examples of promoter bivalency have so far not been found in invertebrates, for example no evidence for widespread bivalency has been observed in early *Drosophila* embryos [44]. These data support the interpretation that bivalency may be vertebrate specific, however more invertebrates clearly need to be studied and "it remains unclear how universal bivalent domains are across species" [43]. Given the role bivalency is thought to play in regulating pluripotency, planarian NBs are a logical place to look for bivalency in invertebrates.

To ask whether bivalency is present in NBs, we needed to combine transcriptomic and epigenetic analyses in the context of the genome. We first identified expressed loci on the genome and annotated transcribed regions using all available planarian RNA-seq data. We define the proportion of expression of every locus in X1, X2 and Xins FACS sorted cell populations allowing us to robustly identify genes silenced or expressed at very low levels in NB that are then actively transcribed during differentiation. We next

developed a robust ChIP-seq protocol for use with X1 NBs and demonstrate clear correlation of conserved epigenetic marks and gene expression based on the distribution of Histone 3 Lysine 4 mono-methylation (H3K4me1), tri-methylation (H3K4me3) and Histone 3 Lysine 27 tri-methylation (H3K27me3) marks. This revealed many bivalent promoters containing both H3K4me3 and H3K27me3 at similar levels, particularly at loci that go on to greatly increase their expression in post-mitotic progeny after asymmetric NB division. This provides strong evidence that this method of epigenetic regulation may in fact be conserved in animal stem cells. Overall, our work provides an essential annotation framework to study coding and non-coding loci in the genome, establishes a robust approach for ChIP-seq in NBs of *S. mediterranea* and reveals the potential for broad conservation of bivalent promoters in animal stem cells.

Results

Establishing a genome wide annotation of transcribed loci in *S. mediterranea*

A growing number of different transcriptome studies have characterized the RNA-seq profiles of NBs and other planarian cells by sequencing either FACS-sorted cell populations (Figure 1A) or animals depleted of stem cells [9,10,13,15,20,21], but these have not been integrated. We performed a comparison of the results of four published transcriptomic studies, and found very poor overlap between those genes defined as having enriched expression in NBs (Figure 1B, Additional File 1).

To address this, we used the rapidly growing collection of publicly available *S. mediterranea* transcriptome sequence data to define a set of genome annotations on the current assembly of the asexual strain to serve as a basis for comparing transcriptional and epigenetic. This annotation then served as a basis for regulation across different planarian research projects. Distinct from previous annotations of the genome [45,46], this annotation includes all transcribed elements present across all available RNA-seq datasets (Additional file 2). As we have integrated all available RNA-seq data, our annotation should be particularly useful for describing potential non-coding RNAs and protein coding genes expressed at low levels, as these may not have been discovered by individual studies with limited numbers of reads and/or reliant on homology of protein coding exons.

Our annotation is markedly different from the current available annotation [46] of the *S. mediterranea* asexual genome sequence. We annotated more than 11,000 potential new

protein coding loci that are expressed at similar overall levels to previously annotated protein coding genes that were also present in our annotation. These new annotations were enriched for less well conserved proteins that may not be predicted by homology based annotation. A total of 6,300 existing annotations were not present in our expression driven annotation and further analysis of these MAKER specific genes shows that they generally have no or very low potential expression within the 164 RNA-seq libraries used for our annotations. (Additional File 2)

In summary, our annotation on the current planarian genome assembly shows regions of active transcription detected by current RNA-seq and transcriptome data, defines many more protein coding regions than currently available annotations and highlights a large number of non-coding transcribed loci. Additionally, it facilitates a consistent comparison specifically between *bona fide* transcriptional activity and the presence of post-translational histone modifications (ChIP-seq), allowing the relationship between epigenetic regulation and gene expression to be studied.

A genome wide expression profile of FACS-sorted cell populations.

The ability to use FACS is a powerful experimental tool for working with *S. mediterranea*, providing convenient access to NBs and other cell populations. A growing number of studies have produced RNA-seq data for the different FACS cell populations that can be differentiated by nuclear to cytoplasmic ratio [9,10,13,21,24]. Given the discrepancies we uncovered between different studies that have taken this approach (Figure 1B, Additional File 1), we decided to reanalyze these datasets and newly available FACS RNA-seq data in public databases. Hierarchical clustering of the normalized expression values of each of these libraries revealed a rough congruence between different FACS cell populations (Additional File 3), and revealed greater heterogeneity among the X1 samples than within the X2 and Xins samples. For example, some X1 samples clustered with X2 samples (Additional File 3).

These inconsistencies are potentially biases introduced by variation in the underlying biological or technical conditions. To mitigate against technical differences affecting absolute expression values we transformed absolute expression values into proportional expression values for each FACs compartment (Figure 1A, Additional File 4). For each locus, we divided each of the three X1, X2, Xins expression values by the sum of expression of all compartments for that locus (Additional File 4), obtaining a proportional

expression value for a total of 27,206 annotated loci that had at least 10 reads mapped in at least one FACS RNA-seq library, 18,010 of these are likely to be protein coding. The advantage of this transformation is that instead of using independent absolute expression values of the various samples, we can use the relationship among the three cell populations in each sample. Given similar FACS gating settings, results should be more consistent between labs despite any technical variations that affect absolute expression levels. Hierarchical clustering of these proportional values showed a consistent clustering of FACS samples by cell type, with good separation between clusters (Additional File 4). We then combined all available FACS RNA-seq data to reach one set of proportional expression values for each locus in our annotation. This gave us a new robust expression metric to compare every transcriptional unit in the genome across FACS cell compartments.

In order to achieve a visual representation of the data, we simply used a line to represent each gene, colored according to the proportion of its total expression in each of the three cell compartments. This allowed us to create genome wide expression spectra as an intuitive visualization and analysis tool. For example, we can sort all genes according to the proportion of their expression in X1 (S/G2/M NBs) (Figure 2A), X2 (G1 NBs and stem cell progeny) (Figure 2B) and Xins (post-mitotic differentiated cells) (Figure 2C).

We used this approach to define which genes were expressed in each FACs compartment, dividing all genes expressed in FACS RNA-seq data into classes of enrichment (Figure 2D, Additional File 5). We confirmed our analysis by checking for the enrichment classes of genes previously described as being expressed in NBs (X1 and X1/X2 classes), in stem cell progeny (X2 and X2/Xins classes) and in differentiated cells (Xins class, Figure 3A-I, Additional File 5). In addition, we performed Gene Ontology(GO) enrichment analyses. We also identified enriched expression for genes not previously called as differentially expressed due to low levels of absolute expression in individual studies. An example of this is *Smed-tert*, the gene encoding the protein subunit of telomerase [47] that is amongst the most enriched X1 genes by proportional expression but does not appear in previous individual studies because of low absolute expression (Figure 3A). We also used the ESCAPE database of human pluripotency factors [48] and found 233 best reciprocal hits to *S. mediterranea*. Looking at the

expression of these genes we found them to be enriched in the X1/X2 expression category (Additional File 5).

Taken together our analyses, using all publicly available data, define the transcribed loci whose expression can be detected in planarian FACS compartments. As well as defining absolute levels of relative expression, we represent data by proportion of expression in each FACS compartment. This allows us to generate expression spectra highlighting loci expressed disproportionately in G2/M stem cells, loci expressed throughout the cell cycle, loci with most of their expression in transient differentiating post-mitotic cells and those expressed mainly in post-mitotic differentiated cells. As our annotations and expression data are in the context of the genome assembly these data can be integrated with ChIP-seq data.

Expression spectra are supported by RNA-seq of RNAi phenotypes and single cell sequencing analyses

As an independent confirmation of our annotation and expression data we re-examined previously published RNA-seq after RNAi datasets and single cell RNA-seq data. For a selection of genes described as being required for stem cell progeny maintenance (Additional File 6) we visualized the RNA-seq profiles in relation to the defined FACS expression categories (Figure 2E) and observed that down-regulation of highly enriched X2 category genes was the main feature of both S*med-mex3* and *Smed-zfp1* RNAi datasets (Additional File 6). From this data, it is straightforward to conclude that both *Smed-mex-3* and *Smed-zfp-1* have a collective effect on many genes that normally switch on in NB progeny as they differentiate and leave the cell cycle, and this correlates with the phenotypic effects of RNAi in both cases causing a depletion in stem cell progeny as stem cell differentiation fails [9,24]. This approach to analyzing RNA-seq data is useful for identifying patterns in the global effects of RNAi.

Recently, two planarian single cell transcriptomic studies have also been used to define expression profiles of single stem cells and other cell types [22,23]. These have helped to reveal heterogeneity of expression profiles in planarian stem cells and provide persuasive evidence for the existence of cycling NB cells that might be committed to particular lineages [23,24]. We re-mapped available single cell RNA-seq data [22,23] to identify the the top one thousand genes ranked by expression for each cell type defined by these two studies. We looked at the position of these genes along expression spectra

sorted by X1 proportion (Additional File 6). The single cell data analyzed in this way follows patterns we would expect and independently validates our proportional expression spectra. For different NB populations defined by single cell studies (sigma, gamma, zeta and head X1) we saw enrichment of genes in the X1 and X1/X2 categories (Additional File 6). Differentiated cell types were enriched for genes in the Xins category. However, all differentiated cell-types, with the exception of the 'epidermis II' class [22], have an enrichment of genes in the X2 category as these genes are amongst those with highest absolute expression in all non-NB cell types, and thus appear amongst the top 1,000 expressed genes in single cell RNA-seq data.

Overall, our annotation and expression analysis is congruent and compatible with independent data from RNAi coupled RNA-seq experiments and single cell sequencing data, further validating the success of our approach.

An optimized ChIP-seq protocol reveals H3K4me3 levels at TSSs in cycling cells correlate with gene expression in NBs.

We next wished to combine our new genome annotation with predicted transcriptional start sites (TSSs) of expressed loci by cell compartment expression with NB derived epigenetic data. Research into epigenetic mechanisms in planarians is still very much in its infancy. Previous work characterized loss of function phenotypes of members of the NuRD complex [32,33,49,50], COMPASS and COMPASS-like families [51,52] and established a lack of endogenous DNA methylation in the *S. mediterranea* genome [32]. With respect to monitoring epigenetic marks, some of the effects of *mll1/2* and *set1* RNAi on the H3K4me3 mark of active transcription have been previously reported [52]). We noted that in this study, the total number of ChIP-seq reads from ~1 million X1 sorted planarian NBs were at relatively low numbers compared to those from *Drosophila* S2 'carrier' cells, which were added at 10x excess to X1 NBs (Additional File 7). These data suggested to us that ChIP-seq experiments with FACS sorted NBs might be very technically challenging.

In order to begin to study epigenetic regulation of NBs, we first developed an optimized protocol for ChIP-seq on FACS sorted X1 cells for H3K4me3 mark. Relatively high levels of H3K4me3 have been shown to be broadly characteristic of active promoters [53,54]. We found we were able to generate 13-26 million high quality *S. mediterranea* uniquely mapped reads using 150,000-200,000 X1 cells per immunoprecipitation, 5-7 times less

starting material compared to the a previous planarian ChIP-seq study [52]. We therefore used *Drosophila* S2 cells to act as a spike-in control for normalization of any technical replicate differences in immunoprecipitation across samples [55,56].

Sequencing X1 sorted cells, we observed high average H3K4me3 peaks around the TSSs of genes categorized as X1 and X1/X2 enriched, indicative of high expression in NBs (Figure 4A, Additional File 8). Conversely, we saw much less H3K4me3 at the promoters of Xins enriched gene. These results validate our planarian ChIP-seq method and confirm that our annotation is useful for studying global correlations between epigenetic marks and gene expression in the context of the genome. We saw intermediate levels of H3K4me3 in the X2 enriched compartment (Figure 4A, Additional File 8), which includes both NBs and recent post-mitotic progeny. A finer grained look at the X2 compartment revealed that genes with the highest proportion of X2 expression had lower levels of H3K4me3 in X1 cells (Figure 4B, Additional File 8), indicative of enriched expression in post mitotic progeny rather than cycling cells of the X1 compartment (see also Figure 3E). The presence of H3K4me3 at X2 enriched gene promoters as a whole is, however, higher than that observed in genes enriched for expression in the differentiated Xins cell compartment (Figure 4B). These observations are broadly in agreement with previous findings from X1 cells [52] using the previously available annotations [45,46].

A base by base correlation analysis of ChIP-seq signal across the promoter region to proportional expression in the X1/X2/Xins FACS compartments shows a positive correlation between X1 proportional expression and levels of H3K4me3 deposition from near the TSS region to ~1kb downstream (Figure 4C). On the other hand, there is a negative correlation between H3K4me3 deposition and the proportion of Xins expression across the same region (Figure 4C). Thus, higher H3K4me3 ChIP-seq signal in X1 cells tends to reflect higher gene expression in X1 cells and lower H3K4me3 signal reflects lower expression in X1 cells and higher expression in the Xins compartment. We also looked at a individual loci of genes known to be expressed in NB and found them all to have relatively high levels of H3K4me3 and low levels of suppressive marks (Figure 4D). Overall, the patterns we observe across the genome are consistent with what would be expected with H3K4me3 being an activating mark. Additionally, it broadly validates our annotation of transcribed loci, our assignment of proportional expression values for each

303 locus to FACS compartments and our method of ChIP-seq using relatively small
304 numbers of starting cells.
305
306 **Levels of the repressive histone marks H3K27me3 and H3K4me1 at TSSs also**
307 **correlate with gene expression in NBs.**
308 With an optimized ChIP-seq protocol, we decided to investigate two additional key
309 histone modifications, the repressive mark H3K27me3 important for the assessment of
310 bivalency [37,57] and H3K4me1 which has also recently been implicated as a repressive
311 mark at promoter regions, mediated by the MLL3/4 family of histone methyltransferases
312 [58].
313 We performed ChIP-seq on these two marks in X1 cells and observed ChIP-seq profiles
314 consistent with these marks being associated with repression of gene expression in
315 NBs. At loci enriched for X1 expression we observed low levels of H3K27me3 around
316 the TSS and higher signal for loci with enriched expression in the Xins FACS
317 compartment. (Figure 5A, Additional File 8). A positive correlation is observed around
318 the TSS and 1 kb downstream between the levels of H3K27me3 and expression in the
319 Xins compartment (figure 5B). This fairly broad domain of H3K27me3 is consistent with
320 previous studies in mammals [59,60] A negative correlation at the TSS is observed
321 between H3K27me3 signal and genes enriched for X1 expression (Figure 5B). Overall,
322 this pattern is the opposite to that observed for H3K4me3.
323 ChIP-seq to detect distribution of the H3K4me1 mark revealed a different pattern to that
324 of either H3K4me3 or H3K27me3. Rather than clear differences in the amount of
325 H3K4me1 signal between loci with different FACS expression profiles, we observed a
326 clear shift in the position of signal peaks (Figure 5C). Loci with a high proportion of
327 expression in the Xins FACS compartment have high levels of H3K4me1 close to the
328 TSS in X1 cells. Conversely, loci that are expressed in cycling cells (X1 and X1/X2
329 enriched) have peaks of H3K4me1 signal on average ~1kb downstream of the TSS.
330 Thus, the peak of H3K4me1 shifts away from the TSS for genes that are actively
331 expressed and is consistent with observations of a previous study looking at H3K4me1
332 levels at promoters in mammalian cells [58]. The Spearman correlation of H3K4me1
333 signal and FACS proportional expression confirms these observations, showing a
334 positive correlation close to the TSS for Xins enriched loci and a negative correlation for
335 X1 enriched loci (Figure 5D). The relationship between H3K4me1 and X1 enriched loci

is positive further downstream, at which position, we therefore conclude, this modification does not broadly exert a repressive effect (Figure 5D). We noticed that for X2 enriched genes H3K4me1 signal had two distinct peaks, one around the TSS and the other downstream. This suggests two populations of loci, one with raised levels of H3K4me1 near the TSS suggesting repression, and the another further downstream suggesting an absence of repression involving H3K4me1 (Figure 5D). One clear possibility is that the repressive peak near the TSS might be for genes that are off in NBs and only switch in in post-mitotic progeny, while the other peak represents X2 enriched genes that are expressed in cycling NBs. We also checked individual loci of genes known to be expressed in differentiated cells, and found they all had relatively high levels of repressive marks at or near the predicted TSS and low levels of H3K4me3 (Figure 5 E). Conversely genes known to be expressed in NBs had low levels of repressive marks (Figure 4D).

A common method of analyzing ChIP-seq data is to perform a cluster analysis on coverage profiles to observe whether groups of similar profiles are enriched for a biological function [61]. While this blind approach to analyzing ChIP-seq profiles can sometimes yield interesting results when manually checking cluster members, it is often the case that the broad biological interpretation of clusters is vague due to low-resolution third party classifications such as gene ontology. Instead of a blind approach, here we have used proportional expression to categorize loci into distinct groups to observe broad trends in the ChIP-seq data. Taken together, our work demonstrates that the dynamics between states of promoter histone methylation are distinct between loci grouped by expression dynamics, and in agreement with previously studied roles of these marks described in mammalian cells [53,54,57-60]. The congruity of our annotation data, expression analysis and ChIP-seq datasets validates our framework for studying epigenetic regulation in NBs. As well as the genome wide analyses presented here, it will now be possible to look at the epigenetic regulation of individual planarians genes or sets of genes of interest in different experimental and environmental conditions using ChIPseq data.

Evidence for the conservation of bivalent promoter regulation in pluripotent animal stem cells

Having validated our epigenetic analysis and demonstrated conservation of activating and suppressive marks we next investigated whether promoter of bivalency could be a regulatory mechanism in NBs. Bivalent promoters were originally observed at genomic loci for genes that were not expressed or expressed at very low levels in mouse ESCs [37], and were surprising because they contain both activating H3K4me3 and repressive H3K27me3 marks. This state is associated with the presence of RNA polymerase in a poised state and may allow rapid transcriptional responses to incoming signals to differentiate, at which point histone marks at bivalent promoters resolve so one of the two marks becomes dominant, resulting in activation or suppression of expression [40,41]. Bivalent promoters have since been described in various stem cells of different developmental origin and potency [39,42]. While they have been described outside of mammals in zebrafish [43], they have not so far been found in any invertebrates, suggesting they may be a novel epigenetic regulatory feature of vertebrates. Our ChIP-seq data from FACS sorted cells makes it possible to detect potential bivalent promoters in NBs if they are present in these cells. We reasoned loci that have relatively low expression in the X1 fraction and are up-regulated during differentiation and highly enriched in post-mitotic progeny (high X2 expression) may be good candidates for regulation by bivalent promoters in NBs.

We analysed the ChIP-seq signal as a continuous dataset by transforming the coverage profile into percent coverage by dividing the coverage at each base by the maximum coverage in the entire dataset. For each FACS category, we took the top one thousand most enriched loci and plotted the percent coverage profiles of H3K4me3, H3K4me1, and H3K27me3 to observe potential bivalency across all these loci. For the top one thousand X1 enriched loci, we see the expected profiles of a high H3K4me3 peak and low H3K27me3 peak (Figure 6A). We observe the opposite pattern for the top one thousand Xins enriched loci (Figure 6B). For the top one thousand X2 enriched loci, which are enriched for expression in post-mitotic progeny, we see similar percent coverage peaks for both H3K4me3 and H3K27me3 across these 1000 genes, consistent with potential bivalency in NBs at many of these promoters (Figure 6C).

As an independent source of validation, we also extracted all genes that were significantly down-regulated more than 2-fold after *Smed-mex3*(RNAi), which blocks the production of post-mitotic progeny. The ChIP-seq profile of these genes in X1 cells shows a similar profile to that of the top one thousand X2 enriched loci and is also

indicative of potential bivalency (Figure 6D). Genes expressed in the X2 compartment may stay on as cells differentiate so that they have an X2/Xins expression profile, some of these genes may also have bivalent promoters. Analysis of this gene sets also showed suggested some of these loci may be bivalent in X1 NBs (Figure 6E)

One caveat of our analysis so far is the possibly that bivalent ChIP signals represent underlying cell heterogeneity in the sampled X1 cell population [40]. While we know that cycling NB have some heterogeneity in gene expression that can describe subclasses with different lineage commitment [24], our focus on promoters of genes that are only upregulated upon differentiation and not expressed in NBs makes it unlikely the patterns we observe represent heterogenous epigenetic regulation in the NB population.. Similarly, given that our analysis identified bivalency across large numbers of promoters it is also possible that our observation is the result of genes that have mostly one or other mark in NBs leading to an average profile that appears bivalent profile when many genes are looked at simultaneously. To rule this possibility out we looked at the correlation (Pearson) between H3K4me3 and H3K27me3 and observed the distribution of correlations for the top 500 ranked amongst X1, X2 and loci with reduced expression in *Smed-mex3*(RNAi) (Figure 6F). For X2 and *Smed-mex3* RNAi category loci, we observe a high density of well correlated H3K4me3 and H3K27me3 profiles indicating similar paired signals for these marks across the TSS, and indeed closer inspection of individual genes confirms this to be the case (Figure 7). For X1 enriched loci, we see a less correlation, with many negatively correlated loci compared to the top X2 enriched genes and *Smed-mex3* RNAi loci. This analysis suggests that many 100s of loci are in fact bivalent with respect to H3K4me3 and H3K27me3 in planarian NBs.

Overall, our data demonstrate the presence of bivalency at promoters in NBs. This suggests that this mechanism of gene regulation may be conserved amongst animals rather than confined to vertebrates [37,62]. It seems likely that the need to have both embryonic and, where appropriate, adult stem cells, capable of sensitive regulatory decisions and responses to incoming signals may have arisen very near the origin of multicellularity. Our work suggests that the evolution of bivalent promoters, arising earlier than previously thought, may have been an important component of achieving stem cell flexibility.

Conclusion

While there have been successful attempts in the model species *S. mediterranea* to integrate transcriptome data from different sources to improve overall representation and annotation [45,46,63-65], different FACS expression datasets from different experiments and laboratories have not been integrated to improve the quality of gene expression profiles across these cell compartments. Additionally, many previous approaches quantifying gene expression have focused on using assembled transcriptomes without the context of a genome assembly. This means that linking these RNA-seq based expression datasets directly to epigenetic or transcription factor based regulation using ChIP-seq is not possible. The goal of our work here was to address these deficits by combining transcriptome and epigenetic approaches to describe the landscape of epigenetic regulation at promoter regions in NBs in the context of expression level data. Our analyses validate our annotation, trancriptome analysis and ChIP-seq protocol and provide clear demonstration of the existence of bivalent promoters in cycling NBs. Our analysis is particularly sensitive for detecting genes that switch on after NB differentiation, due to the structure of the transcriptome and epigenetic datasets available for analysis. Future work can now use planarians as a model for understanding how this mode of regulation works, and the similarities and differences with vertebrates.

The discovery that bivalent promoters exist outside of vertebrates adds to the growing body of evidence that suggests a deeper conservation of stem cell biology amongst animals then previously appreciated. Previously, endogenous genome stability mechanisms, splicing and post-transcriptional regulatory mechanisms have all been shown to be important for NB function [17,28,66]. Additionally, a number of proteins involved in epigenetically mediated gene regulation have also been shown to be essential to maintain NB function [10,13,15]. Particularly the previously described cases of MBNF/CELF mediated splicing regulation [17] and PIWI mediated genome stability [19], these represent deeply conserved processes that likely mediated stem cell function in an ancestral animal. Our work suggests that bivalent promoters represent yet another major conserved mechanism and this regulatory process is not, as previously thought, vertebrate specific. As well as demonstrating bivalency, our work, through establishment of an annotation framework and a robust ChIP-seq protocol for NBs, will allow the use of planarians as a model for epigenetic regulation of stem cell function. For example, the accessibility of the NB population should allow identification of regulatory targets of

chromatin modifying enzymes responsible for pluripotency, self-renewal and proliferation.

Materials and methods

Data Sources for this study

The NCBI Project accession number for ChIP-seq data produced in this study is PRJNA338116. All accession numbers for previously published RNA and ChIP-seq data used in the study are listed in Additional file 9.

Flow cytometry

A modified version of a planarian FACS protocol [27] was used. The modifications were: a 35 µm mesh filter was used instead of 100 µm, staining with Hoechst and calcein was performed simultaneously rather than sequentially and the centrifugation-wash step was omitted. We used Hoechst 34580 instead of Hoechst 33342. A FACS Aria III machine equipped with a violet laser was used for cell sorting. BD FACSDiva and FlowJo software were used for analyses and setting cell population gates.

ChIP-seq

For each experimental replicate 600,000-700,000 planarian X1 cells (enough for Chip-seq of 3 histone marks and an input control sample) were FACS-sorted (using 3-day regenerates) in PBS and pelleted at 4 ºC. The pellet was re-suspended in nuclei extraction buffer (0.5% NP40, 0.25% Triton X-100, 10mM Tris HCl pH 7.5, 3mM CaCl2, 0.25mM sucrose, 1mM DTT, 1/10th Phosphatase Cocktail Inhibitor 2 (Sigma Aldrich), 1/10th Phosphatase Cocktail Inhibitor 3 (Sigma Aldrich)). This was followed by formaldehyde fixation, that was stopped with 2.5M glycine. The pellet was re-suspended in SDS lysis buffer (1% SDS, 50mM Tris HCl pH 8, 10mM EDTA) and incubated on ice. ChIP dilution buffer (0.1% SDS, 1.2mM EDTA, 16.7mM Tris HCl pH 8, 167mM NaCl, 1/1000th Phosphatase Cocktail Inhibitor 2, 1/1000th Phosphatase Cocktail Inhibitor 3, 1mM DTT) was added in a 2.3:1 ratio to the sample. Samples were sonicated and 1/10th volume 10% Triton X-100 was added. Samples were pelleted at 4 ºC and the supernatant kept for further processing. Test de-crosslinkling was performed on 1/8th

volume of the chromatin solution to verify the DNA fragment range following sonication was 100-500 bp.

Protein A-covered Dynabeads (Thermofisher) were used for immunoprecipitation (IP). The amount of reagent used was in a 1:2 ratio to the amount of chromatin used per IP. The Dynabeads were first pre-blocked with 0.5% BSA/PBS and re-suspended in 0.5% BSA/PBS (2.5 times their original volume) containing 7 µg of antibody per IP. ChIP grade antibodies used were anti-H3K4me3 (rabbit polyclonal; Abcam; ab8580), anti-H3K4me1 (rabbit polyclonal; Abcam; ab8895), anti-H3K27me3 (mouse monoclonal; Abcam; ab6002).

After overnight incubation of the Dynabeads at 4 ºC, they were washed 3 times with 0.5% BSA/PBS and re-suspended in 0.5% BSA/PBS, matching their original volume. 1/4th of the total chromatin was used for each IP, leaving a final 1/8th for input control libraries. The IP was done on a rotating wheel overnight at 4 ºC.

Post-IP washes were done 6 times with RIPA buffer (50mM HEPES-KOH pH 8, 500mM LiCl2, 1mM EDTA, 1% NP40, 0.7% Sodium Deoxycholate, cOmplete protease inhibitors – 1 tablet per 50 ml). Beads were then washed in TE/NaCl (50mM NaCl in TE) and re-supended in Elution Buffer (50mM Tris HCl pH 8, 10mM EDTA, 1% SDS). Proteins were separated from the beads via 15-minute incubation at 65 ºC on a shaking heat block (1400 rpm). Supernatant and input samples underwent overnight heat-based de-crosslinking at 65 ºC. RNaseA (0.2 µg/ml) and Proteinase K (0.2 µg/ml) were used for 1 hour each in order to remove residual RNA and protein. DNA was purified with phenol:chloroform extraction and ethanol precipitation. DNA was re-suspended in TE and quantified with Qubit ds DNA HS kit (Thermo Fisher Scientific). The NEBNext Ultra II (NEB) kit was used for library preparation. Manufacturer's instructions were followed. Library clean-up was performed with Becton Coulter AMPureXP beads. Libraries were quantified with Qubit, Agilent Bioanalyzer and using a KAPA Library Quantification qPCR kit. Libraries were sequenced on an Illumina NextSeq machine.

Comparison of previous NB transcriptomes

Independently assembled transcriptomes were downloaded from four previous publications [10,14,20,21]. Transcripts enriched in NBs were extracted based on the classifications provided in respective publications' supplementary information. A clustering of these sequences was done by running CAP3 [67] on all transcripts and

then extracting transcript groups that assembled. Detailed methods are recorded in an IPython notebook (Additional File 9).

Reference assembly and annotations

Transcript sequences from previously assembled transcriptomes (Oxford, Dresden, SmedGD Asexual, SmedGD Unigenes) and known genes were downloaded from SmedGD [46], PlanMine [63] and NCBI. These sequences were mapped to the SmedGD Asexual 1.1 genome with GMAP [68]. PASA [69] was then used to consolidate the annotations. An independent reference assembly was also performed on 164 available RNA-seq libraries with HISAT2 [70] for mapping and StringTie [71] for assembly. PASA consolidated annotations and StringTie reference assembly were merged together with StringTie.

To remove redundancy from the annotations we first calculated an intron jaccard similarity score (intersection of introns / union of introns) for all overlapping transcripts. Pair-wise jaccard similarity scores of 0.9 or more were kept and used to create a graph of similar annotations. Maximal-cliques were extracted from this graph as clusters of redundant annotations. From these cliques, we chose one transcript to be the representative by prioritizing transcript length, ORF length and BLAST homology. Strand information was assigned to each transcript by using strand specific RNA-seq libraries, BLAST homology, and longest ORF length. We ran TransDecoder (utilizing Uniprot and PFAM for coding evidence) [72] to identify protein coding transcripts. Detailed methods are recorded in an IPython notebook (Adddional File file 9). The genome annotations are made available here as a gtf file (Additional File 10).

Proportional expression value generation

Kallisto [73] was used to pseudo-map RNA-seq libraries from four datasets [9,13,24], accession: PRJNA296017, generating estimated counts and TPM values for each transcript. Sleuth [74] was used to calculate a normalization factor for each library. For each locus, the TPM values of member transcripts were summed to generate a loci TPM value and then normalized accordingly.

Not all datasets contained all three X1/X2/Xins populations. The Reddien [24] and Sanchez (accession: PRJNA296017) datasets only had two of the three populations. In order to consolidate proportional expression values among all four datasets, pair-wise

ratios were first calculated for each dataset (X1:X2, X1:Xins, X2:Xins) using normalized TPM values. These ratios were then averaged across the datasets.

Using two of the three ratios, we can calculate a predicted third ratio (i.e., given X1:X2 and X1:Xins, we can calculate X2:Xins). We then correlate the calculated proportion with the actual proportion and kept the pairs of actual proportions (in this case, X1 and Xins) that had the best correlation with the calculated proportion. Detailed methods are recorded in an IPython notebook (Additional File 9).

Single cell RNA-seq analysis

Single cell RNA-seq data were downloaded from short-read archive [22,23]. Reads were pseudo-mapped with Kallisto and the TPM values were used for down-stream analysis. Cell types of each RNA-seq library were previously defined in the respective publications by both FACS (X1/X2/Xins) and by gene markers.

The top 1,000 expressed loci from each cell type cluster were used for generating the spectrum density figure and ternary plots.

ChIP-seq mapping and track generation

Biological triplicate ChIP-seq data from planarian X1 cells for each of three histone marks considered was analyzed in conjunction with *D. melanogaster* S2 spike-in cells, used for downstream between IP replicate normalization. The trimmed reads were mapped to both *S. mediterranea* asexual 1.1 genome (SmedGD, [46]) and *D. melanogaster* r6.10 genome [75] with BWA mem 0.7.12 [76]. Only uniquely mapping reads were considered further. Paired reads that map to both species were also removed. Picard tools 1.115 was used to remove duplicate reads. Reads were separated into the sets that mapped to *D. melanogaster* and *S. mediterranea* respectively so that numbers of mapped reads could be used for downstream normalization calculations. For each paired or single mapped read, coordinates representing the 100bp at the center of the sequenced fragment were parsed and written to a BED file.

To generate coverage tracks in bedgraph format, the bedtools' genomecov function was used. A normalization factor was calculated using the number of mapped reads corresponding to the *D. melanogaster* spike-in [55,56]. A scaling factor for the input ChIP-seq libraries was calculated using the DeepTools [61] python API that utilizes the

SES method [77]. Mean normalized coverage was calculated for each sample and input. The normalized input coverage was then subtracted from the normalized sample coverage to generate the final coverage track for downstream visualization and analysis. The normalization process is detailed in an IPython notebook (Additional file 7)

To calculate correlation of ChIP-seq coverage to proportional expression, two vector of values were used for a group of loci. The first vector is the proportional expression and the second vector is the coverage at a position in the 5kb region of the loci. A spearman correlation was performed on both vectors yielding a correlation value for the assayed position. This correlation value was calculated for every non-overlapping 50 base pair window in the 5kb region around the TSS.

For bivalency, a percent coverage was used instead of the absolute normalized coverage. This was generated by calculating the maximum coverage across all 5kb regions around assay loci. Each absolute coverage value across the loci is then divided by the maximum coverage resulting in a percent coverage.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All data produced in this study is in the form ChiPseq data submitted under The NCBI Project accession number PRJNA338116.

Competing interests

The authors declare that they have no competing interests

Funding

This work was funded by grants from the Medical Research Council (grant number MR/M000133/1) and the Biotechnology and Biological Sciences Research Council (grant number BB/K007564/1) to AA. A.G.L. is funded by a Human Frontier Science Program fellowship.

Authors' contributions

AA conceived designed the study. DK led and performed all data analysis with help from YM and AA. YM led the acquisition of all experimental data, with help from AGL. SH and YM optimized the ChIP-seq protocol. DK, YM and AA wrote the manuscript.

Acknowledgements

We thank members of the AA lab for comments on the manuscript.

References

1. Salo E, Baguna J. Regeneration and pattern formation in planarians. II. and role of cell movements in blastema formation. Development. The Company of Biologists Ltd; 1989;107:69–76.

2. Wagner DE, Wang IE, Reddien PW. Clonogenic Neoblasts Are Pluripotent Adult Stem Cells That Underlie Planarian Regeneration. Science. 2011;332:811–6.

3. Newmark PA, Sánchez Alvarado A. Bromodeoxyuridine Specifically Labels the Regenerative Stem Cells of Planarians. Developmental Biology. 2000;220:142–53.

4. Sánchez Alvarado A, Newmark PA. Double-stranded RNA specifically disrupts gene expression during planarian regeneration. Proceedings of the National Academy of Sciences of the United States of America. National Academy of Sciences; 1999;96:5049–54.

5. Eisenhoffer GT, Kang H, Alvarado AS. Molecular Analysis of Stem Cells and Their Descendants during Cell Turnover and Regeneration in the Planarian Schmidtea mediterranea. Cell Stem Cell. 2008;3:327–39.

6. Guo T, Peters AHFM, Newmark PA. A bruno-like Gene Is Required for Stem Cell Maintenance in Planarians. Developmental Cell. 2006;11:159–69.

7. Cowles MW, Brown DDR, 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;140:4691–702.

8. Lapan SW, Reddien PW. dlx and sp6-9 Control Optic Cup Regeneration in a Prototypic Eye. Desplan C, editor. PLoS Genet. Public Library of Science; 2011;7:e1002226–13.

9. Zhu SJ, Hallows SE, Currie KW, Xu C, Pearson BJ. A mex3 homolog is required for differentiation during planarian stem cell lineage development. eLife. eLife Sciences Publications Limited; 2015;4:304.

10. Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DDR, et al. A Comparative Transcriptomic Analysis Reveals Conserved Features of Stem Cell Pluripotency in Planarians and Mammals. STEM CELLS. 2012;30:1734–45.

11. Scimone ML, Srivastava M, Bell GW, Reddien PW. A regulatory program for

664 excretory system regeneration in planarians. Development. 2011;138:4387–98.

665 12. Rink JC, Vu HTK, Alvarado AS. The maintenance and regeneration of the planarian
666 excretory system are regulated by EGFR signaling. Development. 2011;138:3769–80.

667 13. Onal PO, n DGU, Adamidi C, Rybak A, Solana J, Mastrobuoni G, et al. Gene
668 expression of pluripotency determinants is conserved between mammalian and
669 planarian stem cells. The EMBO Journal. Nature Publishing Group; 2012;31:2755–69.

670 14. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, et al. De novo
671 assembly and validation of planaria transcriptome by massive parallel sequencing and
672 shotgun proteomics. Genome Research. 2011;21:1193–200.

673 15. Solana J, Kao D, Mihaylova Y, Jaber-Hijazi F, Malla S, Wilson R, et al. Defining the
674 molecular profile of planarian pluripotent stem cells using a combinatorial RNAseq, RNA
675 interference and irradiation approach. Genome Biol. BioMed Central Ltd; 2012;13:R19–
676 23.

677 16. Alié A, Hayashi T, Sugimura I, Manuel M, Sugano W, Mano A, et al. The ancestral
678 gene repertoire of animal stem cells. Proceedings of the National Academy of Sciences
679 of the United States of America. 2015;:201514789–8.

680 17. Solana J, Irimia M, Ayoub S, Orejuela MR, Zywitza V, Jens M, et al. Conserved
681 functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative
682 splicing in planarians. eLife. eLife Sciences Publications Limited; 2016;5:1193.

683 18. Juliano CE, Swartz SZ, Wessel GM. A conserved germline multipotency program.
684 Development. 2010;137:4113–26.

685 19. Shibata N, Kashima M, Ishiko T, Nishimura O, Rouhana L, Misaki K, et al.
686 Inheritance of a Nuclear PIWI from Pluripotent Stem Cells by Somatic Descendants
687 Ensures Differentiation by Silencing Transposons in Planarian. Developmental Cell.
688 Elsevier Inc; 2016;37:226–37.

689 20. Blythe MJ, Kao D, Malla S, Rowsell J, Wilson R, Evans D, et al. A Dual Platform
690 Approach to Transcript Discovery for the Planarian Schmidtea Mediterranea to Establish
691 RNAseq for Stem Cell and Regeneration Biology. Jaeger J, editor. PLoS ONE.
692 2010;5:e15617–7.

693 21. Resch AM, Palakodeti D, Lu Y-C, Horowitz M, Graveley BR. Transcriptome Analysis
694 Reveals Strain-Specific and Conserved Stemness Genes in Schmidtea mediterranea.
695 Boutros M, editor. PLoS ONE. 2012;7:e34447–12.

696 22. Wurtzel O, Cote LE, Poirier A, Satija R, Regev A, Reddien PW. A Generic and Cell-
697 Type-Specific Wound Response Precedes Regeneration in Planarians. Developmental
698 Cell. Elsevier Inc; 2015;35:632–45.

699 23. Molinaro AM, Pearson BJ. In silico lineage tracing through single cell transcriptomics
700 identifies a neural stem cell population in planarians. Genome Biol. Genome Biology;
701 2016;:1–17.

24. van Wolfswinkel JC, Wagner DE, Reddien PW. Single-Cell Analysis Reveals Functionally Distinct Classes within the Planarian Stem Cell Compartment. Stem Cell. Elsevier Inc; 2014;15:326–39.

25. 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;48:371–80.

26. Higuchi S, Hayashi T, Hori I, Shibata N, Sakamoto H, Agata K. Characterization and categorization of fluorescence activated cell sorted planarian stem cells by ultrastructural analysis. Development, Growth & Differentiation. 2007;49:571–81.

27. Romero BT, Evans DJ, Aboobaker AA. FACS Analysis of the Planarian Stem Cell Compartment as a Tool to Understand Regenerative Mechanisms. In: Orgogozo V, Rockman MV, editors. Hox Genes. Totowa, NJ: Humana Press; 2012. pp. 167–79.

28. Shibata N, Hayashi T, Fukumura R, Fujii J, Kudome-Takamatsu T, Nishimura O, et al. Comprehensive gene expression analyses in pluripotent stem cells of a planarian, Dugesia japonica. Int. J. Dev. Biol. 2012;56:93–102.

29. Mangel M. Feedback Control in Planarian Stem Cell Systems. BMC Systems Biology. BMC Systems Biology; 2016;:1–18.

30. Salvetti A. DjPum, a homologue of Drosophila Pumilio, is essential to planarian stem cell maintenance. Development. 2005;132:1863–74.

31. Reddien PW. Specialized progenitors and regeneration. Development. 2013;140:951–7.

32. Jaber-Hijazi F, Lo PJKP, Mihaylova Y, Foster JM, Benner JS, Romero BT, et al. Planarian MBD2/3 is required for adult stem cell pluripotency independently of DNA methylation. Developmental Biology. Elsevier; 2013;384:141–53.

33. Scimone ML, Meisel J, Reddien PW. The Mi-2-like Smed-CHD4 gene is required for stem cell differentiation in the planarian Schmidtea mediterranea. Development. 2010;137:1231–41.

34. Barberan S, Fraguas S, Cebrià F. The EGFR signaling pathway controls gut progenitor differentiation during planarian regeneration and homeostasis. Development. 2016;143:2089–102.

35. González-Estévez C, Felix DA, Smith MD, Paps J, Morley SJ, James V, et al. SMG-1 and mTORC1 Act Antagonistically to Regulate Response to Injury and Growth in Planarians. Reddien P, editor. PLoS Genet. 2012;8:e1002619–17.

36. Tu KC, Cheng L-C, T K Vu H, Lange JJ, McKinney SA, Seidel CW, et al. Egr-5 is a post-mitotic regulator of planarian epidermal differentiation. eLife. eLife Sciences Publications Limited; 2015;4:e10501.

37. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A Bivalent

739 Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell.
740 2006;125:315–26.

741 38. McDonald D. Chromatin signatures of pluripotent cell lines. 2006;:1–11.

742 39. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-
743 wide maps of chromatin state in pluripotent and lineage-committed cells. Nature.
744 2007;448:553–60.

745 40. Voigt P, Tee WW, Reinberg D. A double take on bivalent promoters. Genes &
746 development. 2013;27:1318–38.

747 41. Lesch BJ, Page DC. Poised chromatin in the mammalian germ line. Development.
748 2014;141:3619–26.

749 42. Hammoud SS, Low DHP, Yi C, Carrell DT, Guccione E, Cairns BR. Chromatin and
750 Transcription Transitions of Mammalian Adult Germline Stem Cells and
751 Spermatogenesis. Stem Cell. Elsevier Inc; 2014;15:239–53.

752 43. Vastenhouw NL, Schier AF. Bivalent histone modifications in early embryogenesis.
753 Current Opinion in Cell Biology. 2012;24:374–86.

754 44. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, et al.
755 Functional Anatomy of Polycomb and Trithorax Chromatin Landscapes in Drosophila
756 Embryos. Kingston R, editor. PLoS biology. 2009;7:e1000013–8.

757 45. Robb SMC, Ross E, Alvarado AS. SmedGD: the Schmidtea mediterranea genome
758 database. Nucleic Acids Research. 2007;36:D599–D606.

759 46. Robb SMC, Gotting K, Ross E, Sánchez Alvarado A. SmedGD 2.0: The Schmidtea
760 mediterraneagenome database. Vize P, Westerfiled M, editors. genesis. 2015;53:535–
761 46.

762 47. Tan TCJ, Rahman R, Jaber-Hijazi F, Felix DA, Chen C, Louis EJ, et al. Telomere
763 maintenance and telomerase activity are differentially regulated in asexual and sexual
764 worms. Proceedings of the National Academy of Sciences of the United States of
765 America. National Acad Sciences; 2012;109:4209–14.

766 48. Xu H, Baroukh C, Dannenfelser R, Chen EY, Tan CM, Kou Y, et al. ESCAPE:
767 database for integrating high-content published data collected from human and mouse
768 embryonic stem cells. Database. 2013;2013:bat045–5.

769 49. Bonuccelli L, Rossi L, Lena A, Scarcelli V, Rainaldi G, Evangelista M, et al. An
770 RbAp48-like gene regulates adult stem cells in planarians. Journal of Cell Science.
771 2010;123:690–8.

772 50. Vásquez-Doorman C, Petersen CP. The NuRD complex component p66suppresses
773 photoreceptor neuron regeneration in planarians. Regeneration. 2016;3:168–78.

774 51. Hubert A, Henderson JM, Ross KG, Cowles MW, Torres J, Zayas RM. Epigenetic

775 regulation of planarian stem cells by the SET1/MLL family of histone methyltransferases.
776 Epigenetics. 2014;8:79–91.

777 52. Duncan EM, Chitsazan AD, Seidel CW, Alvarado AS. Set1 and MLL1/2 Target
778 Distinct Sets of Functionally Different Genomic Loci In Vivo. CellReports. The
779 Authors; 2015;13:2741–55.

780 53. Ruthenburg AJ, Allis CD, Wysocka J. Methylation of Lysine 4 on Histone H3:
781 Intricacy of Writing and Reading a Single Epigenetic Mark. Molecular Cell. 2007;25:15–
782 30.

783 54. Eissenberg JC, Shilatifard A. Histone H3 lysine 4 (H3K4) methylation in development
784 and differentiation. Developmental Biology. Elsevier Inc; 2010;339:240–9.

785 55. Orlando DA, Chen MW, Brown VE, Solanki S, Choi YJ, Olson ER, et al. Quantitative
786 ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. CellReports. The
787 Authors; 2014;9:1163–70.

788 56. Hu B, Petela N, Kurze A, Chan K-L, Chapard C, Nasmyth K. Biological
789 chromodynamics: a general method for measuring protein occupancy across the
790 genome by calibrating ChIP-seq. Nucleic Acids Research. 2015;:gkv670–20.

791 57. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, et al. Role of
792 Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. Science.
793 2002;298:1039–43.

794 58. Cheng J, Blum R, Bowman C, Hu D, Shilatifard A, Shen S, et al. A Role for H3K4
795 Monomethylation in Gene Repression and Partitioning of Chromatin Readers. Molecular
796 Cell. Elsevier Inc; 2014;53:979–92.

797 59. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, et al. H3K27me3
798 forms BLOCs over silent genes and intergenic regions and specifies a histone banding
799 pattern on a mouse autosomal chromosome. Genome Research. 2008;19:221–33.

800 60. Sparmann A, van Lohuizen M. Polycomb silencers control cell fate, development
801 and cancer. Nat Rev Cancer. 2006;6:846–56.

802 61. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al.
803 deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic
804 Acids Research. 2016;44:W160–5.

805 62. Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, et al. Chromatin
806 signature of embryonic pluripotency is established during genome activation. Nature.
807 Nature Publishing Group; 2010;464:922–6.

808 63. Brandl H, Moon H, Vila-Farré M, Liu S-Y, Henry I, Rink JC. PlanMine – a mineable
809 resource of planarian biology and biodiversity. Nucleic Acids Research. 2016;44:D764–
810 73.

811 64. Kao D, Felix D, Aboobaker A. The planarian regeneration transcriptome reveals a

812 shared but temporally shifted regulatory program between opposing head and tail
813 scenarios. BMC Genomics. BioMed Central; 2013;14:797.

814 65. Rodríguez-Esteban G, González-Sastre A, Rojo-Laguna JI, Saló E, Abril JF. Digital
815 gene expression approach over multiple RNA-Seq data sets to detect neoblast
816 transcriptional changes in Schmidtea mediterranea. BMC Genomics. 2015;16:951–23.

817 66. Solana J, Gamberi C, Mihaylova Y, Grosswendt S, Chen C, Lasko P, et al. The
818 CCR4-NOT Complex Mediates Deadenylation and Degradation of Stem Cell mRNAs
819 and Promotes Planarian Stem Cell Differentiation. Newmark PA, editor. PLoS Genet.
820 2013;9:e1004003–16.

821 67. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome
822 Research. Cold Spring Harbor Laboratory Press; 1999;9:868–77.

823 68. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for
824 Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality.
825 Methods Mol. Biol. New York, NY: Springer New York; 2016;1418:283–334.

826 69. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR. Approaches to Fungal
827 Genome Annotation. Mycology. 2011;2:118–41.

828 70. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory
829 requirements. Nat. Methods. Nature Research; 2015;12:357–60.

830 71. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL.
831 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.
832 Nature Biotechnology. Nature Research; 2015;33:290–5.

833 72. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De
834 novo transcript sequence reconstruction from RNA-seq using the Trinity platform for
835 reference generation and analysis. Nature Protocols. Nature Research; 2013;8:1494–
836 512.

837 73. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq
838 quantification. Nature Biotechnology. Nature Research; 2016;34:525–7.

839 74. Pimentel HJ, Bray N, Puente S, Melsted P, Pachter L. Differential analysis of RNA-
840 Seq incorporating quantification uncertainty. 2016.

841 75. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, et al. FlyBase:
842 establishing a Gene Group resource for Drosophila melanogaster. Nucleic Acids
843 Research. 2016;44:D786–92.

844 76. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler
845 transform. Bioinformatics. Oxford University Press; 2010;26:589–95.

846 77. Diaz A, Park K, Lim DA, Song JS. Normalization, bias correction, and peak calling
847 for ChIP-seq. Stat Appl Genet Mol Biol. 2012;11:Article9.

Figure legends

Figure 1. Planarian FACs compartments and analysis of currently available neoblast transcriptome datasets.
A) Schematic of FACS cell populations and their relationship to stages in the cell cycle, stem cell progeny and differentiated cells. B) Venn diagram describing the overlaps from four independently assembled transcriptomes and genes described as being enriched in neoblasts, produced by the Aboobaker, Pearson, Rajewsky and Graveley labs respectively.

Figure 2. Proportional transformation of gene expression values in planarian FACS compartments.
Spectrum of genes sorted by X1 (A), X2 (B), and Xins (C) proportion of expression where each vertical line in the spectrum represents one expressed loci. The proportion of dark blue, light blue, and orange corresponds proportions of expression in the X1, X2, and Xins FACS compartments respectively. D) A table presenting colour-coded classification groups according to proportional expression in different FACS populations of cells based on the detailed analysis of proportional expression (Additional Files 4 and 5).

Figure 3. Gene categories based on proportional expression values.
Previously described planarian genes are marked in expression profile following panels displaying a gene category. A) Genes with 50% or more X1 proportional expression. B) Genes with 50% or more Xins proportional expression. C) Gene ontology (GO) enrichment of X1 enriched genes showing terms mainly associated with cell division. D) GO enrichment of Xins genes showing terms associated with, for example, the extracellular matrix. E) Genes with 50% or more X2 proportional expression. F) Genes with the of sum X1 and X2 proportional expression more than or equal to 75% and in neither falling into X1 nor X2 categories. Gene names in blue and red are not characterised *S.mediterranea* genes. Blue gene names are genes associated with methyltransferase activity according to GO. Red names are genes associated with

879 mRNA processing according to GO. G) GO enrichment for X1/X2 genes from F)
880 showing enrichment of terms involved in RNA and ribosomal processes. H) Genes with
881 sum X2 and Xins proportional expression more than or equal to 75% and in neither X2
882 nor Xins categories. I) Genes that are expressed in roughly similar proportions among
883 X1, X2, and Xins cells.
884
885
886 Figure 4. ChIP-seq analysis of H3K4me3 in planarian neoblasts.
887 A) ChIP-seq profile of 5kb around predicted transcriptional start sites (TSS) for
888 H3K4me3 histone marks in X1 cells. B) The amount of H3K4me3 signal decreases with
889 increasing proportional expression in the X2 compartment, indicative of expression
890 becoming limited to post-mitotic progeny rather than NBs. C) The correlation for each 50
891 bp window across the 5kb region around TSS to the X1 (dark blue), X2 (light blue), and
892 Xins (orange) proportional expression value. A positive correlation value means that the
893 higher the ChIP-seq signal, the higher the proportional expression value. A negative
894 correlation means that the lower a ChIP-seq signal, the higher the proportional
895 expression value. D) Example ChIP-seq profiles of individual planarian neoblast genes.
896
897 Figure 5. ChIP-seq analysis of repressive marks in planarian neoblasts.
898 A) ChIP-seq profile of 5kb around transcriptional start sites (TSS) for H3K27me3 in X1
899 cells B) The correlation between H3K27me3 ChIP-seq signal with the X1 (dark blue), X2
900 (light blue), and Xins (orange) proportional expression value for each 50 bp window
901 across the 5kb region around TSS. A positive correlation value means that the higher
902 the ChIP-seq signal, the higher the proportional expression value. A negative correlation
903 means that the lower a ChIP-seq signal, the higher the proportional expression value. C)
904 ChIP-seq profile of 5kb around transcriptional start sites (TSS) for H3K4me1 in X1 cells
905 D) The correlation between H3K4me1 ChIP-seq signal with the X1 (dark blue), X2 (light
906 blue), and Xins (orange) proportional expression value for each 50 bp window across
907 the 5kb region around TSS. E) Example ChIP-seq profiles of individual planarian genes
908 expressed in differentiated cells.
909
910 Figure 6. Bivalency of activation and repressive histone marks signal and shifting of
911 H3K4me1 signal. For figure A-D, the mean ChIP-seq profiles shown were transformed

into percent coverage (y axis) by dividing each coverage value by the max coverage among all loci. The percent coverage of genomic region around TSS (x axis) was plotted. A) H3K4me3 and H3K27me3 profile of the top 1000 ranked X1 enriched genes. B) H3K4me3 and H3K27me3 profile of the top 1000 ranked Xins enriched genes. C) H3K4me3 and H3K27me3 profile of the top 1000 ranked X2 enriched genes. D) H3K4me3 H3K27me3 profile of genes down-regulated after *Smed-mex3* RNAi. E) H3K4me3 and H3K27me3 profile of X2/X2ins enriched genes with less than 10% X1 proportional expression. F) The distribution of correlations between H3K4me3 signal 5KB around TSS and H3K27me3 signal 5KB around TSS for X1, X2, and *Smed-mex3* RNAi loci.

Figure 7. ChIP-seq profiles of high ranked X2 genes focusing on annotated transcription factors as examples. H3K4me3 and H3K27me3 profiles are displayed for the 5kb region surrounding the transcription start site of each loci.

Additional files

Additional File 1. PDF Format.
Bar chart showing the number of neoblast transcripts from each of the dataset (blue), as well as the number of neoblast transcripts that have at least one other match in another dataset (red).

Additional file 2. PDF Format.
Asexual genome annotation workflow and metrics. A) A schematic of the workflow used to annotate the asexual genome. This process involved utilizing previously *de novo* assembled transcripts and available RNA-seq datasets. PASA (Program to Assemble Spliced Alignments) was used to create a merged reference assembly. HISAT2 was used for mapping RNA-seq data to the genome and StringTie for defining transcripts. The two sets of reference assemblies (from consolidated transcriptomes and from RNA-seq data) were merged with StringTie and filtered for redundancies, resulting in a final annotation set. B) Proportion of the annotations from the final annotation set that are

likely coding (with TransDecoder evidence) and non-coding (no TransDecoder evidence). Proportion of loci without TransDecoder evidence but with a BLAST hit to the non-redundant (NR) protein database (e-value <= 1e-5) is also shown. C) A comparison of the new annotation in this study with MAKER annotations available on *Schmidtea mediterranea* Genome Database (SmedGD) showing cumulative percentages of annotations at a range of expression value thresholds for SmedGD MAKER annotations as a whole, for the 'Oxford' annotations as a whole, for SmedGD MAKER exclusive annotations, and 'Oxford' exclusive annotations, and 'Oxford' coding annotations.

Additional File 3. PDF Format.
Mapping available FACS RNA-seq libraries to new annotations. A) FACS libraries from the Rajewsky, Reddien, Pearson, and Sanchez labs were downloaded and mapped to the annotations with Kallisto and normalized using Sleuth. Normalization was performed for datasets within each lab. B) A hierarchical clustering of the FACS samples using normalized transcripts per million (TPM). Wherever possible, a Short Read Archive Run (SRR) ID is provided for the original dataset

Additional File 4. PDF Format.
Proportional transformation of expression values. Expression values were converted to proportional expression values resulting in consistent clustering of samples.

Additional File 5. PDF Format.
The ternary plots within this PDF file describe subsets of expressed loci with each of the three axis representing X1, X2, and Xins proportional expression. Dots represent loci, which are categorized according to proportional expression. X1 is dark blue, X2 is light blue, Xins is orange, X1/X2 is green, X2/Xins is red, X1/Xins is purple, and non-enriched is grey.
The second page shows the best reciprocal hits to the ESCAPE database. Human pluripotency factors were obtained from the ESCAPE database. Best reciprocal hits were found between human and *S.mediterranea* genes. A) Ternary plot showing the distribution of the *S.mediterranea* best reciprocal hits of human pluripotency factors. B) Shows the same set of data but extracted from the proportional spectrum. C) Shows a pie chart indicating the percentage of the 233 genes that belong to each category.

The third, fourth, fifth pages show ternary plots of loci with X1/X2, Xins, and X1 enriched GO terms.

Additional File 6. PDF Format.

RNA-seq profiles of selected RNAi datasets. The first page of the PDF file shows RNA-seq profiles divided into four segments representing genes enriched in X1, X2, Xins, and X1/X2. The proportional values of each category are plotted as dark blue (X1 proportion), light blue (X2 proportion), and orange (Xins proportion) on the right of each profile. The RNA-seq profile is displayed as red (up-regulated) and blue (down-regulated) lines that are significantly differentially expressed with a p-value less than 0.05 and fold-change value < -1.5 (log2 fold-change –0.58) or > 1.5 (log2 fold-change 0.58). The length of each line represents the log2 fold-change. A) RNA-seq profile of *Smed-mex3* RNAi performed on whole worms. B) RNA-seq profile of *Smed-zfp-1* performed on X2 and X1 cells. C) RNA-seq time-course data for *Smed-CHD4* RNAi performed on whole worms. D) RNA-seq time-course data for *Smed-p53* RNAi performed on whole worms. E) RNA-seq data for *Smed-coe* RNAi performed on whole worms.

The second page of the PDF file shows single-cell RNA-seq data. A) Transcripts per million (TPM) values for RNA-seq libraries for each cell type were averaged and the top one thousand genes extracted. Each row represents a cell type and the intensity of the color represents the density of genes at the position on the proportional expression spectra. B) Ternary plots of the cell types where the three axes represent X1, X2, and Xins proportional expression.

Additional File 7. PDF format.

Summary of ChIP-seq mapping data from all available planarian ChIP seq data, demonstrating the improved data yield from the methods developed in the current study.

Additional File 8. PDF format.

Genome wide ChIP signal presented for all genes in each proportional gene expression category, average profiles are presented above each genome wide plot.

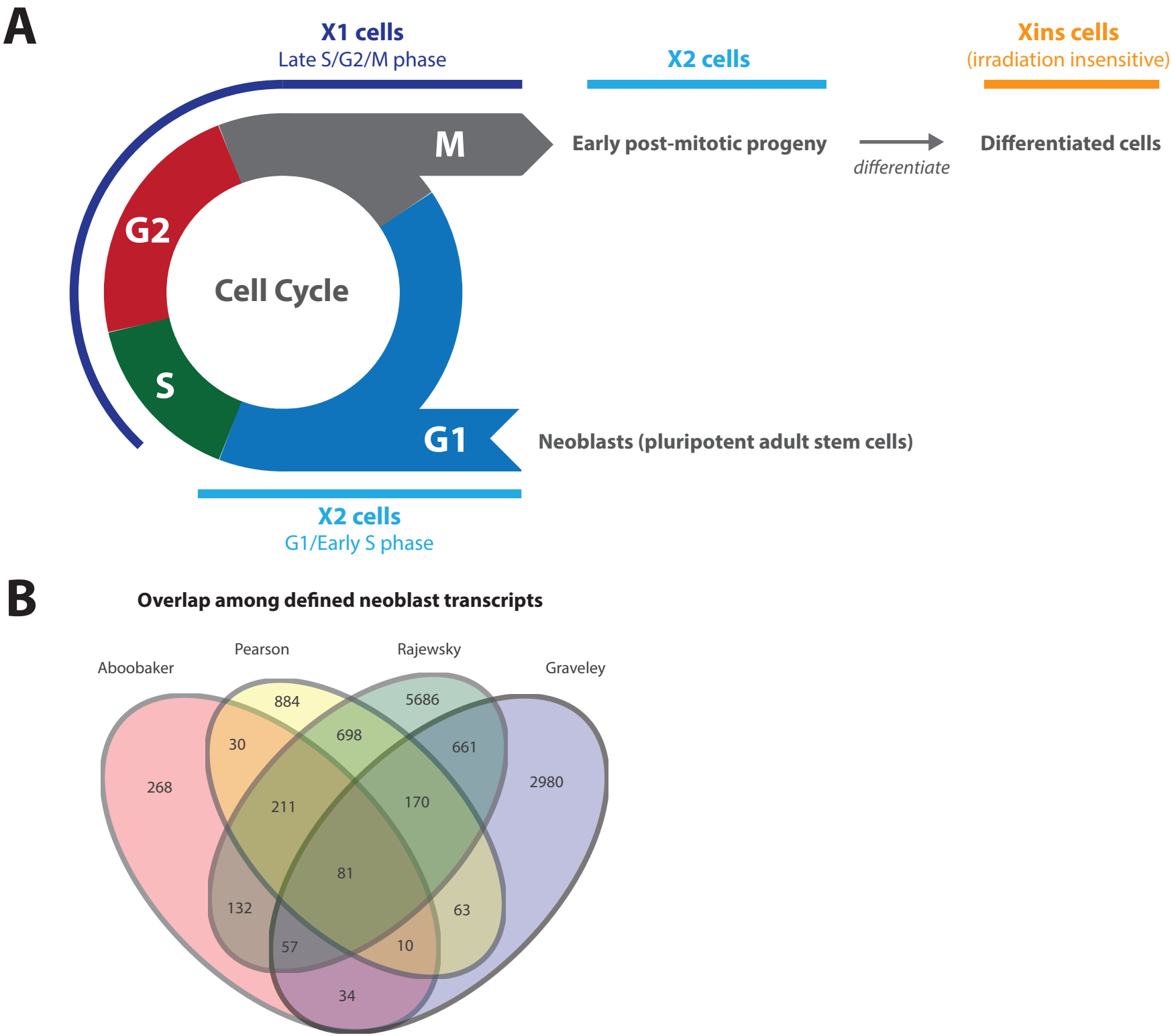
Additional File 9. HTML format.

1009 Jupyter notebook of all analysis performed.

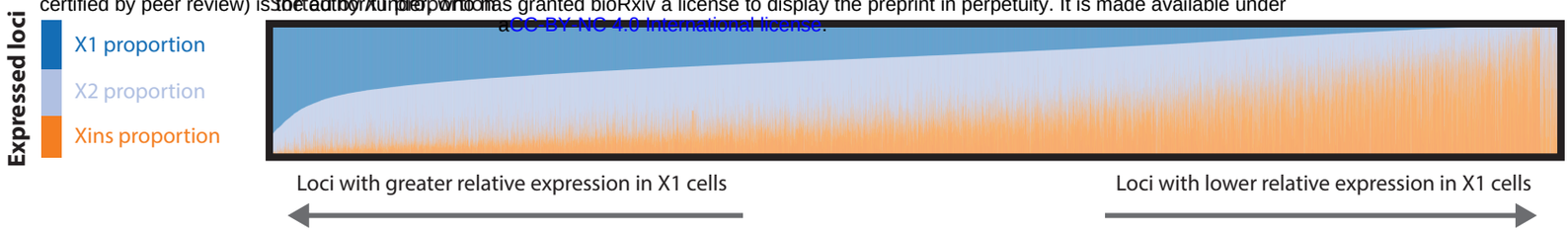
1010

1011 Addition File 10. ZIP format.

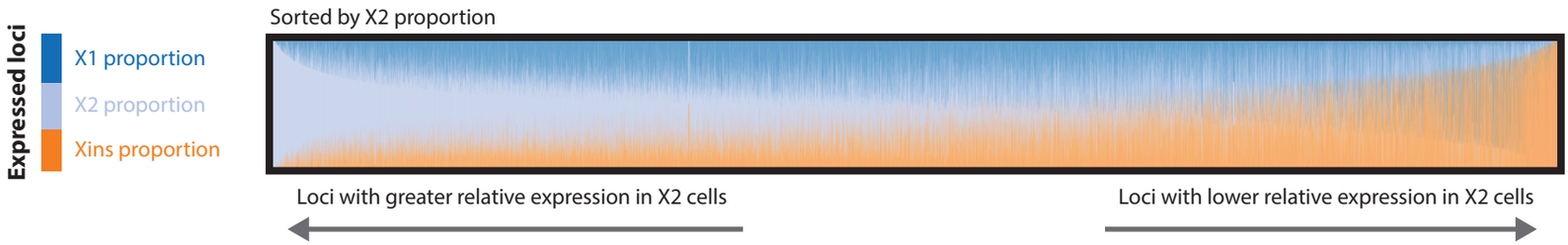
1012 GTF annotation file of the asexual genome


A
X1 cells
Late S/G2/M phase
X2 cells
Xins cells
(irradiation insensitive)
M
G2
Cell Cycle
S
G1
Early post-mitotic progeny
differentiate
Differentiated cells
Neoblasts (pluripotent adult stem cells)
X2 cells
G1/Early S phase
B
Overlap among defined neoblast transcripts
Aboobaker
Pearson
Rajewsky
Graveley
884
5686
30
698
661
268
2980
211
170
81
132
63
57
10
34

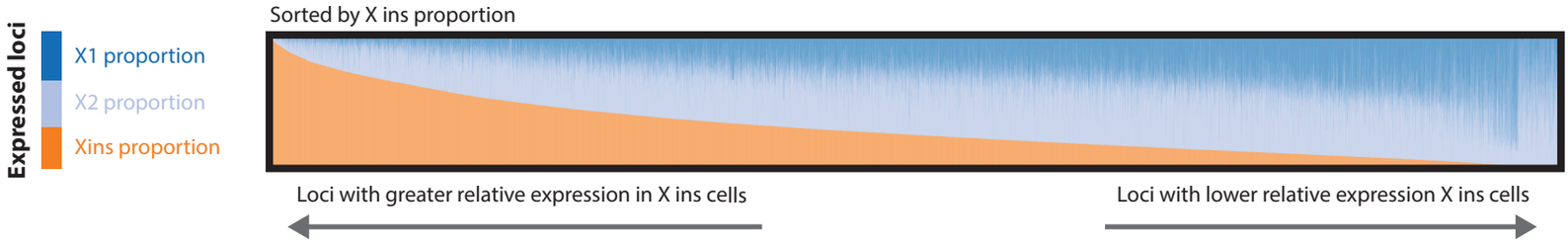
A



B



C



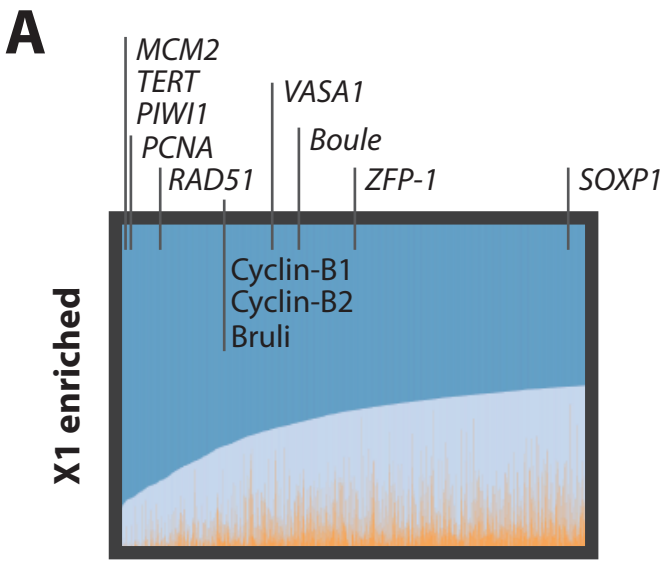
D

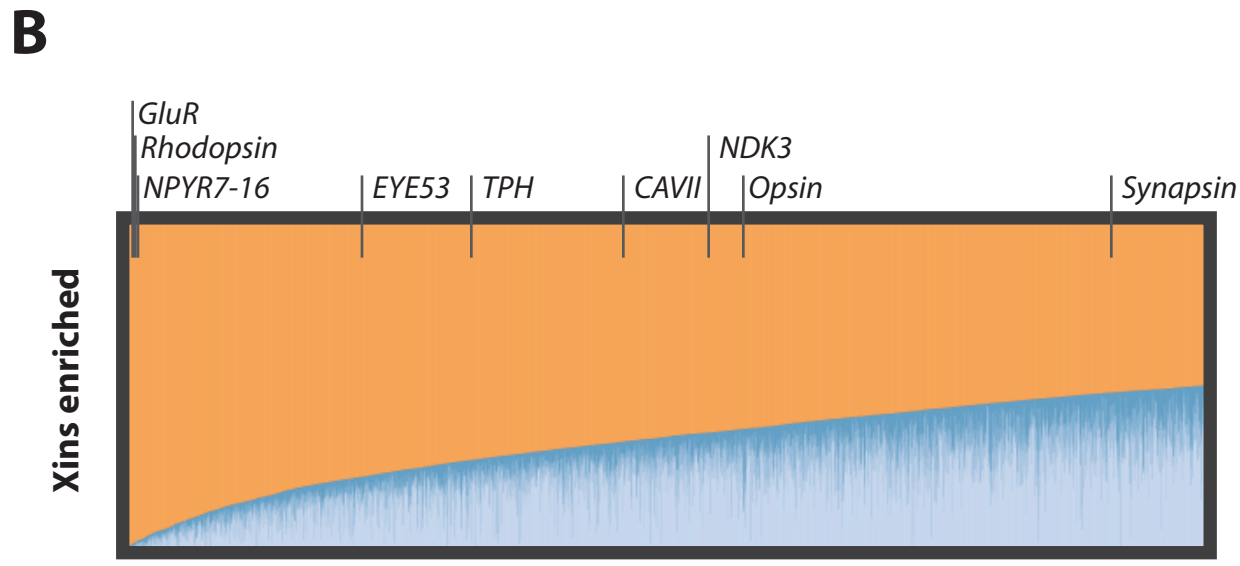
38,771 loci total

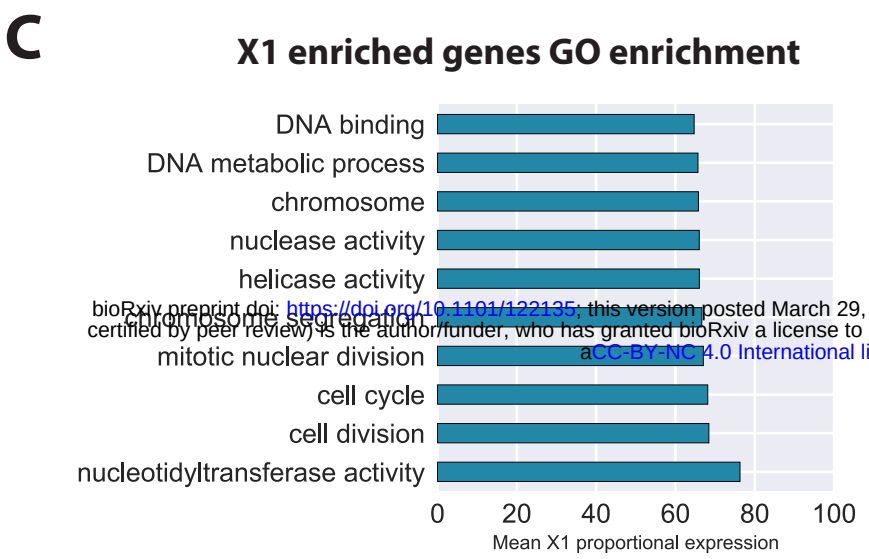
Categories defined by FACS RNA-seq data

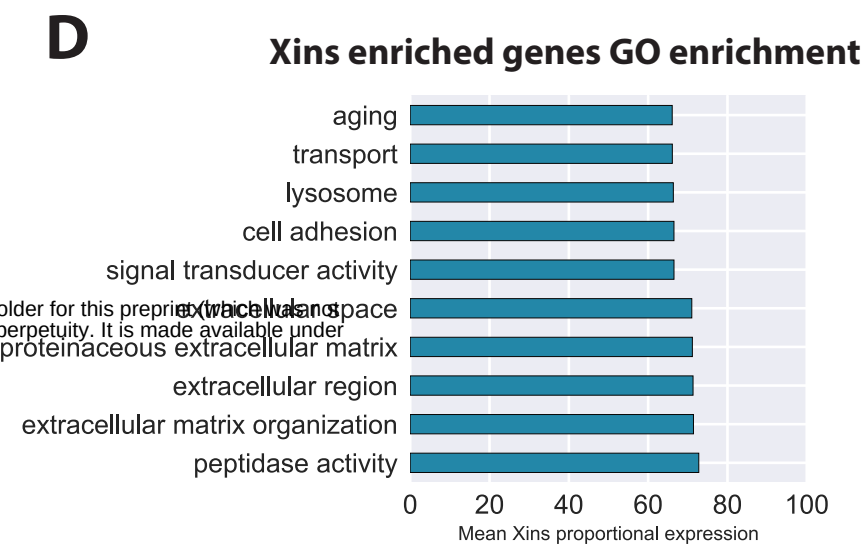
Category	Criteria	# Loci	# Coding* loci (% of category)
X1 enriched	X1 proprtional expression >= 50%	2,253	1,544 (68%)
X2 enriched	X2 proprtional expression >= 50%	8,444	4,781 (57%)
Xins enriched	Xins proprtional expression >= 50%	5,119	3,887 (76%)
X1/X2 enriched	X1 + X2 proprtional expression >= 75% and not in X1 enriched nor X2 enriched	4,538	3,107 (68%)
X2/Xins enriched	X2 + Xins proprtional expression >= 75% and not in X2 enriched nor Xins enriched	3,652	2,688 (74%)
X1/xins enriched	X2 + Xins proprtional expression >= 75% and not in X2 enriched nor Xins enriched	303	0 (0%)
unenriched	Remaining loci with roughly equal proportions among X1, X2, and Xins	2,897	2,003 (69%)
unclassified	Loci with # of reads less than 10 in all FACS RNA-seq libraries	11,565	3,762 (33%)

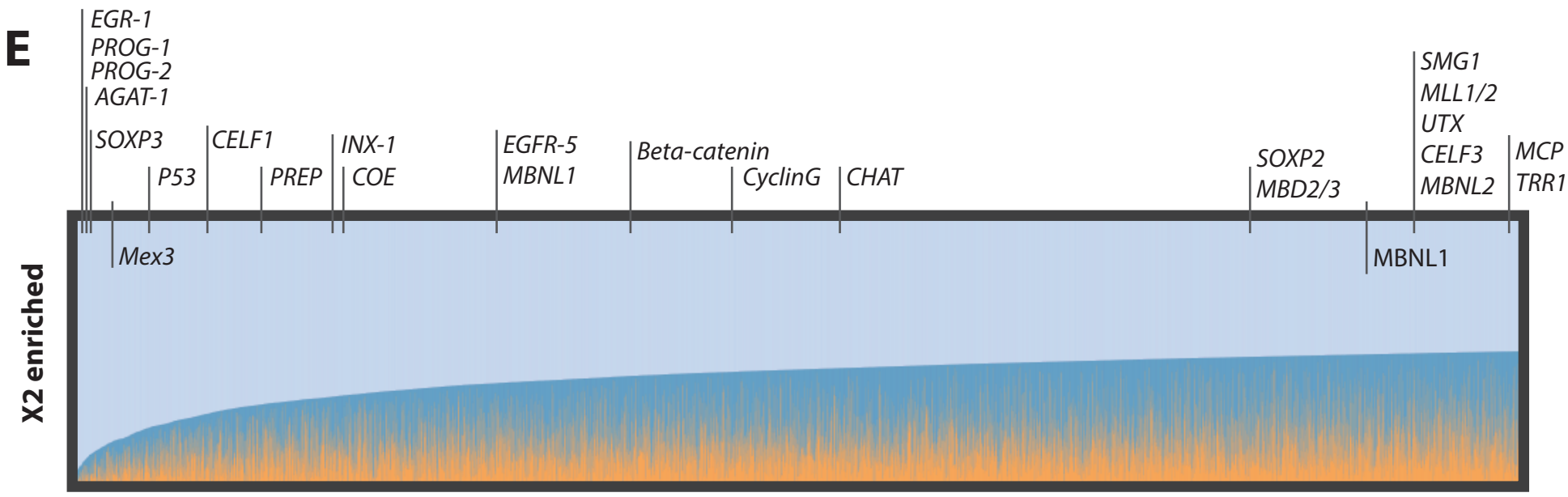
* Coding is defined by having TransDecoder evidence which includes homology (Uniprot, PFAM), hexamer frequency, and ORF length

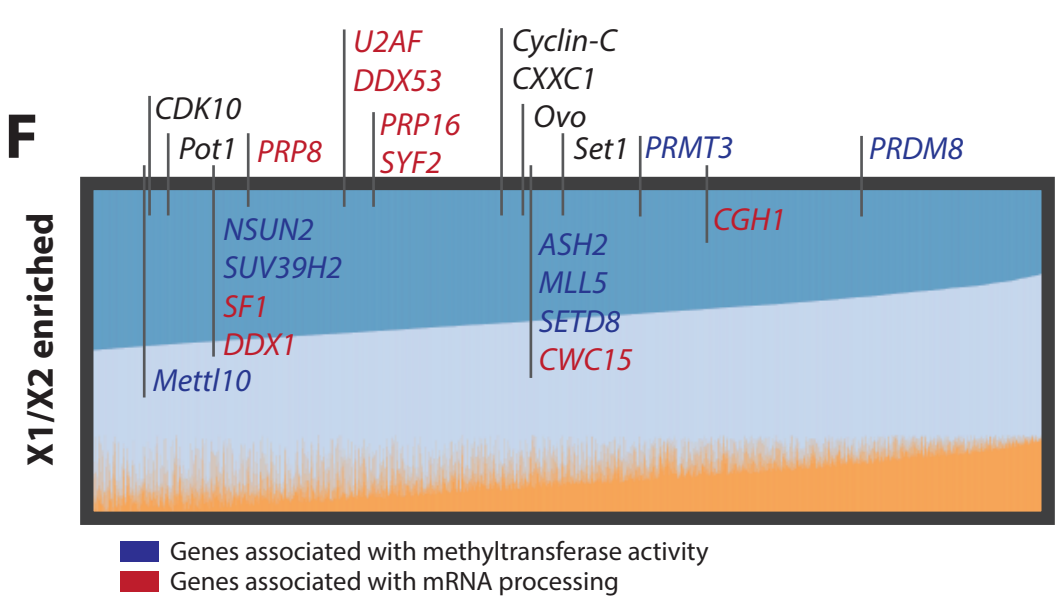


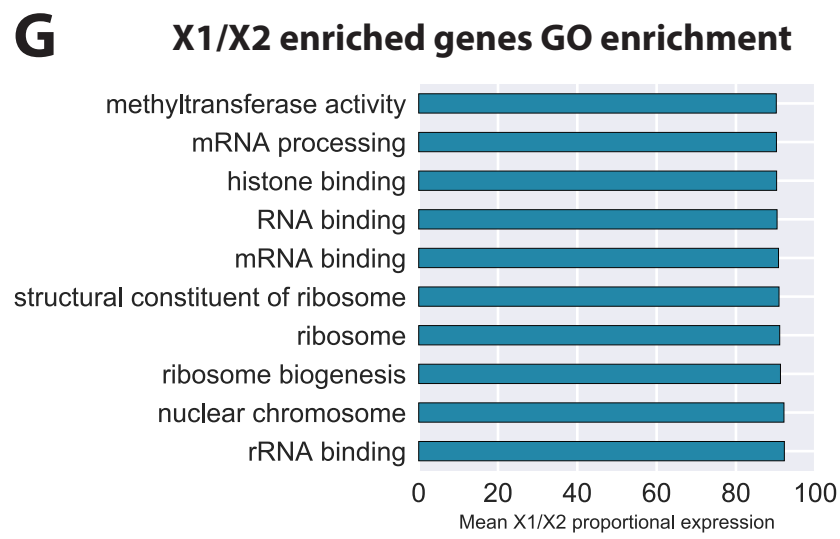


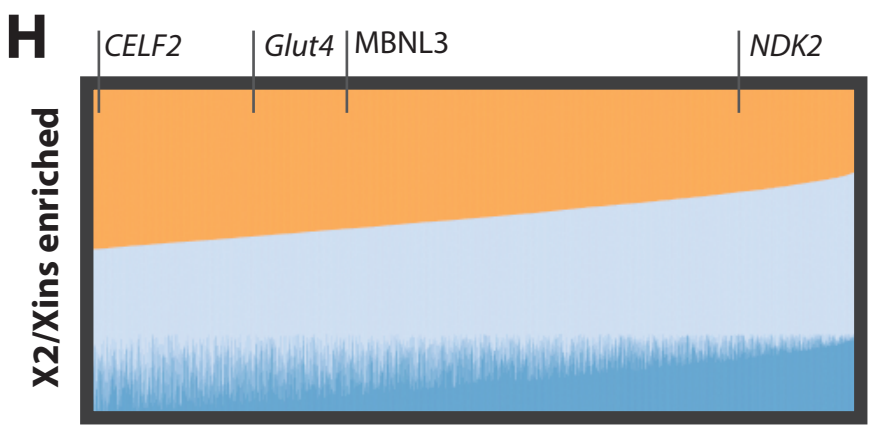


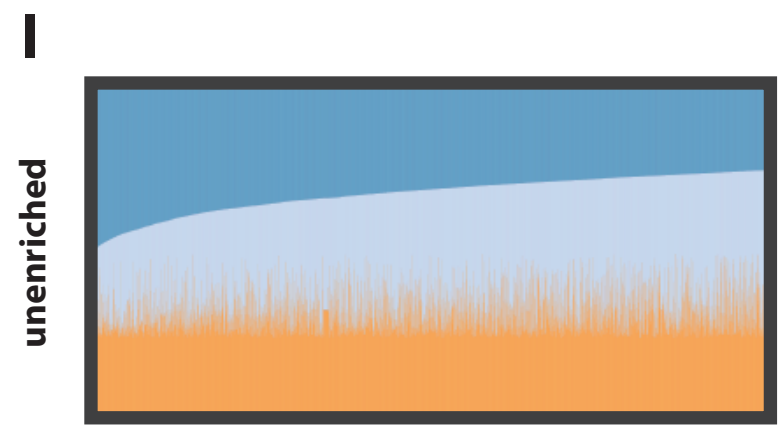




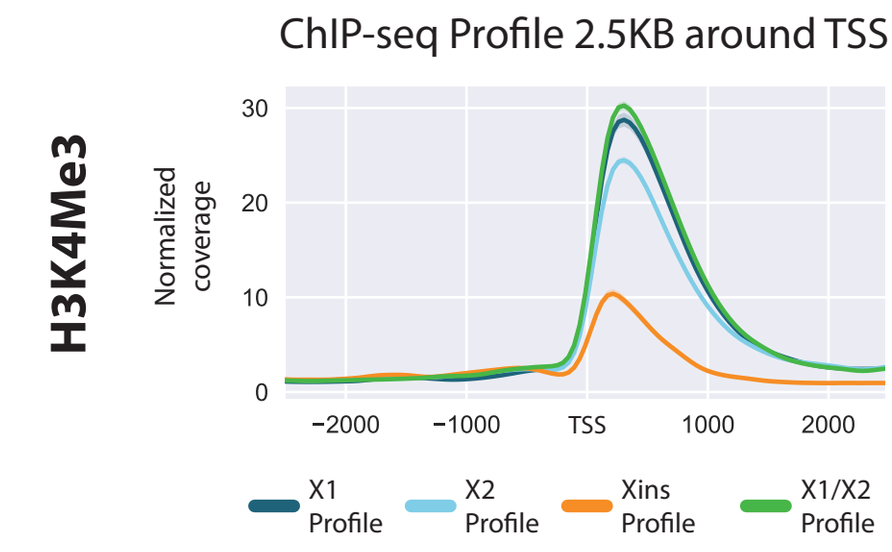




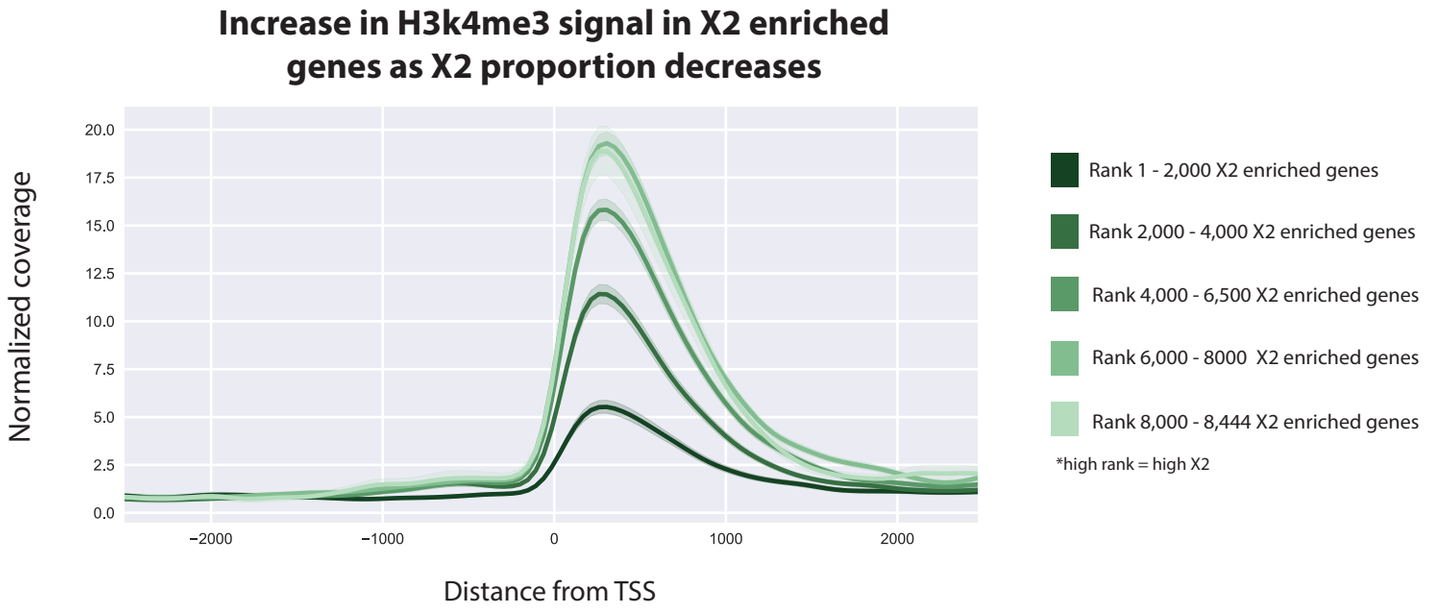




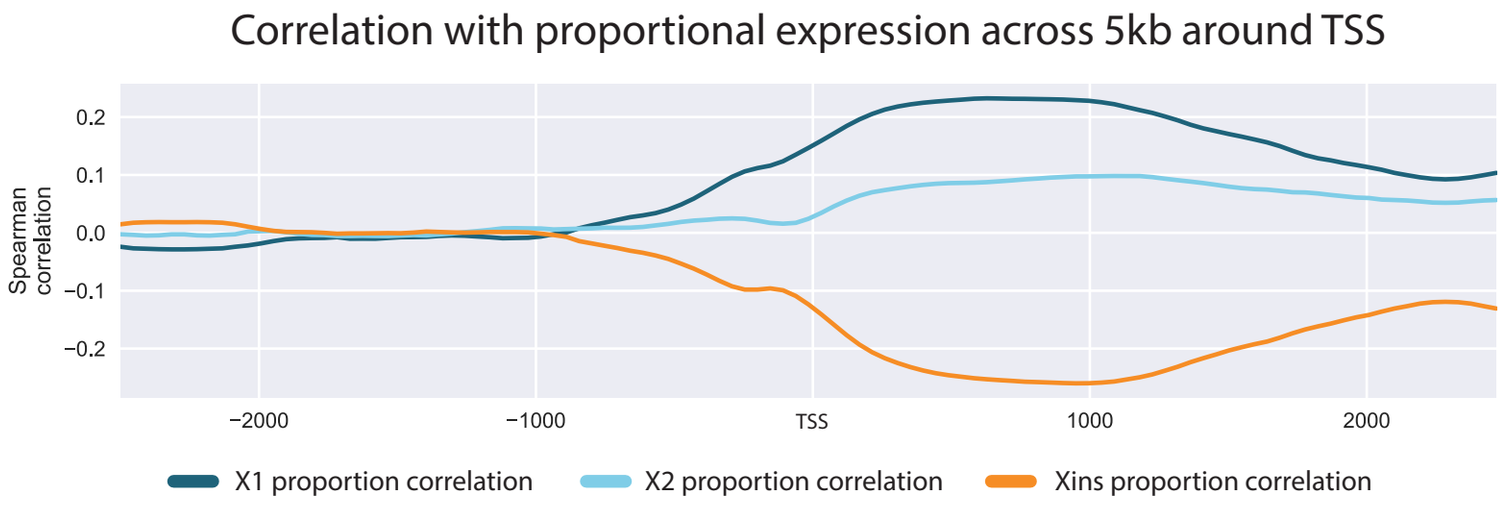
A



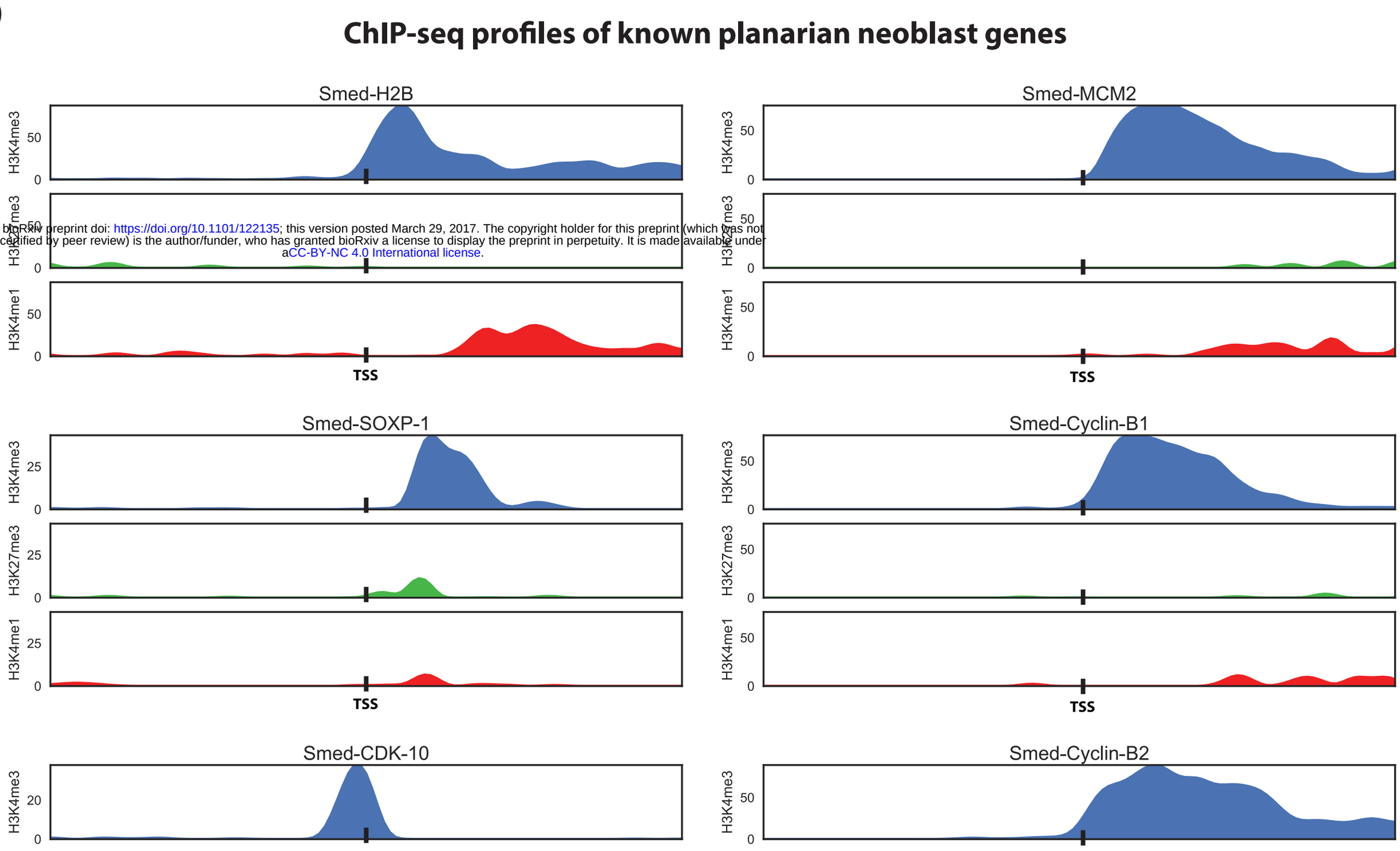
B

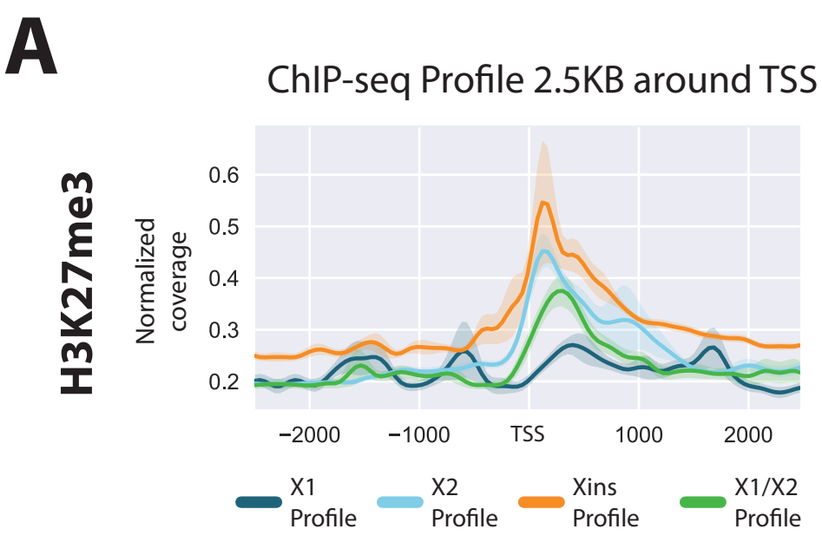


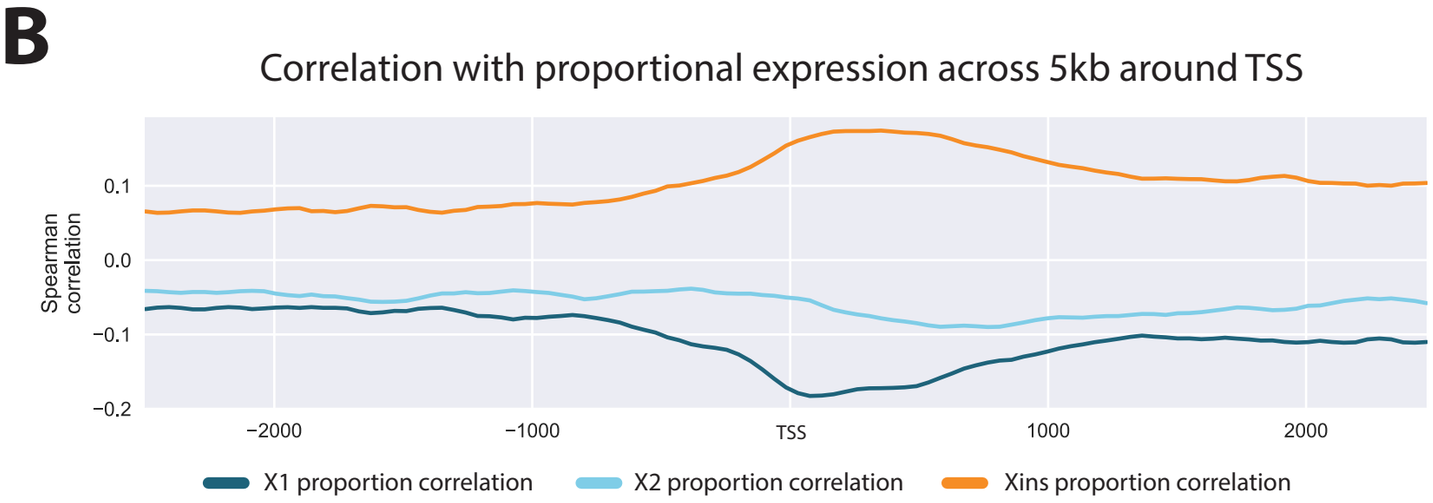
C

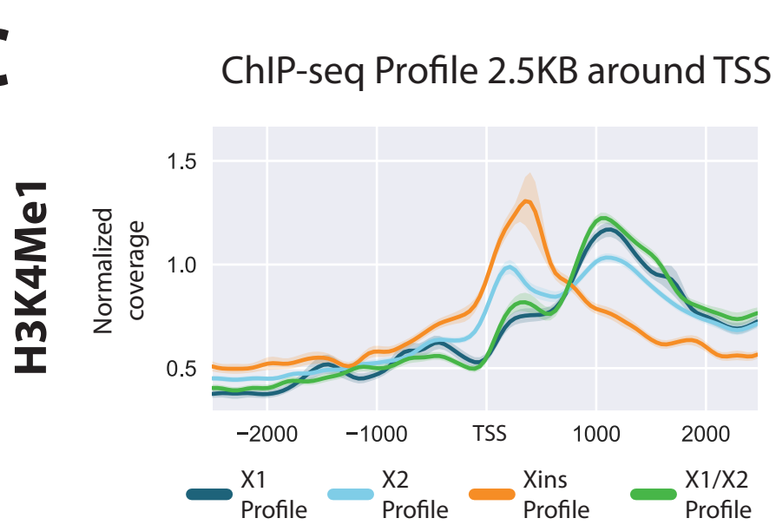


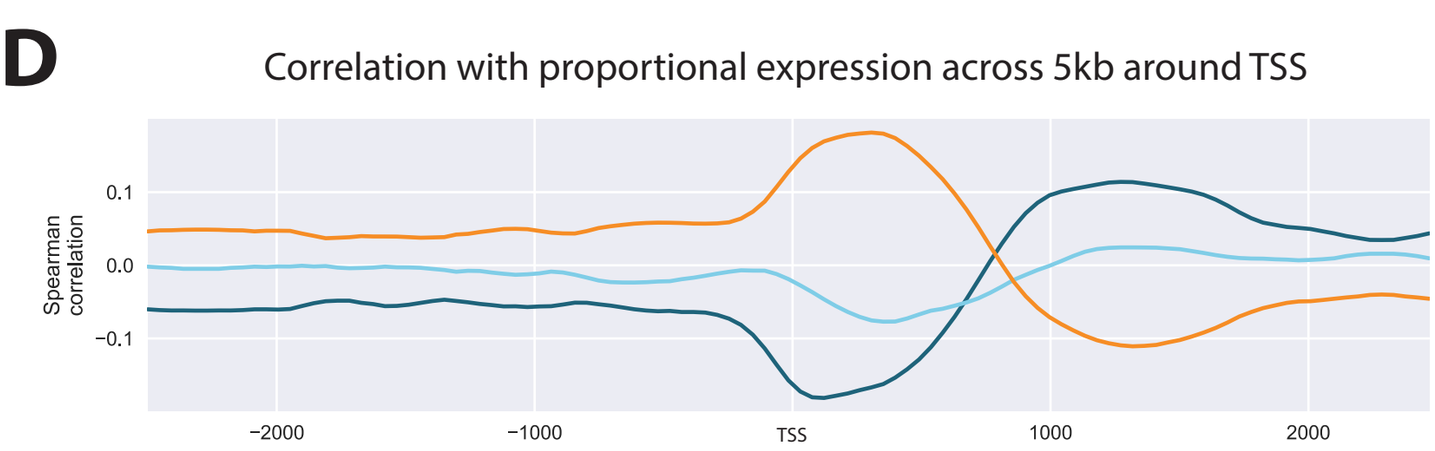
D

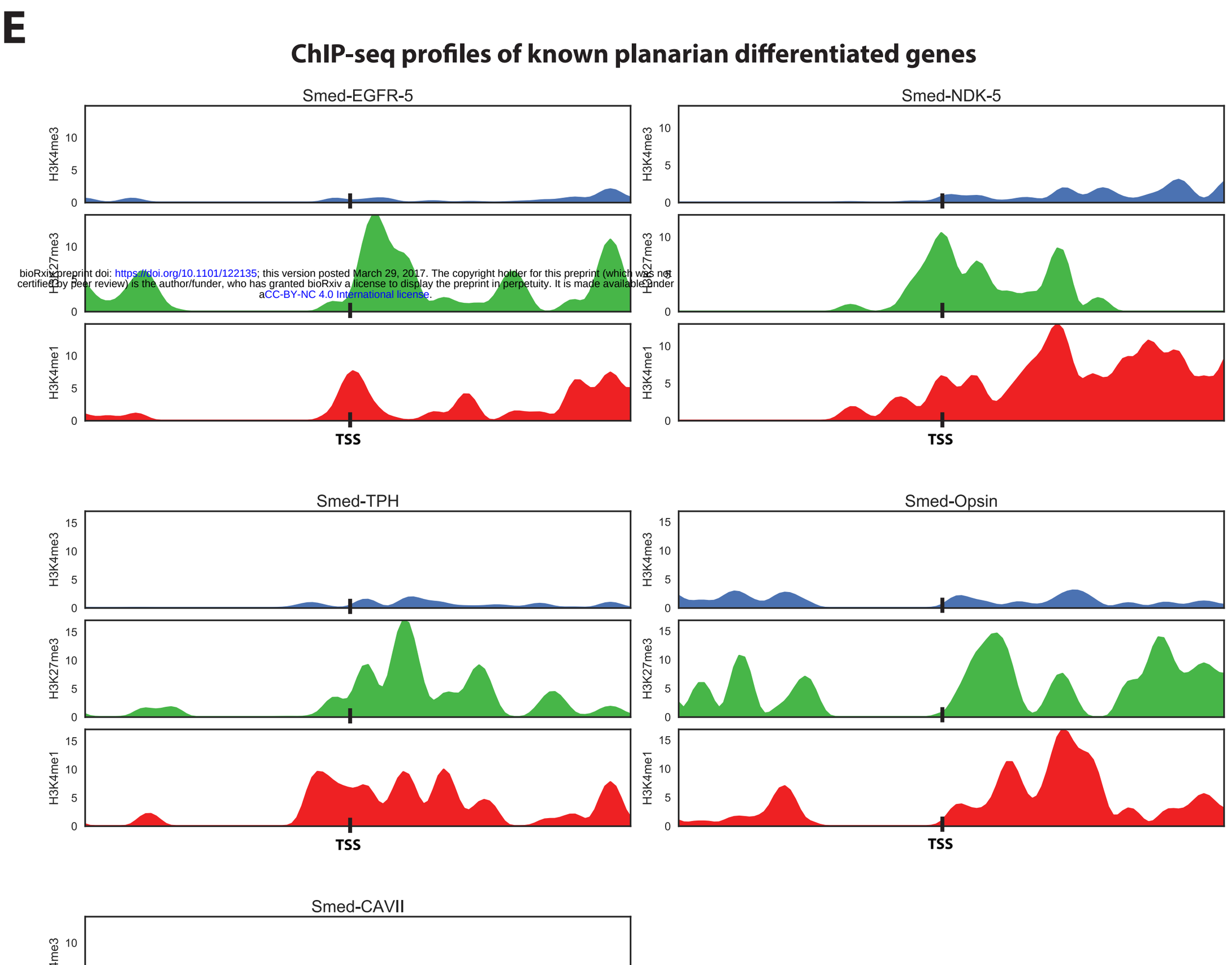




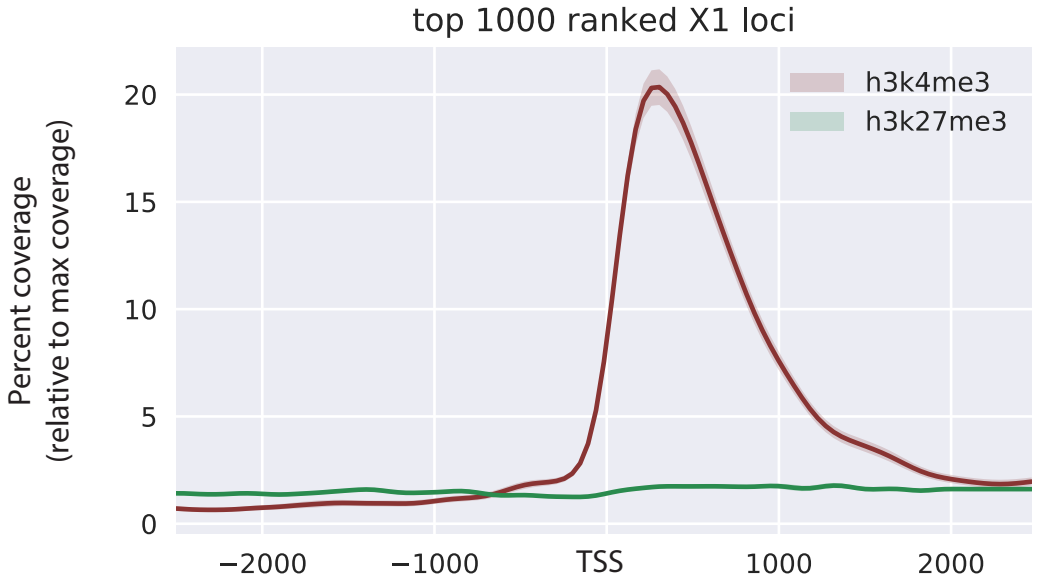




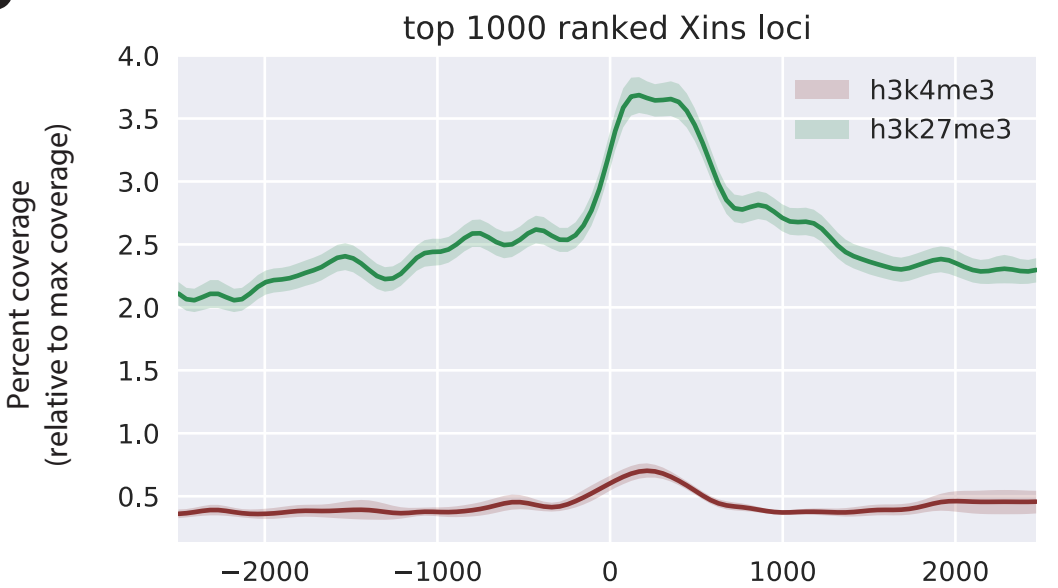




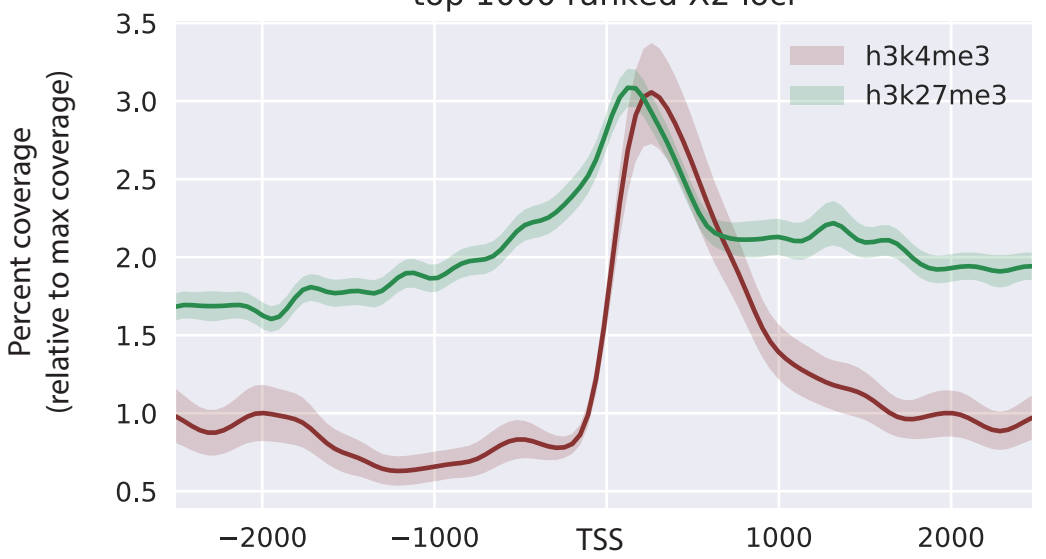
A



B

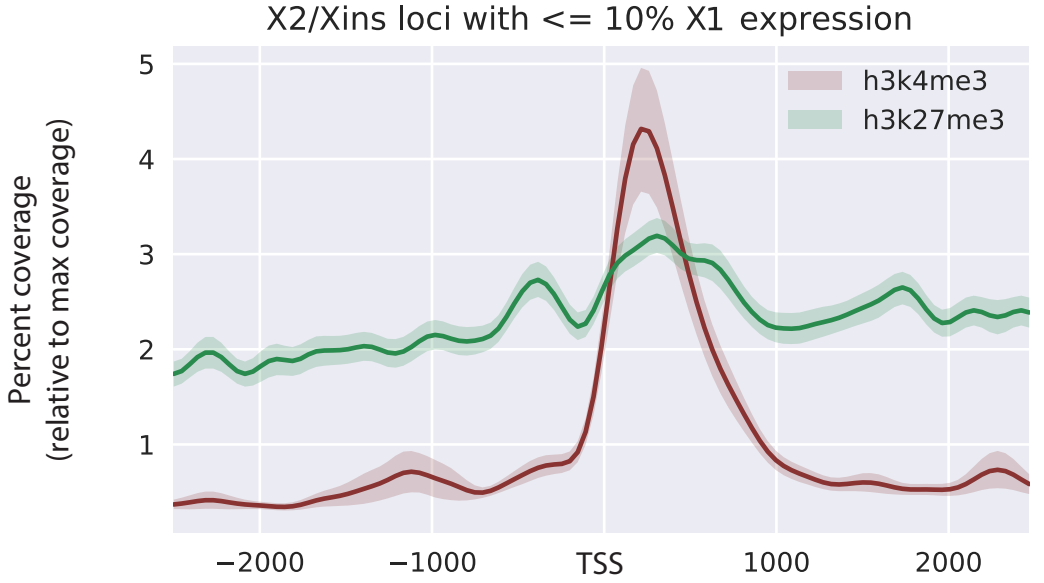


C

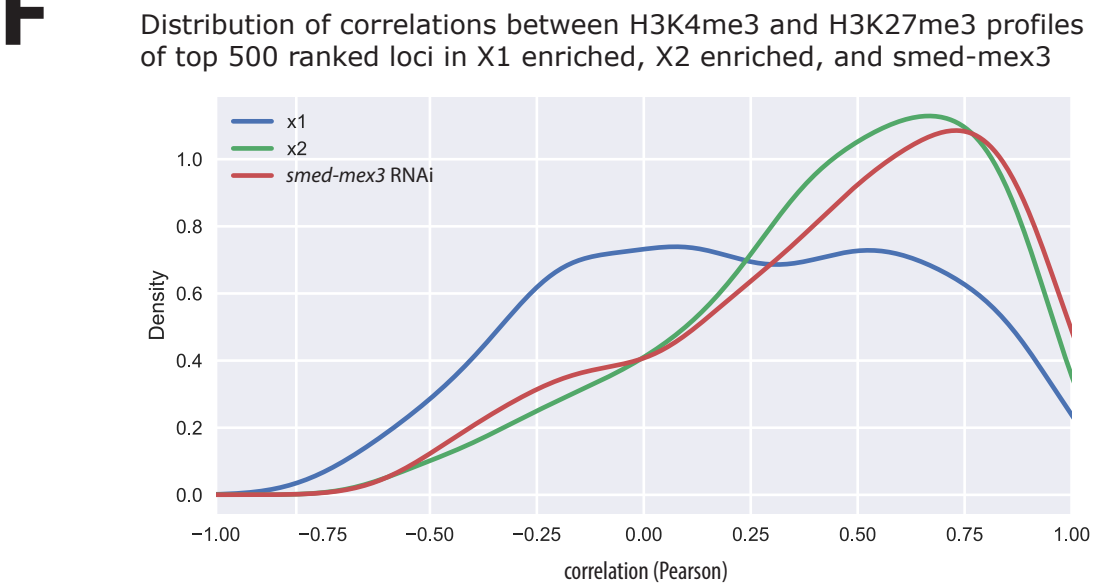


D

E



F



ChIP-seq profiles of high ranked X2 loci