Anastasiadi *et al. Epigenetics & Chromatin (2018) 11:37*
https://doi.org/10.1186/s13072-018-0205-1

Epigenetics & Chromatin

RESEARCH

Open Access

Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species

Dafni Anastasiadi[1], Anna Esteve-Codina[2,3] and Francesc Piferrer[1*]

Abstract

Background: DNA methylation is one of the main epigenetic mechanisms for the regulation of gene expression in eukaryotes. In the standard model, methylation in gene promoters has received the most attention since it is generally associated with transcriptional silencing. Nevertheless, recent studies in human tissues reveal that methylation of the region downstream of the transcription start site is highly informative of gene expression. Also, in some cell types and specific genes it has been found that methylation of the first intron, a gene feature typically rich in enhancers, is linked with gene expression. However, a genome-wide, tissue-independent, systematic comparative analysis of the relationship between DNA methylation in the first intron and gene expression across vertebrates has not been explored yet.

Results: The most important findings of this study are: (1) using different tissues from a modern fish, we show a clear genome-wide, tissue-independent quasi-linear inverse relationship between DNA methylation of the first intron and gene expression. (2) This relationship is conserved across vertebrates, since it is also present in the genomes of a model pufferfish, a model frog and different human tissues. Among the gene features, tissues and species interrogated, the first intron's negative correlation with the gene expression was most consistent. (3) We identified more tissue-specific differentially methylated regions (tDMRs) in the first intron than in any other gene feature. These tDMRs have positive or negative correlation with gene expression, indicative of distinct mechanisms of tissue-specific regulation. (4) Lastly, we identified CpGs in transcription factor binding motifs, enriched in the first intron, the methylation of which tended to increase with the distance from the first exon–first intron boundary, with a concomitant decrease in gene expression.

Conclusions: Our integrative analysis clearly reveals the important and conserved role of the methylation level of the first intron and its inverse association with gene expression regardless of tissue and species. These findings not only contribute to our basic understanding of the epigenetic regulation of gene expression but also identify the first intron as an informative gene feature regarding the relationship between DNA methylation and gene expression where future studies should be focused.

Keywords: DNA methylation, Gene expression, First intron, Regulation, Gene features

*Correspondence: piferrer@icm.csic.es
[1] Institute of Marine Sciences (ICM-CSIC), Passeig Marítim de la Barceloneta, 37-49, 08003 Barcelona, Spain
Full list of author information is available at the end of the article

Background

DNA methylation is one of the main epigenetic mechanisms for the regulation of gene expression [1]. Under the so-called standard model of gene expression regulation, methylation of cytosine–guanine dinucleotides (CpGs) in the promoter regions of genes has received the most attention since it is generally associated with repression of transcription, either directly, by blocking the access of transcription factors (TFs), or indirectly, by recruiting other repressive proteins with methyl-binding domains [2, 3]. Regions rich in CpGs that typically span 200–1000 bp are called CpG islands (CGI), usually remain unmethylated, overlap with gene promoters and are associated with gene transcription regulation [4, 5].

Nevertheless, recent studies in human tissues reveal that methylation of the region downstream of the transcription start site (TSS) is highly informative of gene expression.

Thus, in addition to promoters, enhancers also bind TFs, interact with the promoter, and exhibit widespread hypo-methylation during development [6] and dynamic changes during oncologic transformation [7, 8]. Also, studies using mammalian cells have shown differences in methylation levels between the first exon and the rest of exons and, further, gene expression levels are better inversely correlated with the methylation of the first exon than with that of the promoter [9]. Furthermore, between gene body methylation and gene expression, a positive correlation has been demonstrated [10]. These studies suggest that DNA methylation of distal or intragenic regulatory elements with different degrees of CpG density are involved in the regulation of gene expression and that DNA methylation has dual roles, both inhibitory and permissive, depending on the genomic region.

Differences in the contribution of DNA methylation to gene expression regulation among distinct genomic features are also evident in the so-called tissue-specific differentially methylated regions (tDMRs), which are located both upstream and downstream of the transcription start site [11]. These tDMRs contain binding sites for different TFs and overlap with regions of variable CpG density, and although their hypo-methylation is thought to be related to tissue-specific functions, they can also exhibit positive or negative correlation with gene expression levels [11, 12].

Recent comparative epigenomic studies using non-model organisms have shown that epigenetic divergence follows the genetic phylogenetic patterns across species [13, 14]. Thus, across vertebrates there are global differences in the methylation content of warm-blooded versus cold-blooded species [15]. Research in epigenetics of non-model vertebrates including fish [16–22], birds [23, 24] and mammals [13, 25–30] is generally undertaken with the main objective to correlate DNA methylation patterns with a specific phenotypic trait. However, a genome-wide, tissue-independent, systematic comparative analysis of the relationship between DNA methylation in defined and distinct genomic features and gene expression across vertebrates has not been explored yet.

To address these questions, here we used the European sea bass (*Dicentrarchus labrax*), a modern teleost and one of the fish species with more genomic resources available [31, 32]. To account for the possible influence of cellular diversity when compared to cell lines, we selected the muscle, where myocytes clearly dominate, hence a tissue of very low cellular diversity, and the adult testis, where different types of somatic and germ cells coexist, thus a tissue of high cellular diversity. We constructed reduced representation bisulfite sequencing (RRBS) libraries to measure the genome-wide DNA methylation and RNA-seq libraries to measure gene expression levels. We determined the relationship between DNA methylation and transcriptomic profiles in different genomic features including not only promoters but also introns and exons. We found that a clear inverse correlation between DNA methylation and gene expression is present in the first intron. Results were contrasted not only with results obtained in mammals but also with those obtained in other vertebrates, including the model fish *Tetraodon nigroviridis* (pufferfish) and model frog *Xenopus tropicalis*. Then, we investigated the functional properties of this relationship and we identified CpGs in TF-binding motifs enriched in the first intron, which were close to the beginning of first intron and were indicators of gene expression. Lastly, we detected tDMRs between the two tissues of different transcriptomic complexity which correlated with gene expression.

Results

The relation of DNA methylation and gene expression depends on the gene feature

In whole genes, DNA methylation patterns followed a bimodal distribution, with high (> 80%) or low (~ 10%) levels of DNA methylation in the majority of CpG sites in both muscle (Fig. 1a) and testis (Fig. 1b). Separating the whole gene in specific gene features exposed distinct patterns. A similar bimodal DNA methylation pattern was observed in introns and exons. However, in promoters, most of CpG sites were unmethylated. In addition, by partitioning data from exons into first exon and the rest of exons, a contrasting pattern was revealed. The majority of unmethylated cytosines were restricted to the first exon and methylated cytosines almost exclusively localized in the rest of exons, while the peak of unmethylated cytosines was sharper in the first exon than in the promoter. Likewise, partitioning the introns showed a

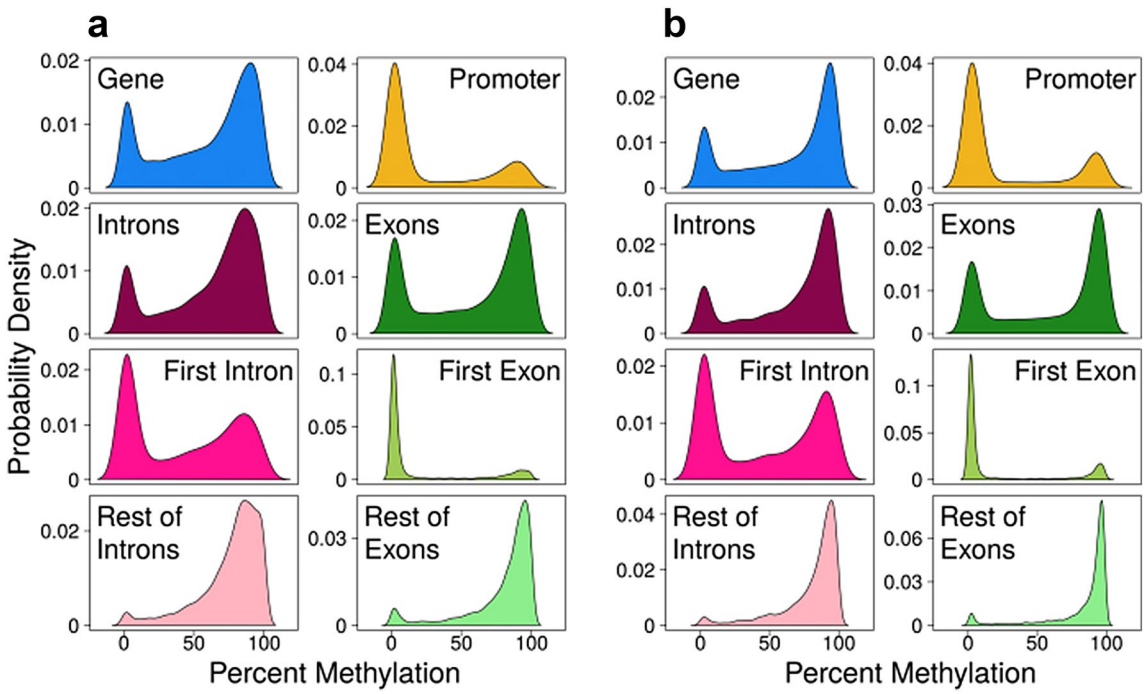

Fig. 1 DNA methylation per gene in gene features in muscle (**a**) and in testis (**b**). Kernel density plots for DNA methylation in genes ($n = 15{,}456$), promoters (− 1000 bp from the transcription start site; $n = 5034$), all introns ($n = 9184$) and all exons ($n = 12{,}317$). Separation of exons in first exon ($n = 5790$) and rest of exons ($n = 8798$) and of introns in first intron ($n = 4387$) and rest of introns ($n = 5646$)

majority of highly methylated cytosines in all except the first intron. In the first intron, the distribution was still bimodal but skewed toward the unmethylated sites and smoother than in the first exon and the promoter (Fig. 1a, b). The distribution of DNA methylation in specific gene features was similar in liver and spleen, for which one RRBS library per tissue was also constructed (Additional file 1: Fig. S1). The RRBS libraries for liver and spleen were constructed for a preliminary study and were not further analyzed since no biological replicates were sequenced. Thus, regardless of tissue and cellular diversity, the majority of CpG sites were unmethylated in the promoter and first exon and, to a lesser degree, also in the first intron.

In order to relate the gene expression levels with the DNA methylation levels of specific gene features, we divided the gene expression levels in deciles based on the increasing distribution of $\log_2$-transformed copy million number (cpm) values. In muscle, median DNA methylation levels were low regardless of gene expression in promoter and first exon (Fig. 2). In the first exon, there was also a weak but significant negative correlation of DNA methylation with gene expression (Spearman's rank correlation coefficient $[\rho] = -0.08$, p value < 0.001). By contrast, in the first intron DNA methylation levels decreased with increasing expression levels (Fig. 2) and there was the strongest among gene features negative correlation of DNA methylation with gene expression ($\rho = -0.15$, $p < 0.001$). In the rest of exons and introns, DNA methylation levels were high independently of gene expression and there were no significant correlations of DNA methylation with gene expression.

In testis, in the promoter and first exon, median DNA methylation levels were also low in all expression deciles (Fig. 2). However, in the genes belonging to the first and second expression deciles, there was significantly more variance in the DNA methylation levels in comparison with the muscle (ANOVA on residuals followed by Tukey's HSD; p adjusted < 0.001) and also compared to the rest of expression deciles ($p < 0.001$). In the testis, significant negative correlations of DNA methylation with gene expression were evident in the promoter ($\rho = -0.19$; $p < 0.001$), first exon ($\rho = -0.27$; $p < 0.001$) and first intron ($\rho = -0.25$; $p < 0.001$). The negative correlation in the first intron was stronger than the one in the promoter, similarly to what was observed in muscle, but slightly weaker than in the first exon, in contrast to what was observed in muscle. In the rest of exons and in the rest of introns, median DNA methylation levels were

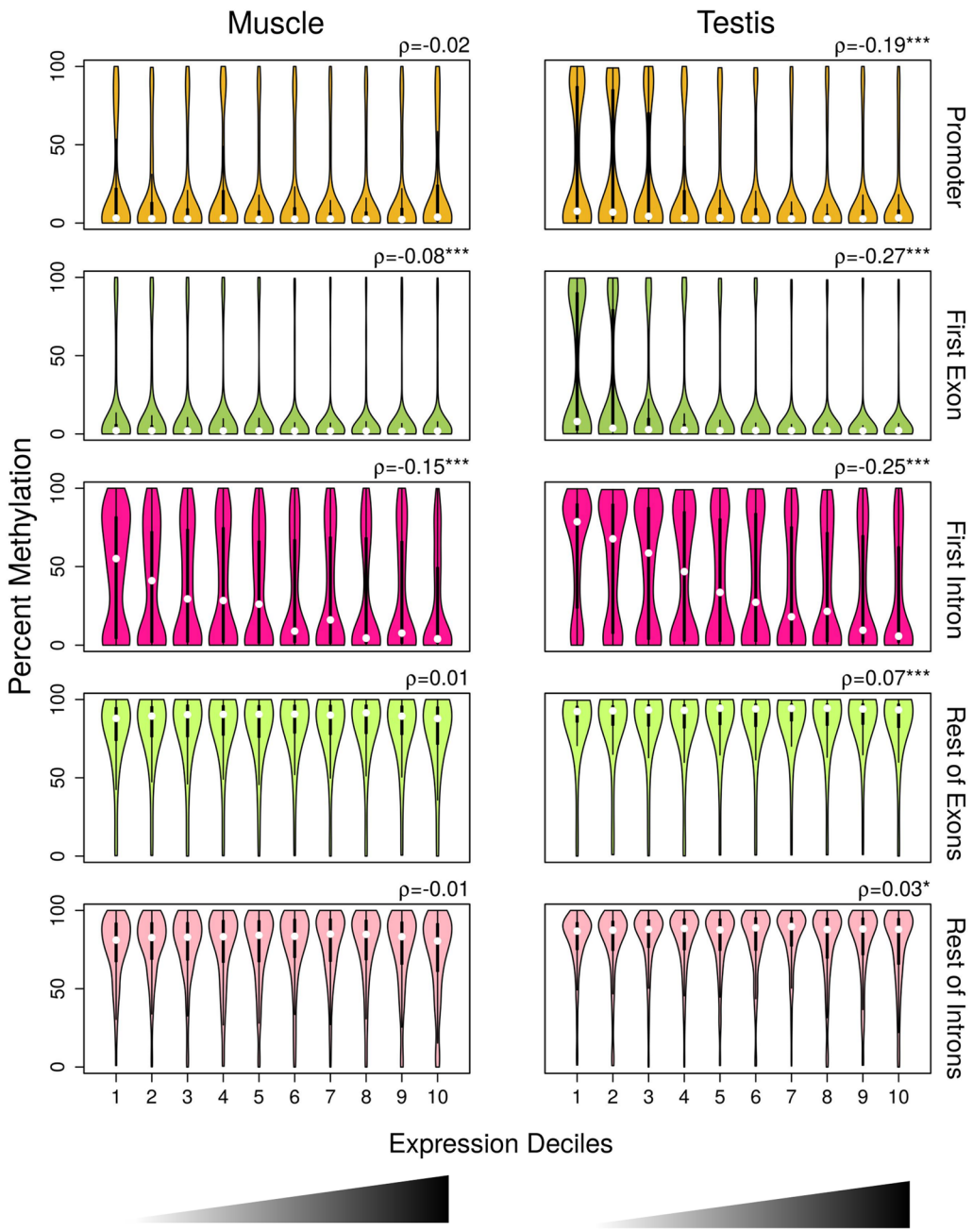

Fig. 2 DNA methylation in gene features by expression deciles in muscle and in testis. Violin plots of DNA methylation in promoter (muscle, $n = 2745$; testis, $n = 3345$), first exon (muscle, $n = 3537$; testis, $n = 4064$), first intron (muscle, $n = 2801$; testis, $n = 3122$), rest of exons (muscle, $n = 5523$; testis, $n = 6398$) and rest of introns (muscle, $n = 4043$; testis, $n = 4897$) divided into deciles based on increasing ranking of gene expression measured as $\log_2$-transformed count per million (cpm) values. Box plots with rotated kernel density plots at both sides indicate the interquartile range, and white central dots the median of the distribution. Correlations between DNA methylation and gene expression were measured using Spearman's rank correlation coefficient (ρ), and the significance levels are reported as follows: $^*p < 0.05$; $^{***}p < 0.001$

high regardless of gene expression levels and the correlations were weakly positive, although significant ($p < 0.05$). Thus, gene expression was clearly inversely correlated with DNA methylation levels across the two tissues only in the first intron.

This inverse relationship became more evident when we only considered genes at the extremes of the expression range in both tissues. For example, genes with low expression (members of the first and second expression deciles) or with high expression (members of the

ninth and tenth expression deciles) exhibited similar and clearer patterns of DNA methylation (Additional file 1: Fig. S2).

The inverse relationship of DNA methylation and gene expression is present in other vertebrate genomes

In order to investigate this inverse correlation in other vertebrate species, we, then, used whole-genome bisulfite sequencing (WGBS) and RNA-seq data from a pufferfish, whole *Tetraodon nigroviridis* [20] (NCBI's Gene Expression Omnibus [33]; GEO with accession number GSE19824), from a frog, *Xenopus tropicalis* [34] (GEO with accession number GSE67974) and from human liver and lung [35] (GEO with accession number GSE70091). WGBS data of *Xenopus* were obtained from gastrula stage 10.5 and RNA-seq data from gastrula stage 11, therefore during development. In the promoters and first exons of these species, DNA methylation showed a decreasing relationship with gene expression in the first three to four deciles and then remained low in the deciles of higher expression (Fig. 3). The correlation of gene expression with DNA methylation was negative in all cases. In the first intron, DNA methylation was decreasing with the expression decile in all vertebrate datasets tested. In contrast, in an invertebrate species, *Ciona intestinalis*, [20] (GEO with accession number GSE19824), the correlations were always positive and significant in promoters, first exons and first introns (Additional file 1: Fig. S3).

The DNA methylation of TF-binding motifs located at the beginning of the first intron is informative of gene expression

Next, we focused exclusively on the first intron and searched for potential enrichment of specific TF-binding motifs associated with the identified negative correlation of gene expression with DNA methylation. This analysis was performed using sequences of ±50 bp from the CpGs with methylation values in the first introns of expressed genes, i.e., the genes of Fig. 2. This distance was chosen to encompass the maximum TF-binding motif length (31 nucleotides [36]) and an arbitrary +20 nucleotides more. The objective was to identify TF-binding motifs that were independent of the tissue under question; therefore, after performing the enrichment compared to input shuffled sequences, we selected only the 10 TF-binding motifs that were enriched in both muscle and testis (Additional file 1: Fig. S4).

The methylation status of the CpGs inside TF-binding motifs may directly affect the binding affinity. Therefore, we then focused on the 4 TFs, among the 10 enriched, in the binding motifs of which CpGs were present: CREB1, ZBTB33, ZBTB7A and E2F4. The first introns of muscle and testis were, then, screened for these 4 specific motifs, and the methylation status of the target CpGs was identified (Fig. 4). The CpGs were classified as unmethylated if their methylation was below the first quartile of the total distribution or as methylated if their methylation was above the third quartile of the total distribution for each tissue. The expression of genes with unmethylated CpGs in the target TF-binding motifs was significantly higher than the expression of genes with methylated CpGs in muscle (one-sided Wilcoxon rank sum test with continuity correction; $W=1432.5$, $p=0.0171$) and in testis ($W=1552$, $p=0.0003$). In addition, the unmethylated CpGs were located closer to the first exon–first intron boundary than the methylated CpGs in both muscle ($W=705$, $p=1.222^{-13}$) and testis ($W=738$, $p=3.887^{-12}$). Analysis of covariance revealed a significant effect of the interaction between relative distance of the CpG and methylation status on gene expression in both muscle ($F=6.264$, $p=0.013$; Table 1) and testis ($F=5.781$, $p=0.018$; Table 1).

The DNA methylation of the first intron associates with the upstream features and is independent of its length

In order to test whether there is association of the methylation state of two gene features, we calculated the odds ratio (OR) as representative of the odds that a gene feature A is also methylated when gene feature B is methylated. In both tissues, there was strong evidence for statistically significant association of the DNA methylation of the promoter, the first exon and the first intron (Fig. 5), since the 99.9% confidence intervals (CIs) were far from overlapping the value 1. The gene body methylation, including all exons and introns of a gene, showed strong association with the DNA methylation of all gene features tested, while the methylation of the rest of exons and the rest of introns was also associated. In testis, the methylation of the first intron was associated with the methylation of the rest of exons as well.

Since DNA methylation occurs in the CpG context, we wanted to exclude biases potentially affecting our results regarding the importance of the first intron for gene expression, mainly the CpG density in the gene features of interest and the length of the first intron. To address potential CpG density bias, we looked to the distribution of DNA methylation as a function of CpG density for the promoter, first exon and first intron, for all genes across the expression deciles. The CpG density was higher in the first exon, followed by the promoter and the first intron (pairwise Wilcoxon rank sum tests with continuity correction, $p<2.2^{-16}$ in all cases; Additional file 1: Fig. S5). The first intron showed a less dynamic range of CpG density and a more uniform distribution of the DNA methylation. To contemplate first intron length's bias, we

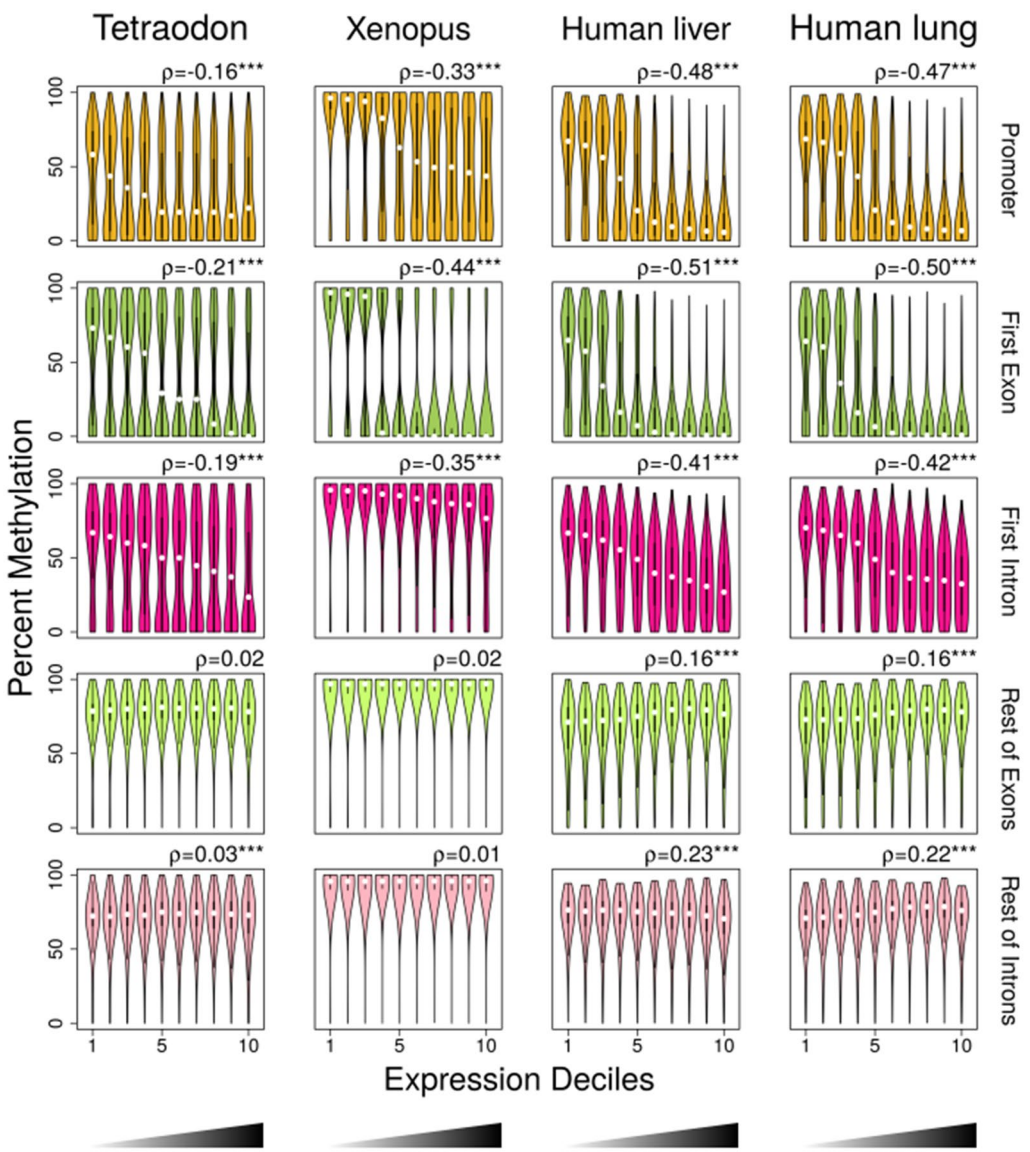

Fig. 3 Violin plots of DNA methylation in promoter (Tetraodon, $n = 12{,}896$; Xenopus, $n = 12{,}704$; human liver, $n = 22{,}680$; human lung, $n = 23{,}012$), first exon (Tetraodon, $n = 11{,}887$; Xenopus, $n = 10{,}361$; human liver, $n = 20{,}383$; human lung, $n = 20{,}704$), first intron (Tetraodon, $n = 11{,}420$; Xenopus, $n = 12{,}202$; human liver, $n = 20{,}029$; human lung, $n = 20{,}757$), rest of exons (Tetraodon, $n = 12{,}618$; Xenopus, $n = 12{,}662$; human liver, $n = 18{,}961$; human lung, $n = 19{,}331$) and rest of introns (Tetraodon, $n = 11{,}840$; Xenopus, $n = 11{,}905$; human liver, $n = 16{,}930$; human lung, $n = 17{,}007$) divided into deciles based on increasing ranking of gene expression. Box plots with rotated kernel density plots at both sides indicate the interquartile range, and white central dots the median of the distribution. Correlations between DNA methylation and gene expression were measured using Spearman's rank correlation coefficient (ρ), and the significance levels are reported as follows: $^{***}p < 0.001$

divided the dataset of genes in four quartiles according to the distribution of the length of their first intron. The correlation between DNA methylation and gene expression was negative independently of the length in both tissues, even in the fourth quartile of intron length that consisted of introns with median length of 18,746 bp in the muscle and 25,989 bp in the testis (Additional file 1: Table S1).

To further decipher the relationship of gene expression with DNA methylation in the first intron, we focused only on the extreme situations. Therefore, we selected the genes with the lowest (expression deciles 1 and 2) and the highest (deciles 9 and 10) expression and with mean DNA methylation below 10% or above 90% (Additional file 1: Fig. S6). The vast majority of the highest expressed genes in both tissues had DNA methylation below 10%

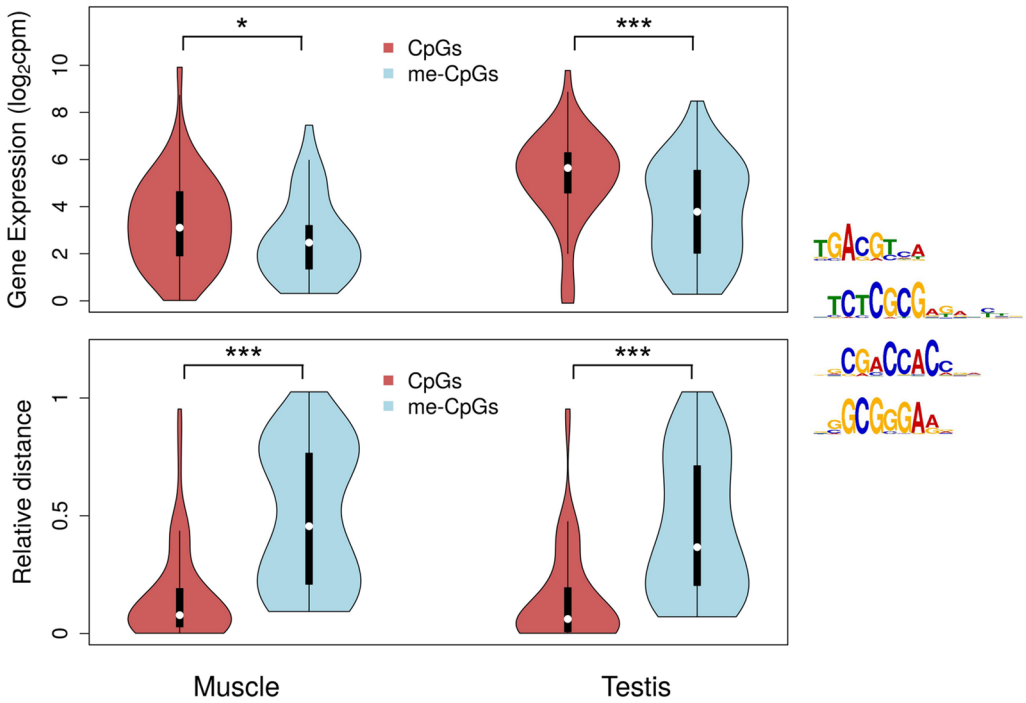

Fig. 4 CpGs in enriched transcription factor (TF) binding sites of the first intron. CpGs were classified as unmethylated (below the first quartile of the total distribution; CpGs, dark red) or methylated (above the third quartile of the total distribution; me-CpGs, light blue) for muscle and for testis. The expression of genes measured as $\log_2$-transformed count per million (cpm) values is shown in the upper panel depending on the type of CpGs these genes contained in their first intron. In the lower panel, the relative distance of the CpGs and me-CpGs from the first exon-first intron boundary was calculated as distance from nucleotide 0 (bp)/width of the intron (bp). The sequences of the four enriched TF-binding motifs that contained CpGs are also shown. The Wilcoxon rank sum test with continuity correction was used to test for statistical differences of gene expression and relative distance between CpGs and me-CpGs, which are reported with the following equivalence: ***$p < 0.001$; *$p < 0.05$

Table 1 Effects on gene expression of the methylation status and the relative distance of the CpGs inside the four transcription factor binding motifs enriched in the first introns

Tissue	Factors	SS	*d.f.*	*F* value	*p* value
Muscle	Relative distance	5.18	1	1.872	0.173
	Methylation status	10.76	1	3.885	0.051
	Interaction of relative distance with methylation status	17.35	1	6.264	*0.013*
	Residuals	387.65	140		
Testis	Relative distance	7.31	1	1.855	0.176
	Methylation status	121.86	1	30.900	*0.000*
	Interaction of relative distance with methylation status	22.8	1	5.781	*0.018*
	Residuals	540.28	137		

The effects were tested using analysis of covariance, and the statistically significant ones are shown in italics

d.f. degrees of freedom, *SS* sums of squares

(89.5% of genes in muscle and 84.9% in testis). However, in the lowest expressed genes, DNA methylation was more equally distributed to the two extremes, with the 75.2% of genes in the muscle and the 48.2% of genes in the testis having less than 10% methylation, and the 24.8% of genes in the muscle and the 51.8% of genes in the muscle having more than 90% methylation.

More genes contain tDMRs in their first intron than in other gene features

In general, there were distinct patterns between the tissue-specific genes and the non-tissue-specific genes that were obvious from the gene expression data. Therefore, we focused on tDMRs between testis and muscle and explored the relationships between differential DNA methylation and differences in gene expression. Both directions of correlation were evident between DNA methylation and gene expression. There was strong negative correlation ($\rho = -0.73$, $p < 0.001$) for up-regulated genes, in either testis or muscle that contained hypo-methylated tDMRs inside their gene body or 4 kb upstream of the TSS or downstream of the 3′ UTR

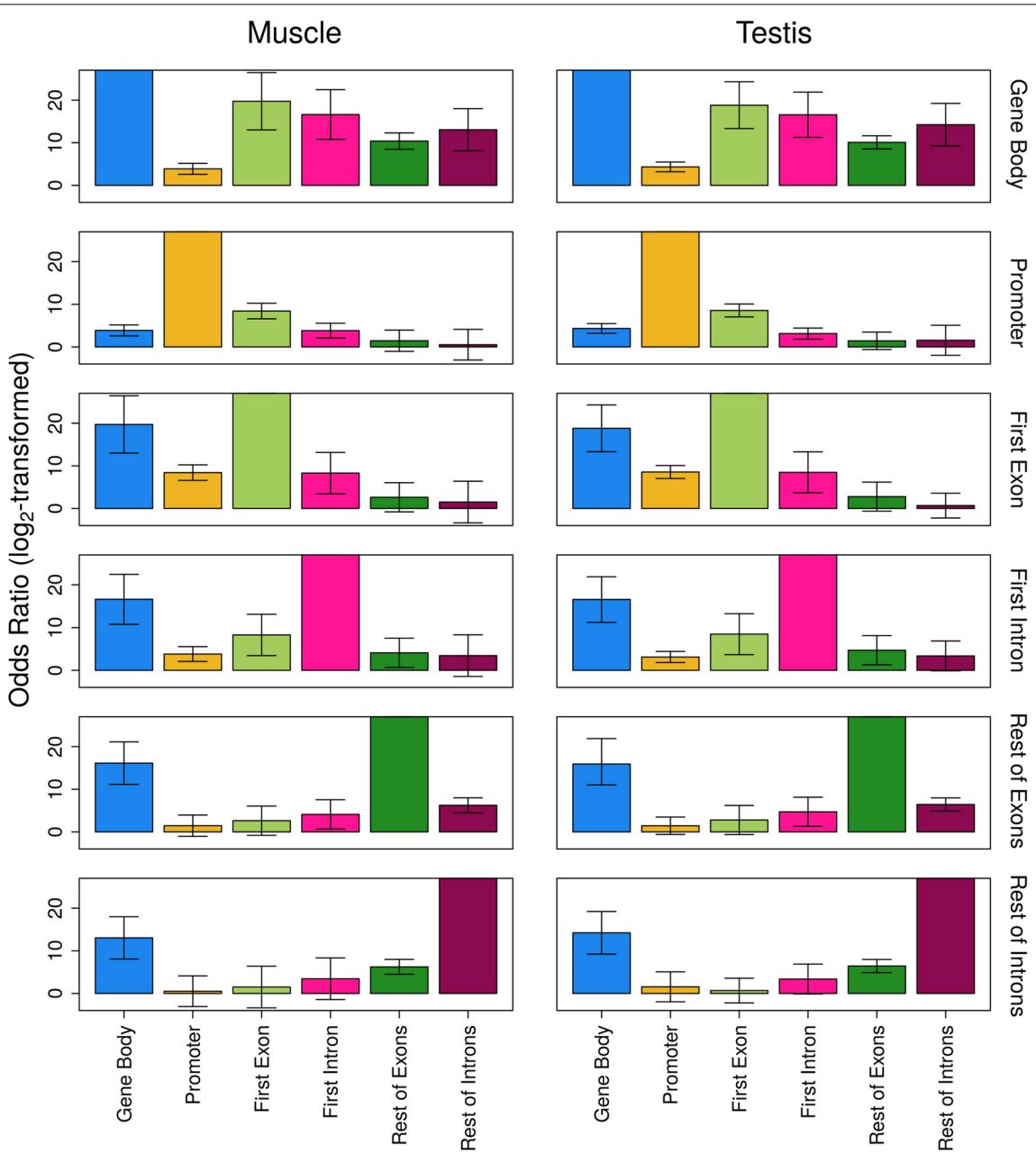

Fig. 5 Association of DNA methylation between pairs of gene features as measured by odds ratio. A gene feature was considered methylated if DNA methylation > 90% and unmethylated if DNA methylation < 10%. The odds ratio (OR) indicates the pairwise association between the methylation states of the gene features of interest, including gene body (all exons and introns), promoter, first exon, first intron, rest of exons and rest of introns in muscle and testis. The odds ratio is represented as $\log_2$-transformed values, and the bars indicate the 99.9% confidence intervals based on the Wald approximation. For the associations of gene feature A versus gene feature A, values were set to maximum and confidence intervals are not applicable

(squares in Fig. 6). Nonetheless, there was also weaker positive correlation ($\rho = 0.26$, $p < 0.001$) for up-regulated genes, in either testis or muscle that contained hypermethylated tDMRs (circles in Fig. 6). The same two categories of genes showing either negative or positive correlation between DNA methylation and gene expression were obvious in genes that contained tDMRs only in the promoter, first exon or first intron (Fig. 6). There were 187 genes that contained tDMRs in their first intron, while there were fewer genes that contained tDMRs in

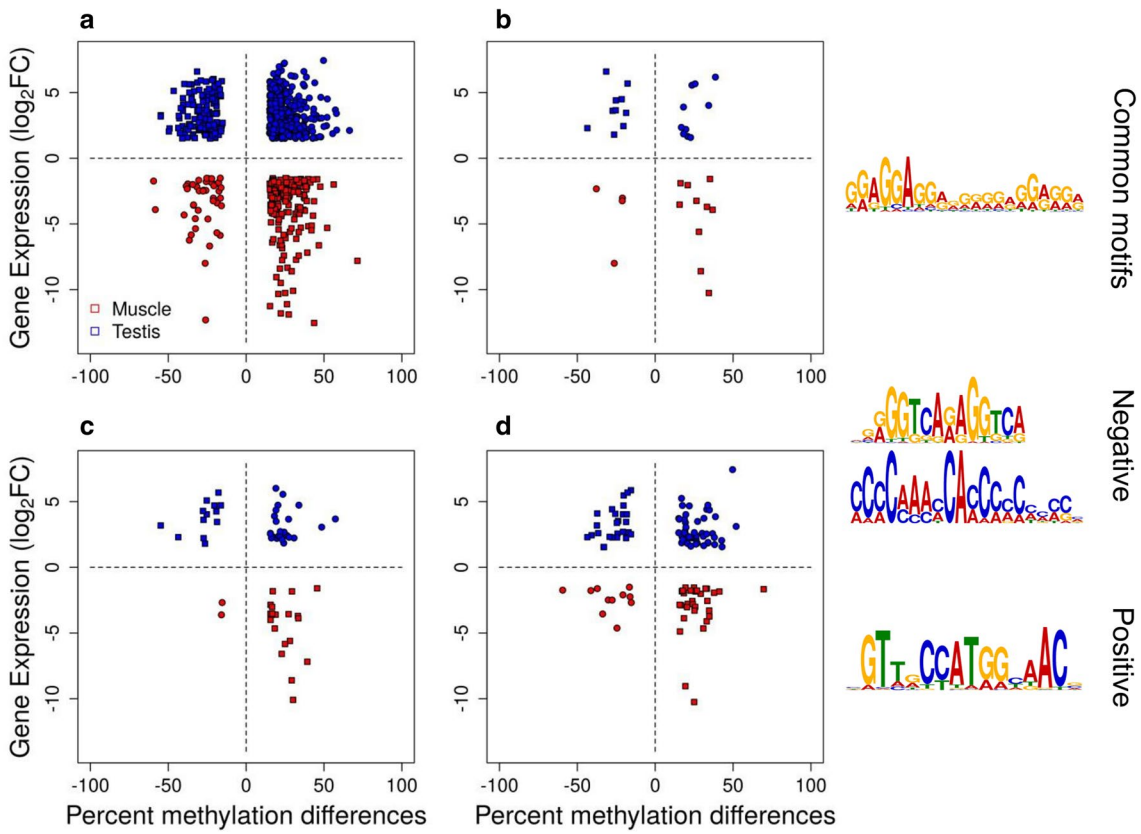

Fig. 6 Differentially expressed genes with differentially methylated regions (tDMRs) between tissues. tDMRs overlap with the gene body and/or 4 kb upstream from the transcription start site or downstream of the 3′ UTR (**a**, $n = 1044$), the promoter (**b**, $n = 47$), the first exon (**c**, $n = 75$) or the first intron (**d**, $n = 187$). Positive (circles) and negative (boxes) correlation is shown for up-regulated genes in muscle (red) and up-regulated genes in testis (blue). Hyper-methylated tDMRs and up-regulated DEGs in testis (blue circles), hypo-methylated tDMRs and up-regulated DEGs in testis (blue squares), hypo-methylated tDMRs and up-regulated DEGs in muscle (red squares) and hyper-methylated tDMRs and up-regulated DEGs in muscle (red circles). Differentially expressed were considered the genes with $\log_2$ fold change > |1.5| and false discovery rate < 0.05. tDMRs were defined as regions showing more than 15% methylation difference between tissues and q value < 0.001, with a minimum number of 5 CpGs and 3 differentially methylated cytosines (DMCs), where a DMC showed more than 15% methylation difference between tissues. Transcription factor binding motifs present in the tDMRs of gene bodies and/or ±4 kb that are common between positive and negative correlation of DNA methylation with gene expression or that are correlation specific (negative or positive)

their promoter (47) and the first exon (75). The majority of the tDMRs in the first intron were located in proximity to the first exon-first intron boundary (Additional file 1: Fig. S7).

Next, we scanned the tDMRs present inside genes or 4 kb upstream of the TSS or downstream of the 3' UTR for TF-binding sites in an attempt to identify features that characterize the type of correlation. The binding motif of the TF ZNF263 was common between the tDMRs associated with genes showing either positive or negative correlation. However, there were also correlation-specific motifs in each case, with the positive correlation-specific motif for binding of the regulatory factor X3 (RFX3) and two negative correlation-specific motifs for binding of the nuclear receptor subfamily 2 group C member 2 (NR2C2) and the Ras-responsive element binding protein 1 (RREB1; Fig. 6).

Discussion

In this study, we present an inverse correlation of the DNA methylation in the first intron and gene expression, which is conserved in tissues with different levels of cellular complexity and across vertebrates. Furthermore, we detect CpGs in enriched TF-binding sites close to the first exon–first intron boundary indicative of gene

expression and tDMRs the methylation of which correlates with gene expression.

Recent studies on the relationship between DNA methylation and gene expression have revealed a key role for the region from +0.5 to +2.5 kb downstream of the TSS in transcriptional regulation of different human tissues [37, 38]. In addition, the methylation of the first exon was shown in mammalian cell lines to be negatively correlated to gene expression, in a more pronounced way than the promoter region [9]. In parallel, the methylation of the first intron has been shown by functional studies to have both positive and negative correlation with gene expression in specific genes in cancer cell lines, fetal and adult tissues [39], CD4+ cell lines isolated from mice [40], during lineage specification in T cells [41], in multiple myeloma cell lines [42], in leukocytes of patients with schizophrenia [43] and in blood samples isolated from children [44]. Furthermore, it has been suggested that the first intron contains distinct properties from the rest of introns and that is linked to transcriptional regulation [45, 46]. These properties may be linked to the closer proximity of the first intron to the TSS, since in many species the CpG-rich regions expand from the CGI promoter to the surrounding sequences [47]. Other active chromatin marks are enriched in conserved parts of the human first intron [48]. However, to the best of our knowledge, an association at the genome-wide scale of the methylation of the first intron as an outlined gene feature has not been demonstrated to date.

In two complex tissues of a phylogenetically distant species, we present the same pattern previously observed in humans. In the European sea bass genome, the median methylation of the first intron showed, consistently in both tissues, the most clear inverse relationship with gene expression among all gene features. We further show that the inverse relationship between DNA methylation of the first intron and gene expression is conserved across vertebrate species, since it is evident in the model fish *Tetraodon*, in the model frog *Xenopus* and two healthy human tissues. This does not seem to hold in invertebrates, which actually exhibit distinct mechanisms of DNA methylation and extreme diversity [49, 50]. In general, across vertebrates, several characteristics of the sequences related to DNA methylation, like the CpG density, are conserved around the TSSs. These characteristics slightly differ for fish, where there is higher CpG density, but still signatures of CGI promoters [47]. Nevertheless, even though the CpG density differs, conserved mechanisms of regulation are present as reflected in this study. Overall, there seems to be a conserved role for the DNA methylation of the first intron across tissues, vertebrates and developmental stages.

The classic promoter methylation–gene expression model seems to hold only in extreme cases and specific genes, while a "triple-inverse" model was suggested, where the methylation of the promoter and the gene body exerts separate influences on gene expression [51]. The variability of DNA methylation in the first intron from 0 to 100% independently of the expression level could be linked to the more variable and larger length of the intron and suggestive of a more complex role of this methylation. In our case, a general pattern would be a negative association of the methylation of the first intron with gene expression in the majority of genes, which is clearest at the extremes of the gene expression range. Nevertheless, subclasses of genes escape this rule and exhibit positive correlations. Indeed, at the genome-wide level there are clusters of genes, each one showing different DNA methylation patterns associated with gene expression [38, 52]. DNA methylation in positive correlation with gene expression has been suggested to appear either as cause or consequence of transcription [53]. Our results also suggest a permissive state of gene expression linked with low methylation, but not a linear inhibitory link with high methylation. These observations are in support of the increasing evidence of a far more complex relationship of the epigenetic modifications—including histone and DNA modifications and noncoding RNAs—with gene expression at a spatiotemporal scale [52].

The function of the inverse relationship between DNA methylation of the first intron and gene expression could be partially explained by the presence of intronic enhancers interacting with the promoters of their corresponding genes. Indeed, the intron-mediated enhancement is a well-described phenomenon [48, 54, 55]. Silencing of intragenic enhancers is considered to play a role even more significant than promoter methylation in the silencing of their target genes [7]. In addition, intronic enhancers present tissue-specific methylation status associated with gene expression [56, 57]. Our results also revealed enriched TF-binding motifs common between tissues. Moreover, the unmethylated CpGs tend to be located closer to the beginning of the first intron and associated with higher gene expression levels, while the methylated CpGs tend to be further downstream and associated with lower gene expression levels. Therefore, the methylation status of the CpGs at shorter or greater distances from the beginning of the first intron which belong to a TF-binding motif is indicative of the gene expression level. Taken together, these results further support a regulatory role for the DNA methylation in the first intron region, although further experiments are needed to demonstrate the mechanistic relationships at the functional level.

tDMRs located in the whole gene showed both positive and negative correlation with gene expression as

in human tissues [11]. Here, in addition to confirm this in a phylogenetically distant species, we partition the genomic localization of tDMRs in three important gene features: promoter, first exon and first intron. tDMRs are distributed across all these gene features and exhibit both directions of correlation, but no enrichment of positive or negative correlation depending on the location. This is in accordance with the latest findings in human tissues [11] and in contrast to the standard model of gene regulation by DNA methylation. However, in the first intron, there are more tDMRs, in agreement with our finding of the importance of the first intron in the regulation of gene expression by DNA methylation. In human macrophages after bacterial infection, the majority of gene body DMRs were also located in the first introns of genes [58]. Taken together, these results suggest an overlooked role for the DNA methylation of the first intron in the tissue-specific regulation of gene expression.

We used RRBS to assess DNA methylation and RNA-seq to measure the gene expression levels in the European sea bass genome which is one of the best annotated teleost genomes [31]. Nevertheless, the precision of the annotation of genes and their regulatory elements is not comparable to model species, like human or mouse. Therefore, we defined promoters as the region −1000 kb upstream the predicted TSSs, as commonly arbitrarily defined [59–64], but without excluding the possibility of alternative TSSs or variable promoter lengths. The limitations of the sea bass genome annotation may influence also the gene expression data, where the analysis could only be performed at the gene level, based on the current annotation. RRBS allows for enrichment of the standard relevant parts of the genome for DNA methylation, e.g., promoters and CpG islands, and requires only a modest amount of sequencing [65, 66], making it a cost-effective alternative to WGBS which is considered generally inefficient since only 20–30% of the reads provide relevant information [67, 68]. Our RRBS results, including the genome representation and the actual methylation values, are comparable to other teleost fish, like the stickleback [69], the Atlantic salmon [70] and the zebrafish [71]. Regardless of the limitations of this study related to the main species in question and the techniques used, our key results were confirmed in other vertebrate species, corroborating the general trends shown in the sea bass, even accepting the possibility that some genes may be not well annotated.

Conclusions

Our integrative analyses clearly reveal the important and conserved role of the methylation level of the first intron and its inverse association with gene expression regardless of tissue and species. Notably, the first intron exhibits a tissue-independent enrichment for TF-binding motifs and the methylation of the CpGs they contain is indicative of the gene expression level. Furthermore, the first intron presents a higher number of tDMRs than other gene features, suggestive of a regulatory role in tissue-specific expression. These findings not only contribute to our basic understanding of the epigenetic regulation of gene expression but also identify the first intron as an informative gene feature regarding the relationship between DNA methylation and gene expression where future studies could be focused, e.g., for the design of target sequences or for the analysis of genome-wide data throughout the region downstream the TSSs.

Methods

Animals

Wild European sea bass (*Dicentrarchus labrax*) adults with body weight = 1000 ± 109.5 g (mean ± SEM), standard length = 39.3 ± 1.4 cm and gonadosomatic index = 0.076 ± 0.009, the latter calculated as in [72], were captured by speargun at the Montgrí, Medes Islands and Baix Ter Natural Reserve (NE Spain) during the non-reproductive season (June 2013). Since the fish were caught in the wild, even if they were size-matched, they may have shown variation due to age, status or environment. Therefore, in further DNA methylation and gene expression analyses, biological variation was taken into account. Tissues were dissected immediately upon capture and stored in RNAlater® (ThermoFisher Scientific).

RNA isolation

Total mRNA was isolated from testis and muscle of five fish. Tissues were removed from RNAlater®, dried, immersed into TRIzol® Reagent (ThermoFisher Scientific) and homogenized by the Polytron PT 1200 CL (Kinematica AG). RNA extraction was performed according to the manufacturer's instructions. RNA was quantified by the Qubit® RNA BR Assay Kit (ThermoFisher Scientific), and RNA quality was evaluated by the Agilent RNA 6000 Nano Kit (Agilent). Samples with RNA integrity number (RIN) > 8 were used for library construction.

RNA-seq

The libraries were prepared using the mRNA-Seq sample preparation kit (Illumina Inc., Cat. # RS-122-2001x2) according to the manufacturer's protocol. Briefly, 0.5 μg of total RNA were used for poly-A-based mRNA enrichment selection using oligo-dT magnetic beads followed by fragmentation by divalent cations at elevated temperature resulting into fragments of 80–250 nt, with the major peak at 130 nt. First-strand cDNA synthesis by random hexamers and reverse transcriptase was followed by the second-strand cDNA synthesis. Double-stranded

cDNA was end-repaired and 3′-adenylated, and the 3′-"T" nucleotide at the Illumina adapter was used for the indexed adapters ligation. The ligation product was amplified using 15 PCR cycles. Each library was sequenced using the TruSeq SBS Kit v3-HS, in 76-bp paired-end mode on an Illumina HiSeq 2000 instrument following the manufacturer's protocol. Images from the instrument were processed using the manufacturer's software to generate FASTQ sequence files.

RNA-seq analysis

RNA-seq reads were aligned with the GEMtools RNA-seq pipeline v1.7 (http://gemtools.github.io), which is based on the GEM mapper [73]. The pipeline aligns the reads in a sample in three phases, mapping against the reference genome (dicLab v1.0c, Jul. 2012), against a reference transcriptome (COMBINED ANNOTATION track) and against a de novo transcriptome, generated from the input data to detect new junction sites. The sea bass genome used here is one of the best in silico annotated fish genomes [31]. After mapping, all alignments were filtered to increase the number of uniquely mapped reads. The filtering criteria included a minimum intron length of 20 bp, a maximum exon overlap of 5 bp and a filter step against a reference annotation checking for consistent pairs and junctions where both sites align to the same annotated gene. The libraries' statistics including the number of raw reads and the average reads aligned can be found in Additional file 2. The same pipeline was used to quantify expression at the gene level. Similarity across RNA-seq samples was investigated with principal component analysis (PCA; Additional file 1: Fig. S8A), where the first principal component separated the samples by tissue type and explained almost 96% of the variance. One muscle sample was excluded from further analysis since it was clearly an outlier in the quality clustering. The variation in gene expression values was higher and with more extreme values in muscle (Levene's test; $p < 0.001$), whereas the testis had higher expression median (Mood's median test; $\chi^2 = 3372.7$, $p < 0.001$; Additional file 1: Fig. S9A). Subsequently, the TMM method [74] was used for gene expression normalization, which takes into account not only library size (sequencing depth) of the samples but also the composition of the RNA population. The EdgeR robust method [75] was used for differential expression analysis. Genes with p adjusted < 0.05 were considered significant. Positive $\log_2$-transformed fold change (FC) indicates up-regulation in testis, and negative $\log_2$FC indicates down-regulation in testis. There were 9449 up-regulated genes and 6220 down-regulated genes in testis compared to muscle (FDR < 0.05). Furthermore, most of the genes expressed in both tissues had higher expression levels in the testis than in the muscle (Additional file 1: Fig. S9B). Tissue-specific genes were considered the genes that were expressed in only one of the two tissues, regardless of the actual expression level. Approximately 20 and 2000 genes accounted for half of the number of reads mapped in the muscle and testis, respectively (Additional file 1: Fig. S9C). However, the testis-specific genes had lower median expression (Mood's median test; $\chi^2 = 32.282$, $p < 0.001$) than the muscle-specific genes (Additional file 1: Fig. S9D). These results confirmed that the testis and muscle constitute two tissues with very different transcriptomic complexity and validated our choice of tissues for the purposes of this study.

DNA isolation

Genomic DNA was extracted by phenol/chloroform/isoamyl alcohol (PCI) from 3 of the samples of testis and muscle used to prepare RNA-seq libraries from a fragment contiguous to the one used for RNA extraction. DNA was extracted also from liver and spleen of one of the same fish. In brief, tissue samples were dried out from RNAlater® and immersed into digestion buffer (0.1 M NaCl, 10 mM Tris–HCl, 1 mM EDTA pH 8, 0.5% SDS), and proteins were digested by 1 µg of proteinase K (Sigma-Aldrich) and RNA by 0.5 µg of ribonuclease A (PureLink RNase A; Life Technologies). DNA was precipitated by 95% ethanol, eluted in Milli-Q® water (Merck, Millipore) and cleaned up with 2× AMPure XP beads (Beckman Coulter) to ensure purity. DNA was quantified three times by independent means, being by ND-spectrophotometer (NanoDrop Technologies) or Qubit™ fluorometric quantitation (ThermoFisher Scientific), each time followed by dilutions with nuclease-free water in order to normalize DNA quantities across samples.

RRBS libraries preparation

RRBS libraries were prepared as in Klughammer et al. [76]. One hundred nanograms of genomic DNA was digested by 20 units of *Msp*I (NEB) overnight at 37 °C. Five units of Klenow fragment (3′ → 5′exo-; NEB) and dNTP mix (final concentration: 300 µM dATP, 30 µM dCTP and 30 µM dGTP) were added to the reaction. End fill-in was performed for 20 min at 30 °C, A-tailing for 20 min at 37 °C and inactivation of the enzyme for 20 min at 75 °C. Ligation of Illumina TruSeq Adapters v2 was performed by Quick Ligase (NEB) for 20 min at 25 °C, followed by heat inactivation of the enzyme for 10 min at 65 °C. Libraries were size-selected by 0.75× 1:5 diluted AMPure XP beads, quantified by qPCR, pooled based on qPCR values and cleaned up with 2.5× 1:5 diluted AMPure XP beads. Samples were subjected to bisulfite conversion using the EZ DNA Methylation-Direct kit

(Zymo Research) with 0.9x CT Conversion Reagent, 20 cycles of 95 °C for 1 min and 60 °C for 10 min and desulphonation time extended to 30 min. Libraries were enriched by the PfuTurbo Cx HotStart Polymerase (Agilent Technologies) with the following cycling parameters: 95 °C for 2 min, followed by the optimal number of cycles of 95 °C for 30 s, 65 °C for 30 s and 72 °C for 45 s, and a final step at 72 °C for 7 min. The optimal number of cycles for the enrichment PCR was calculated based on qPCR values. A final clean-up step was performed by 1x AMPure XP beads. The quantity of the libraries was measured by Qubit High Sensitivity assays (ThermoFisher Scientific), and the quality was evaluated by Experion DNA 1 k assays (BioRad). Sequencing of RRBS libraries was performed on an Illumina HiSeq 2000 platform in 50-bp single-end mode.

DNA methylation analysis

RRBS raw reads were quality trimmed by the Trimmomatic v. 0.32 [77] using a sliding window trimming with window size 4 and required quality 15, an adaptive quality trimming with the target length set at 20 and the strictness at 0.50 and a minimum read length of 18 bp. Trimmed reads were aligned to the reference genome of sea bass using BSMAP v. 2.90 [78] in RRBS mode requiring a minimum coverage of 5 reads. Methylation calling was performed by the *methratio.py* python script that accompanies BSMAP. The bisulfite conversion rate was calculated using the *bsrate* script of the MethPipe pipeline v. 3.4.3 [79]. In brief, the RRBS libraries showed a mean of 40,342,296 reads, a mean alignment rate of 81.26%, 1,122,487 covered CpGs, a mean fold coverage of 60.16 and 99.1% of bisulfite conversion ratio. The statistics of the libraries per sample including the number of raw reads, the alignment rates, the number of covered cytosines and the bisulfite conversion ratio can be found in Additional file 2. All subsequent bioinformatics analyses were performed using R v. 3.4.1 and Rstudio v. 1.0.143 [80, 81] and Bioconductor packages [82], unless stated otherwise. The package *methylKit* v. 1.2.0 [83] was used for DNA methylation analysis. Called bases with less than 10 reads or more than the 99.9th percentile of coverage distribution were filtered out. Coverage values were normalized as by default and bases were united in order to retain the ones that were covered in all samples which were 529,070. Pairwise comparisons of RRBS DNA methylation values for testis and muscle showed good correlation between biological replicates within each tissue and higher for testis (Pearson's correlation scores: testis ≥ 0.97; muscle ≥ 0.83, Additional file 1: Fig. S8B). Overall, DNA methylation levels were similar between the two tissues and showed a strong positive correlation (Pearson's product-moment correlation $= 0.95$, p value < 0.001; white to blue scale in Additional file 1: Fig. S10A). Differentially methylated cytosines (DMCs) between tissues were defined as CpGs with more than 15% methylation differences and q value < 0.01 after applying logistic regression using the SLIM method for p value adjustment. Positive values indicate hyper-methylation in testis, and negative values indicate hypo-methylation in testis. Among the 500 top-differentially methylated CpG (DMC) sites, there were ~2.8 times more hyper-methylated CpGs in the testis than in the muscle (Additional file 1: Fig. S10B), while there were no CpG sites with >90% methylation in muscle and <10% methylation in the testis (green to red scale in Additional file 1: Fig. S10A). Differentially methylated regions (DMRs) between tissues were identified using the weighted optimization algorithm for empirically based DMRs of the package *edmr* v. 0.6.4.1 [84] with default parameters, except for DMC differences cutoffs which were set to 15% and DMR differences cutoffs set to 15%. We used a 15% cutoff for defining differential methylation after exploratory analyses with variable thresholds, since it represented a good compromise between robust differential methylation and retention of potentially interesting loci.

Combined analysis of DNA methylation and gene expression

A BSgenome package [85] was created for use when required using the full sea bass genome and masks from the UCSC server (dicLab v1.0c, Jul. 2012). Annotations of gene features were based on the COMBINED ANNOTATION track. Promoters were defined as 1000 bp upstream the in silico annotated transcription start sites (TSSs) from the COMBINED ANNOTATION track. The DNA methylation levels of gene features were calculated by averaging the methylation values per gene feature. Genomic overlaps of features were identified using the GenomicRanges (v. 1.28.4) package [86].

The vioplot (v.0.2) package was used for visualizing methylation data by expression decile [87]. For positive and negative correlations of methylation differences and gene expression, first we identified the DMRs located in genomic regions encompassing the whole gene bodies and 4 kb upstream from the TSSs or downstream of the 3' UTR. Then, DMRs overlapping with promoters, first exons or first introns were identified. Only genes with $\log_2$ FC $> |1.5|$ and FDR < 0.05 were considered as differentially expressed. Enrichment of transcription factor (TF) binding motifs was performed using the ame tool v. 4.9.1 [88] of the MEME suite [89] using as input the JASPAR CORE 2016 vertebrates database [90] and shuffled input sequences as controls for enrichment. AME shuffles the input sequences while preserves the dinucleotide

frequencies. TF-binding motifs were considered enriched if p value < 0.001 after Bonferroni correction of multiple Fisher's exact test. The sequencing scoring method was the average odds score and the input sequences contained ±50 bp from the CpGs of the first introns. This distance was chosen to encompass the maximum TF-binding motifs length (31 nucleotides [36]) and an arbitrary +20 nucleotides more. Four enriched TF-binding motifs common between muscle and testis that contained CpG sites were selected for screening in the sequences of the first introns using the fimo tool v. 4.9.1. CpGs present in these sequences were classified as unmethylated if their methylation was below the first quartile of the distribution and as methylated if their methylation was above the third quartile of the distribution. The relative distance of these CpGs from the start of the first intron was calculated as the distance from nucleotide 0 (bp)/width of the intron (bp). Detection of TF-binding motifs inside the sequences of tDMRs in gene bodies 4 kb upstream from the TSS or downstream of the 3' UTR was performed using the fimo tool v. 4.9.1 to scan for the motifs of the JASPAR CORE 2016 vertebrates database. Only TF-binding motifs with a q value < 0.01 for both muscle and testis were considered.

Other species data

WGBS-seq and RNA-seq data of whole *Tetraodon nigroviridis* and of muscle tissue of *Ciona intestinalis* were obtained from the study [20] (NCBI Gene Expression Omnibus [33, 91] with accession number GSE19824). WGBS-seq from gastrula stage 10.5 and RNA-seq data from gastrula stage 11 of *Xenopus tropicalis* were obtained from the study [34] (GEO with accession number GSE67974). WGBS-seq and RNA-seq data of normal human lung and liver were obtained from the study [35] (GEO with accession number GSE70091). For the human data, 3 replicates were available; therefore, after excluding positions with less than 10 reads coverage from the WGBS data, the methylation per position was calculated as 100*methylated_read_count/total_read_count and averaged for the three replicates. Gene annotations were read with the readTranscriptFeatures function of the genomation v.1.8.0 package [92], and the first intron was selected. The methylation per gene feature was calculated as the average of CpGs covered in each gene feature. Expressed genes (cpm > 0) with methylation in the first intron were ordered, split in deciles according to their expression levels and plotted using vioplot.

Statistical analysis of the data

Statistical analyses of the data were performed by R v. 3.4.1 and Rstudio v. 1.0.143 [80, 81]. The association of DNA methylation for pairs of gene features was calculated as the odds ratios: $(N_{00} \times N_{11})/(N_{01} \times N_{10})$, where N_{00} = gene feature (GF) 1 < 10% and GF2 < 10%, N_{11} = GF1 > 90% and GF2 > 90%, N_{01} = GF1 < 10% and GF2 > 90% and N_{10} = GF1 > 90% and GF2 < 10%. The quantification of the strength of the association by odds ratio was chosen as done before for the same purpose [9]. The Wald approximation was used to calculate the confidence intervals at alpha = 0.001.

Correlations between DNA methylation data were measured using Pearson's product-moment correlation coefficient. Correlations between DNA methylation and gene expression were measured using Spearman's rank correlation coefficient because the relationship of DNA methylation with gene expression data is not necessarily expected to be linear. To compare the medians of gene expression between the two tissues, the Mood's median test was used. Homogeneity of variances was checked by Levene's test. To compare that variance of DNA methylation in gene features between tissues and expression deciles, an ANOVA on the residuals followed by Tukey's honest significant differences was performed. For pairwise comparisons of DNA methylation values between the extreme expression groups, the Wilcoxon rank sum test with continuity correction was used, after removing outliers as defined by Tukey fences (values below $Q_1 - 1.5(Q_3 - Q_1)$ or above $Q_3 - 1.5(Q_3 - Q_1)$). The effects of relative distance and methylation status on gene expression using analysis of covariance after checking for normality of the residuals.

Additional files

Additional file 1. Supplementary Figures and Tables.

Additional file 2. RRBS and RNA-seq libraries basic statistics. This file includes the raw number of reads, the alignment rates and other basic statistics of the NGS libraries prepared for this study.

Authors' contributions

DA conceived the study, performed the experimental part, analyzed the RRBS methylation and expression data, interpreted results and drafted the article; AEC analyzed the RNA-Seq data, interpreted results and revised the article; and FP conceived the study, interpreted results and drafted the article. All authors read and approved the final manuscript.

Author details

[1] Institute of Marine Sciences (ICM-CSIC), Passeig Marítim de la Barceloneta, 37-49, 08003 Barcelona, Spain. [2] CNAG-CRG, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain. [3] Universitat Pompeu Fabra (UPF), Barcelona, Spain.

Acknowledgements

We would like to thank Oscar Sagué and Álex Lorente for assistance with the capture of the wild fish; Dr. Noelia Díaz and Sílvia Joly (ICM, Barcelona) for help with the samplings; Dr. Christoph Bock, Dr. Matthias Farlik, Paul Datlinger, Johanna Klughammer and Charles Dietz (CeMM, Austrian Academy of Sciences) for help with the RRBS libraries preparation and basic bioinformatics; and Dr. Esteban Ballestar (IDIBELL, Barcelona) for helpful comments.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The RRBS and RNA-seq data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus [33, 91] and are accessible through GEO Series accession number GSE104366.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The fish used in this study were wild fish captured by speargun at the Medes Islands and Baix Ter Natural Reserve (NE Catalonia).

Funding
DA was supported by a Ph.D. scholarship from the Spanish Government (BES–2011–044860). AEC is funded by the RED-BIO project of the Spanish National Bioinformatics Institute (INB) under grant number PT13/0001/0044. The INB is funded by the Spanish National Health Institute Carlos III (ISCIII) and the Spanish Ministry of Economy and Competitiveness (MINECO). Research supported by MINECO grant "Epifarm" (ref. AGL2013–41047–R) to FP. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

Publisher's Note

Received: 18 April 2018 Accepted: 19 June 2018
Published online: 29 June 2018

References

1. Lowdon RF, Jang HS, Wang T. Evolution of epigenetic regulation in vertebrate genomes. Trends Genet TIG. 2016;32:269–83.
2. Gilbert SF, Epel D. Ecological developmental biology: integrating epigenetics, medicine, and evolution. Sunderland: Sinauer Associates; 2008.
3. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38:23–38.
4. Illingworth RS, Bird AP. CpG islands—'a rough guide'. FEBS Lett. 2009;583:1713–20.
5. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, et al. Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol. 2009;16:564–71.
6. Bogdanović O, Smits AH, de la Calle Mustienes E, Tena JJ, Ford E, Williams R, et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet. 2016;48:417–26.
7. Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, et al. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. Genome Biol. 2014;15:1.
8. Tomazou EM, Sheffield NC, Schmidl C, Schuster M, Schönegger A, Datlinger P, et al. Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. Cell Rep. 2015;10:1082–95.
9. Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS ONE. 2011;6:e14524 **(Papavasiliou N, editor)**.
10. Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009;27:361–8.
11. Wan J, Oliver VF, Wang G, Zhu H, Zack DJ, Merbs SL, et al. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. BMC Genom. 2015;16:49.
12. Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, Kolde R, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol. 2014;15:r54.
13. Hernando-Herraez I, Heyn H, Fernandez-Callejo M, Vidal E, Fernandez-Bellon H, Prado-Martinez J, et al. The interplay between DNA methylation and sequence divergence in recent human evolution. Nucleic Acids Res. 2015;43:8204–14.
14. Zhong X. Comparative epigenomics: a powerful tool to understand the evolution of DNA methylation. New Phytol. 2016;210:76–80.
15. Varriale A. DNA methylation, epigenetics, and evolution in vertebrates: facts and challenges. Int J Evol Biol. 2014;2014:e475981.
16. Baerwald MR, Meek MH, Stephens MR, Nagarajan RP, Goodbla AM, Tomalty KMH, et al. Migration-related phenotypic divergence is associated with epigenetic modifications in rainbow trout. Mol Ecol. 2016;25:1785–800.
17. Jiang L, Zhang J, Wang J-J, Wang L, Zhang L, Li G, et al. Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell. 2013;153:773–84.
18. Potok ME, Nix DA, Parnell TJ, Cairns BR. Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell. 2013;153:759–72.
19. Shao C, Li Q, Chen S, Zhang P, Lian J, Hu Q, et al. Epigenetic modification and inheritance in sexual reversal of fish. Genome Res. 2014;24:604–15.
20. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328:916–9.
21. Chen X, Wang Z, Tang S, Zhao Y, Zhao J. Genome-wide mapping of DNA methylation in *Nile tilapia*. Hydrobiologia. 2016;791:247–57. https://doi.org/10.1007/s10750-016-2823-6
22. Sun L-X, Wang Y-Y, Zhao Y, Wang H, Li N, Ji XS. Global DNA methylation changes in *Nile tilapia* gonads during high temperature-induced masculinization. PLoS ONE. 2016;11:e0158483.
23. Nätt D, Rubin C-J, Wright D, Johnsson M, Beltéky J, Andersson L, et al. Heritable genome-wide variation of gene expression and promoter methylation between wild and domesticated chickens. BMC Genom. 2012;13:59.
24. Derks MFL, Schachtschneider KM, Madsen O, Schijlen E, Verhoeven KJF, van Oers K. Gene and transposable element methylation in great tit (*Parus major*) brain and blood. BMC Genom. 2016;17:332.
25. Cao J, Wei C, Liu D, Wang H, Wu M, Xie Z, et al. DNA methylation landscape of body size variation in sheep. Sci Rep. 2015;5:13950.
26. Couldrey C, Brauning R, Bracegirdle J, Maclean P, Henderson HV, McEwan JC. Genome-wide DNA methylation patterns and transcription analysis in sheep muscle. PLoS ONE. 2014;9:e101853 **(Niemann H, editor)**.
27. Choi M, Lee J, Le MT, Nguyen DT, Park S, Soundrarajan N, et al. Genome-wide analysis of DNA methylation in pigs using reduced representation bisulfite sequencing. DNA Res. 2015;22:343–55.
28. Janowitz Koch I, Clark MM, Thompson MJ, Deere-Machemer KA, Wang J, Duarte L, et al. The concerted impact of domestication and transposon insertions on methylation patterns between dogs and grey wolves. Mol Ecol. 2016;25:1838–55.
29. Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C, et al. Dynamics of DNA methylation in recent human and great ape evolution. PLoS Genet. 2013;9:e1003763.
30. Lea AJ, Altmann J, Alberts SC, Tung J. Resource base influences genome-wide DNA methylation levels in wild baboons (*Papio cynocephalus*). Mol Ecol. 2016;25:1681–96.
31. Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun. 2014;5:5770.
32. Louro B, Power DM, Canario AVM. Advances in European sea bass genomics and future perspectives. Mar Genomics. 2014;18:71–5.
33. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
34. Hontelez S, van Kruijsbergen I, Georgiou G, van Heeringen SJ, Bogdanovic O, Lister R, et al. Embryonic transcription is controlled by maternally defined chromatin state. Nat Commun. 2015;6:10148.
35. Li X, Liu Y, Salz T, Hansen KD, Feinberg A. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. Genome Res. 2016;26:1730–41.
36. Stewart AJ, Hannenhalli S, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. Genetics. 2012;192:973–85.

37. Schlosberg CE, VanderKraats ND, Edwards JR. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. Nucleic Acids Res. 2017;45:5100–11.
38. VanderKraats ND, Hiken JF, Decker KF, Edwards JR. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. Nucleic Acids Res. 2013;41:6816–27.
39. Unoki M, Nakamura Y. Methylation at CpG islands in intron 1 of EGR2 confers enhancer-like activity. FEBS Lett. 2003;554:67–72.
40. Hashimoto S, Ogoshi K, Sasaki A, Abe J, Qu W, Nakatani Y, et al. Coordinated changes in DNA methylation in antigen-specific memory CD4 T cells. J Immunol. 2013;190:4076–91.
41. Sellars M, Huh JR, Day K, Issuree PD, Galan C, Gobeil S, et al. Regulation of DNA methylation dictates *Cd4* expression during the development of helper and cytotoxic T cell lineages. Nat Immunol. 2015;16:746.
42. Hayami Y, Iida S, Nakazawa N, Hanamura I, Kato M, Komatsu H, et al. Inactivation of the E3/LAPTm5 gene by chromosomal rearrangement and DNA methylation in human multiple myeloma. Leukemia. 2003;17:1650–7.
43. Yoshino Y, Ozaki Y, Yamazaki K, Sao T, Mori Y, Ochi S, et al. DNA Methylation changes in intron 1 of triggering receptor expressed on myeloid cell 2 in Japanese Schizophrenia subjects. Front Neurosci. 2017 [cited 2017 Dec 1];11. https://www.frontiersin.org/articles/10.3389/fnins.2017.00275/full.
44. Kim J, Bhattacharjee R, Khalyfa A, Kheirandish-Gozal L, Capdevila OS, Wang Y, et al. DNA methylation in inflammatory genes among children with obstructive sleep apnea. Am J Respir Crit Care Med. 2012;185:330–8.
45. Li H, Chen D, Zhang J. Analysis of intron sequence features associated with transcriptional regulation in human genes. PLoS ONE. 2012;7:e46784.
46. Majewski J, Ott J. Distribution and characterization of regulatory elements in the human genome. Genome Res. 2002;12:1827–36.
47. Hartono SR, Korf IF, Chédin F. GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. Nucleic Acids Res. 2015;43:9729–41.
48. Park SG, Hannenhalli S, Choi SS. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics. 2014 [cited 2018 Feb 1];15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4085337/.
49. Song X, Huang F, Liu J, Li C, Gao S, Wu W, et al. Genome-wide DNA methylomes from discrete developmental stages reveal the predominance of non-CpG methylation in *Tribolium castaneum*. DNA Res Int J Rapid Publ Rep Genes Genomes. 2017;24:445–57.
50. Takayama S, Dhahbi J, Roberts A, Mao G, Heo S-J, Pachter L, et al. Genome methylation in *D. melanogaster* is found at specific short motifs and is independent of DNMT2 activity. Genome Res. 2014;24:821–30.
51. Lou S, Lee H-M, Qin H, Li J-W, Gao Z, Liu X, et al. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. Genome Biol. 2014;15:408.
52. Yu P, Xiao S, Xin X, Song C-X, Huang W, McDee D, et al. Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. Genome Res. 2013;23:352–64.
53. Rountree MR, Selker EU. DNA methylation inhibits elongation but not initiation of transcription in *Neurospora crassa*. Genes Dev. 1997;11:2383–95.
54. Jeong Y-M, Mun J-H, Lee I, Woo JC, Hong CB, Kim S-G. Distinct roles of the first introns on the expression of Arabidopsis profilin gene family members. Plant Physiol. 2006;140:196–209.
55. Rose AB. Intron-mediated regulation of gene expression. Curr Top Microbiol Immunol. 2008;326:277–90.
56. Hoivik EA, Bjanesoy TE, Mai O, Okamoto S, Minokoshi Y, Shima Y, et al. DNA methylation of intronic enhancers directs tissue-specific expression of steroidogenic factor 1/adrenal 4 binding protein (SF-1/Ad4BP). Endocrinology. 2011;152:2100–12.
57. Rico D, Martens JH, Downes K, Carrillo-de-Santa-Pau E, Pancaldi V, Breschi A, et al. Comparative analysis of neutrophil and monocyte epigenomes. bioRxiv. 2017;237784.
58. Sharma G, Sowpati DT, Singh P, Khan MZ, Ganji R, Upadhyay S, et al. Genome-wide non-CpG methylation of the host genome during *M. tuberculosis* infection. Sci Rep. 2016;6:25006.
59. Vernimmen D, Bickmore WA. The hierarchy of transcriptional activation: from enhancer to promoter. Trends Genet. 2015;31:696–708.
60. Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type–specific gene expression from regions of open chromatin. Genome Res. 2012;22:1711–22.
61. Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, et al. Sequence features that drive human promoter function and tissue specificity. Genome Res. 2010;20:890–8.
62. Weigelt K, Moehle C, Stempfl T, Weber B, Langmann T. An integrated workflow for analysis of ChIP-chip data. BioTechniques. 2008;45:131–2, 134, 136 passim.
63. Akan P, Deloukas P. DNA sequence and structural properties as predictors of human and mouse promoters. Gene. 2008;410:165–76.
64. Karlsson K, Lönnerberg P, Linnarsson S. Alternative TSSs are co-regulated in single cells in the mouse brain. Mol Syst Biol. 2017;13:930.
65. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010;28:1106–14.
66. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016;9:26.
67. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.
68. Ziller MJ, Stamenova EK, Gu H, Gnirke A, Meissner A. Targeted bisulfite sequencing of the dynamic DNA methylome. Epigenetics Chromatin. 2016;9:55.
69. Metzger DCH, Schulte PM. Persistent and plastic effects of temperature on DNA methylation across the genome of threespine stickleback (Gasterosteus aculeatus). Proc Biol Sci. 2017. https://doi.org/10.1098/rspb.2017.1667.
70. Moghadam HK, Johnsen H, Robinson N, Andersen Ø, Jørgensen EH, Johnsen HK, et al. Impacts of early life stress on the methylome and transcriptome of Atlantic Salmon. Sci Rep. 2017;7:5023.
71. Chatterjee A, Ozaki Y, Stockwell PA, Horsfield JA, Morison IM, Nakagawa S. Mapping the zebrafish brain methylome using reduced representation bisulfite sequencing. Epigenetics. 2013;8:979–89.
72. Navarro-Martín L, Blázquez M, Viñas J, Joly S, Piferrer F. Balancing the effects of rearing at low temperature during early development on sex ratios, growth and maturation in the European sea bass (*Dicentrarchus labrax*). Aquaculture. 2009;296:347–58.
73. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012;9:1185–8.
74. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.
75. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42:e91.
76. Klughammer J, Datlinger P, Printz D, Sheffield NC, Farlik M, Hadler J, et al. Differential DNA methylation analysis without a reference genome. Cell Rep. 2015;13:2621–33.
77. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.
78. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinform. 2009;10:232.
79. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS ONE. 2013;8:e81148 **(El-Maarri O, editor)**.
80. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. https://www.R-project.org.
81. RStudio Team. RStudio: integrated development environment for R. Boston, MA: RStudio, Inc.; 2015. http://www.rstudio.com/.
82. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
83. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13:R87.
84. Li S, Garrett-Bakelman FE, Akalin A, Zumbo P, Levine R, To BL, et al. An optimized algorithm for detecting and annotating regional differential methylation. BMC Bioinform. 2013;14:S10.

85. Pagès H. BSgenome: infrastructure for biostrings-based genome data packages and support for efficient SNP representation. 2016. https://bioconductor.org/packages/release/bioc/html/BSgenome.html.
86. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118 **(Prlic A, editor)**.
87. Adler D. vioplot: violin plot. 2005 [cited 2016 Aug 26]. https://cran.r-project.org/web/packages/vioplot/index.html.
88. McLeay RC, Bailey TL. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. BMC Bioinform. 2010;11:165.
89. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015;43:W39–49.
90. Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016;44:D110–5.
91. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–5.
92. Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics. 2015;31:1127–9.