Title

Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications

Author List

Magdalena J. Koziol[1,2], Charles R. Bradshaw[1*], George E. Allen[1*], Ana S.H. Costa[3*], Christian Frezza[3], John B. Gurdon[1,2]

Institutions

1- Wellcome Trust Cancer Research UK Gurdon Institute, University of Cambridge

2- Department of Zoology, University of Cambridge

3- Medical Research Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre

Present addresses: Wellcome Trust Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, United Kingdom (M.J.K., C.R.B., G.E.A., J.B.G.), Department of Zoology, University of Cambridge, Cambridge, United Kingdom (M.J.K., J.B.G.), Medical Research Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre, Cambridge, United Kingdom (A.S.H.C., C.F.).

*These authors contributed equally to this work

Correspondence should be addressed to M.J.K. (m.koziol@gurdon.cam.ac.uk).

Abstract

Methylation of cytosine deoxynucleotides (dC^{5m}) is a well-established epigenetic mark, but in higher eukaryotes much less is known about modifications affecting other deoxynucleotides. Here, we report the detection of N-6-methyl-deoxyadenosine (dA^{6m}) in vertebrate DNA, specifically in *Xenopus laevis*, but also in other species including mouse and human. Our methylome analysis reveals that dA^{6m} is widely distributed across the eukaryotic genome, is present in different cell types, but commonly depleted from gene exons. Thus, direct DNA modifications might be more widespread than previously thought.

More than 60 years ago, it was discovered that cytosine deoxynucleotides could be methylated in eukaryotic genomic DNA[1]. Since then, dC^{5m} has been extensively studied, revealing it as a major genetically heritable regulatory modification for gene transcription[2–4]. Up to now, not much is known in higher eukaryotes about modifications affecting other deoxynucleotides. In contrast, RNA, a molecule that is built up from similar molecules as DNA, is known to have more than 60 modifications in eukaryotes, and when including different organisms, the number is greater than 110 (ref. 5). Due to the strong similarity between the DNA and RNA building blocks, we found it surprising that higher eukaryotic DNA is not known to be diverse. In order to determine if there are other direct DNA modifications, we used dA^{6m} as an example to discover if the higher eukaryotic genome is more diverse than previously thought.

Methylation of deoxyadenosines has been identified and is a well-described epigenetic feature in bacteria. In these prokaryotes, dA^{6m} is known to regulate various biological pathways such as the restriction-modification system, replication, repair, transcription and transposition[6–11]. Two reports, using restriction enzyme digests, suggested that dA^{6m} might exist in higher eukaryotes, but no direct evidence or global pattern has ever been reported[12,13]. Other initial analytical approaches to assess the presence of dA^{6m} in higher eukaryotes were unsuccessful, possibly because these approaches were constrained by the detection limit of 0.1–0.01% of total deoxynucleotides[14–16]. Only very recently, it was reported that dA^{6m} is present in the genome of the algae *Chlamydomonas*, in the insect *Drosophila* and in the nematode *C. elegans*[17–19]. In contrast to that work, we focused on higher eukaryotes instead.

Here, we report the identification of dA^{6m} in higher eukaryotes, using the same approach used by the other reports. However, ours was developed independently, before the recent publications concerning this modification were published. In order to determine the presence and distribution of dA^{6m} in eukaryotic genomes, we used dot blots, ultra-high performance liquid chromatography-tandem mass spectrometry (UHPLC-MS/MS) and applied a dA^{6m} enrichment approach. Using an antibody against dA^{6m} (dA^{6m} Ab), we carried out DNA immunoprecipitation (DIP) to enrich for genomic DNA fragments containing dA^{6m} that allowed us to identify and describe dA^{6m} genome-wide (Fig. 1a). Here, we identified dA^{6m} not only in the genomes of the frog *X. laevis*, but also in all genomes we analyzed, such as the mouse *M. musculus* and human tissues. We showed that this mark is widely distributed across the genome, but is depleted in exonic regions, and appeared to have a preference for TAGG sites, and possibly contain AG as a core motif.

Results

To identify dA^{6m} in higher eukaryotic genomes, we applied an antibody enrichment approach. First, we verified that an antibody reported to bind to methylated adenosines could in fact recognize the dA^{6m} modification[20]. Dot blot experiments and DIP using synthetic oligonucleotides confirmed that this Ab indeed recognizes dA^{6m} (Supplementary Fig. 1a–c). We then asked if the *X. laevis* sperm genome contains dA^{6m}. To address this, we isolated DNA from different samples and removed all proteins and

RNA. We performed dot blots with *X. laevis* sperm genomic DNA and stained with the dA^{6m} Ab (Fig. 1b). Importantly, we detected a dA^{6m} signal with the dA^{6m} Ab on *X. laevis* sperm genomic DNA (Fig. 1b and Supplementary Fig. 1d–h). As controls, we used bacterial genomes from deoxyadenosine methylase (Dam) positive (Dam+) and negative (Dam–) bacteria. We detected dA^{6m} not only in Dam+ bacteria, but also in Dam– bacteria (Fig. 1b). The dA^{6m} signal in Dam– bacteria could be explained by the presence of the other deoxyadenosine methylase EcoKI, which maintains some level of dA^{6m} in the genome even in the absence of Dam[21,22].

Genomes of higher eukaryotes contain dA^{6m}

To further confirm the results from the dot blot screen, genomic DNA was digested into its individual nucleosides and analyzed by UHPLC-MS/MS (Fig. 1a). As a positive UHPLC-MS/MS reference, we used a synthetic dA^{6m} standard dilution series, as well as a water negative control. dA^{6m} was identified in a given sample only when the retention time as well as its fragmentation pattern both matched the synthetic dA^{6m} standard. Analogous to the dot blot results, dA^{6m} was detected in both Dam+ and Dam– bacteria controls. As expected, the level of dA^{6m} differed between these two bacteria. We encountered a lower level of dA^{6m} in Dam– bacteria in comparison to Dam+ bacteria. Importantly, we did not detect dA^{6m} in our processed negative control, but detected dA^{6m} in the processed DNA isolated from eukaryotic tissues (Fig. 1c and Supplementary Fig. 2a–d). These results substantiate the dot blot approach and strongly support the presence of dA^{6m} in the genome of a higher eukaryotic organism.

We next tested if the dA^{6m} Ab can in fact enrich for dA^{6m}. We carried out dA^{6m} Ab DIP on sheared *X. laevis* DNA. The DNA recovered from the dA^{6m} Ab DIP was then further processed into its individual nucleosides and analyzed by UHPLC-MS/MS. The results validated that the dA^{6m} Ab DIP strongly enriches for the low level of dA^{6m} in higher eukaryotes, namely 14,152 times under the conditions applied (Fig. 1d–e and Supplementary Fig. 2a–c, e–g). To estimate the abundance of dA^{6m} in the higher eukaryotic genome, we used the data obtained from the non-enriched dA^{6m} Ab DIP samples. Our results show that dA^{6m} is found 1 in 84 dA in Dam+ bacteria (1.19%), 1 in 4,215 dA in Dam– (0.02%) bacteria and only 1 in 1,172,141 dA (0.00009%) in higher eukaryotic samples (Fig. 1f). This corresponds to 27,238 dA^{6m} in Dam+ bacteria, 542 dA^{6m} in Dam– bacteria and 1,654 dA^{6m} in one *X. laevis* genome, or 6,616 dA^{6m} in one *X. laevis* tetraploid cell.

To determine if dA^{6m} is only a feature of *X. laevis* testes or if it is present in other higher eukaryotes, we extended our dot blot screen to search for the presence of dA^{6m} in other organisms. Our results suggest that dA^{6m} is not only present in various *X. laevis* tissues, but is also found in all higher eukaryotes we tested, such as in *D. rerio, M. musculus* and tissue culture cells derived from mouse and humans (Fig. 1g). We decided to focus our studies on *X. laevis* and used *M. musculus* to generalize our findings for higher eukaryotes.

Few genes are associated with dA^{6m}

To study the location and distribution of dA^{6m} containing regions across the genome, we generated high throughput sequencing libraries (Seq) from dA^{6m} Ab DIP-enriched and input fractions (dA^{6m} Ab DIP-Seq and input-Seq, respectively). We analyzed the genomes of *X. laevis* testes, fat and oviduct, and of *M. musculus* kidney by dA^{6m} Ab DIP-Seq (Fig. 2a–d, Supplementary Table 1). For all *X. laevis* tissues, we processed 2 biological replicates that were obtained from different animals. In the case of *M. musculus*, we used 3 biological replicates that were also isolated from different animals. We compared the dA^{6m} Ab DIP-Seq to the corresponding input-Seq controls in order to determine which regions in the genome were enriched by the dA^{6m} Ab, hence, contained the dA^{6m} mark. Based on our dA^{6m} Ab DIP-Seq data, we identified in total 27,374 dA^{6m} peaks in *X. laevis* testes, 20,160 in oviduct, 47,834 in fat, and 27,374 in *M. musculus* kidney (Fig. 2d, Supplementary Table 2). In dA^{6m} Ab DIP-Seq experiments, dA^{6m} peaks obtained from different cell types add up, aberrantly increasing the total abundance of dA^{6m}. Therefore, such peak data should be used only to estimate the distribution of dA^{6m} genome wide, rather than to determine the absolute levels of dA^{6m} in the tissue. To determine if our sequencing data is of good quality for subsequent genome-wide analyses, we determined if it is consistent and reproducible.

By comparing samples to each other using the Pearson Correlation Coefficient and scatter plots, we showed that biological replicates (tissues from different animals) as well as experimental conditions such as pulldown and input correlate with each other more than

between different experimental conditions or between different biological sources, corroborating the robustness of our sequencing data (Fig. 2b–c). Further, we determined the number of identified dA^{6m} peaks in individual replicates, and asked how many of these peaks overlap between biological replicates (Fig. 2d). First, the number of peaks identified in biological replicates is similar, which supports reproducibility (Fig. 2d). Second, the overlap between all biological replicates (tissues from different animals) is much higher than one would expect at random (χ2-test, **P-value $< 1 \times 10^{-16}$), strengthening the reproducibility of our approach. Next, we determined the overlap of dA^{6m} peaks between different tissues (Fig. 2d). We took the overlapping peaks between replicates of *X. laevis* testes, oviduct and fat, and overlapped them between the different tissues. We found that some of them overlapped between all tissues, suggesting that some dA^{6m} peaks are present at the same location in the genome, irrespective of tissue type. This is in particular true for *X. laevis* oviduct, where 85% of the dA^{6m} peaks identified seem to also be present in at least one of other tissues analyzed, namely fat and testes (Fig. 2d). However, many dA^{6m} peaks seem also to be tissue type specific. For example, 52% (7,207) of all dA^{6m} peaks identified in fat are only found in *X. laevis* fat, while the remainder 48% (6,620) are also present in testes, in oviduct or in both (Fig. 2d). Overall, we conclude that some of the dA^{6m} peaks are the same in different cell types, but many are different, indicating some degree of cell type specificity.

dA^{6m} is predominantly excluded from coding regions

Despite the high number of dA^{6m} peaks identified, only a small fraction of all genes have a dA^{6m} peak. This was observed in all tissues and in samples from both *X. laevis* and *M. musculus*. Between 6.7% and 20.6% of all genes have a dA^{6m} peak, while the rest of the peaks lie in non-genic regions (Supplementary Table 3). In *X. laevis*, the few genes that are found to be associated with dA^{6m} are strongly linked to pathways such as nucleic acid binding, metabolic processes and transcription, as determined by gene ontology analysis. This was found across all tissues (Supplementary Table 4). In contrast, the genes that are associated with dA^{6m} in *M. musculus* kidney are linked to different pathways, for example to ion channel activity, cell adhesion and ATP binding (Supplementary Table 4). The different pathways found in *M. musculus* could either indicate a tissue specific role of dA^{6m} in kidney, or be due to the possibility that dA^{6m} regulates different pathways in *M. musculus* than it does in *X. laevis*. When we analyzed the gene regions further, we observed that few dA^{6m} peaks are located in exonic regions. Only 0.1–0.6% of all exons have dA^{6m} peaks. In contrast, dA^{6m} peaks are more frequent within introns. We found that 6.4–17.6% of all genes have dA^{6m} peaks in introns (Supplementary Table 3). This lack of dA^{6m} in exonic regions is in accordance with transcriptional start site (TSS) plots (Fig. 3a, Supplementary Figure 3a). The TSS plots showed a strong decrease of dA^{6m} levels just after the TSS of genes. In addition, occasionally, a small increased abundance of dA^{6m} upstream of TSS was detected, in comparison to the more downstream 3' region. This TSS plot pattern, where a strong decrease of dA^{6m} level is observed just after the TSS, was encountered in all *X. laevis* tissues analyzed and also in all *M. musculus* kidney samples (Fig. 3a, Supplementary Figure 3a). This suggests that the absence of dA^{6m} in coding regions might be a general feature of dA^{6m} in higher eukaryotes.

To obtain a better understanding of the dA^{6m} distribution, we further analyzed the abundance of dA^{6m} in the vicinity of genes. We divided regions that are in the vicinity of genes into different groups, for example those consisting of 1kb areas upstream and downstream of coding genes, and those that distinguish exons and introns. We next determined the ratio of methylated versus non-methylated deoxyadenosines in these regions, based on our DIP-Seq data. As a control, we also analyzed the dC^{5m} distribution in the same way[23]. Our analysis revealed an enrichment of dC^{5m} in exons (Fig. 3b), but in contrast to this we observed depletion of dA^{6m} marks in exonic regions in all *X. laevis* testes replicates (Fig. 3c, Supplementary Fig. 3b). Further, this observation was confirmed in all *X. laevis* and *M. musculus* tissues analyzed, and in all replicates (Supplementary Fig. 3c–e). This suggests that depletion of dA^{6m} in exonic regions is a distinct feature of this epigenetic modification.

To further corroborate our findings, we carried out DIP-Seq on *X. laevis* testes with 2 other antibodies that are known to recognize dA^{6m}. These are referred to as dA^{6m} Ab* and dA^{6m} Ab**. Importantly, dA^{6m} signals identified were irrespective of the antibody used, excluding an antibody bias. As a control, we used the corresponding input, but also IgG for further validations (Fig. 4, Supplementary Fig. 4, Supplementary Table 1, Supplementary Table 5). Our analysis showed that irrespective of the dA^{6m} recognizing antibody used, and independently of whether we compared our dA^{6m} recognizing antibodies to input or IgG controls, the distribution of dA^{6m} remained the same

(Supplementary Fig. 4). In all cases we found that the level of dA^{6m} decreased in exons, strengthening our previous dA^{6m} DIP-Seq results.

Next, we asked if our dA^{6m} peaks are conserved. For this purpose, PhyloP scores across 30 vertebrate species were compared to our dA^{6m} *M. musculus* data[24]. This analysis was not possible for *X. laevis*, as PhyloP data is not available for this species. Using the Top 300 dA^{6m} peak overlaps between the kidney *M. musculus* replicates, we found that the conservation score means of dA^{6m} enriched regions (0.08) are smaller than and differ significantly (**P-value < 2.2×10^{-16}) from the scores when the dA^{6m} enriched regions were shifted by 10kb (0.12). This suggested that although there is some conservation, it was relatively weak. Bearing in mind that most of our peaks were excluded from coding regions, which were considered conserved, it is not too surprising that the dA^{6m} peak regions showed weak conservation. This in fact confirmed our previous observations.

"AG" could be a putative consensus site for dA^{6m}

We then wanted to identify putative dA^{6m} consensus sequence motifs. To verify our approach, we first tested the abundance of any 4bp motifs in bacteria. We have carried out dA^{6m} Ab DIP-Seq and input-Seq experiments on Dam+ and Dam– *E. coli* genomes and identified dA^{6m} peaks (Supplementary Table 6). We then asked how abundant any 4bp motif is in these peaks. Out of the 256 possible combinations, we found enrichment for the GATC sequence in the dA^{6m} peaks of Dam+ bacteria (Fig. 5a). Importantly, this GATC sequence is the known target recognition sequence of the Dam methylase[25]. In

Dam– bacteria, this GATC motif, as expected, is no longer the most abundant motif encountered in dA^{6m} peaks (Fig. 5a–b). We also applied the MEME prediction program to the bacterial dA^{6m} peaks that were at least enriched by a magnitude of 2 (ref. 26). This analysis confirmed the GATC Dam motif in Dam+, but not in Dam– bacteria (Fig. 5c). This further confirmed the validity of our data and approach to predict the consensus sequence of dA^{6m}. Next, we tried to identify potential dA^{6m} consensus sequences for all our *X. laevis* and *M. musculus* samples. Using MEME, we obtained significant putative 8bp consensus motifs for all *X. laevis* tissues (**E-value < 1.2×10^{-8}) (Fig. 5d). Also, forcing a shift of dA^{6m} peaks by 5kb led to an inability of MEME to identify these sequences, showing that these putative motifs were not identified at random (Supplementary Fig. 5). Our analysis was not successful on *M. musculus*, as it seemed to be embedded in sequences that are repetitive. This made it impossible to reliably predict a motif in *M. musculus*, even when we tried to remove repeats from the input sequences for MEME analysis. As a result, we decided to focus on *X. laevis*. Interestingly, overlapping all of the tissues gives the putative dA^{6m} 8bp motif TAGGAAGG (**E-value < 6.7×10^{-141}) (Fig. 5d). This sequence was very similar to the ones identified in individual tissues, suggesting that this might be or contain a basic motif that is present in all tissue types. The sequences found by MEME in different tissues were variable enough to make us believe that the core motif was shorter. To determine if we can narrow down the 8bp putative motif to 4bp, we determined, as we have done with bacteria, how abundant any 4bp motif is in our peaks. However, out of the 256 possible combinations, not all are feasible as a potential motif, as the putative motif should at least in part overlap with the 8bp motif that was generated with MEME and has to contain a

deoxyadenosine[26]. We calculated the frequency of all 256 possible 4bp in the peaks, and the noise, namely the frequency one would expect under those peaks when they were shifted (Supplementary Table 7). The ratio of those revealed that TAGG is likely to be a potential motif, as it is in top 4 most enriched sequences in all overlapped peaks. The other most enriched sequences are not applicable, as they are not found by MEME to be statistically significant (E-value > 0.05). Interestingly, we did not identify the bacterial motif GATC, also confirming we did not have bacterial contamination in our eukaryotic datasets. Although MEME and our 4 base predictions showed that the TAGG sequence is enriched under our peaks, we are hesitant to claim this is a consensus motif. However, all our results point towards part of it being a consensus. We therefore postulate that AG, with the major fraction being TAG, forms part of the motif. However, further elaborate experimental evidence is required to determine if TAG or AG is in fact a *bona fide* consensus motif in which dA^{6m} is found.

Discussion

Epigenetic modifications can cause changes to the genome without altering the DNA sequence. These are known to occur on histones, RNA and DNA. Most of the epigenetic modifications studied to date are those of histone and RNA modifications. Both molecules can bind to specific DNA sequences and subsequently change the accessibility of that region, but do not directly modify the DNA itself. Up to date, only dC^{5m} has been studied extensively in higher eukaryotes, which directly affects the DNA itself[2–4]. Although intermediate forms of dC^{5m}, such as 5-formylcytosine and 5-carboxylcytosine

have been discovered and are increasingly being studied, not much is known in higher eukaryotes about modifications affecting other deoxynucleotides[27,28]. We found it surprising that so little attention has been given to direct epigenetic modifications. In order to determine if there are in fact no other modifications, we used dA^{6m} as an example and discovered that the higher eukaryotic genome is more diverse than previously thought.

dA^{6m} is a modification found in bacterial DNA and affects gene expression and virulence[6,10]. However, its presence in higher eukaryotes has been debated[12,13]. Its identification was likely constrained by the low abundance of this modification[14–16]. With technological advancements, the detection limits improved, allowing us to directly identify dA^{6m} in the genome of higher eukaryotes. Very recently, other reports were published making observations similar to ours, using the same approach[17–19]. However, ours was developed independently, before the recent publications concerning this modification were published. We have discovered dA^{6m} in higher eukaryotic organisms, while the recent publications reported the presence of dA^{6m} in the genome of the algae *Chlamydomonas*, in the insect *Drosophila* and in the nematode *C. elegans*[17–19]. In agreement to previous work, we find that dA^{6m} is a low abundant modification, even less abundant in higher eukaryotes than in other organisms. For example, in *Drosophila*, the frequency of the dA^{6m} /dA ratio varies between 0.07–0.001%, in *C. elegans* between 0.01–0.4%, and in *Chlamydomonas* it is 0.4% (refs. 17–19). In the organisms that we investigated, dA^{6m} was substantially less abundant, namely 0.00009%. The reason for the difference in abundance of dA^{6m} among these organisms is unclear, and may be ascribed

to inherent differences in genome organization and epigenetic regulators. The high abundance of dA^{6m} in *C. elegans* might also be explained by the fact that these animals were fed with Dam– bacteria that still contained dA^{6m}. Indeed, the genome of these bacteria still possess residual dA^{6m} due to the presence of the other known deoxyadenosine methylase EcoKI, which we have also confirmed by UHPLC-MS/MS[121,22]. This might have interfered with the determination of total dA^{6m} levels, as well as any functional tests performed in the presence of these bacteria.

When comparing the genome wide distributions of dA^{6m} in *M. musculus* and *X. laevis* to the other organisms, we encountered a pattern different from what we saw in the in higher eukaryotes. Indeed, we found that dA^{6m} is absent from areas downstream of TSS and from exons in mouse and frog genomes. In *C. elegans*, no appreciable distinct pattern near genes is observed[19]. In contrast, in *Drosophila* and *Chlamydomonas* genomes, dA^{6m} is enriched at or following TSS sites[17,18]. This is the opposite of what we found in mouse and frogs. This different pattern of dA^{6m} suggests that this modification may have distinct roles across eukaryotes. However, any functions of dA^{6m} in higher eukaryotes remain to be investigated. Key aspects of this investigation are the identification of epigenetic modifiers that deposit (methylase) and remove (demethylases) the modification, and of possible dA^{6m} interacting proteins. The latter aspect is of particular importance since dA^{6m} might serve as a DNA anchor for regulatory proteins to bind, which could then trigger various downstream pathways and regulate gene transcription. Also, the presence of dA^{6m} could ultimately cause different chromatin landscapes, influence nucleosome positioning, or insulate different DNA regions from each other.

Overall, our findings suggest that direct epigenetic modifications might be more widespread than previously thought in higher eukaryotes. RNA, a molecule that is built up from similar molecules as DNA, is known to have more than 60 modifications in eukaryotes, and when including different organisms, the number is greater than 110 (ref. 5). Due to the strong similarity between the DNA and RNA building blocks, we have shown that DNA is much more diverse than has been previously believed. Overall, we believe it is very unlikely that DNA is so simplistic while RNA is so diverse. Hence, we hypothesize that many of such 'apegenetic' (from Greek *apeftheias*, meaning direct) modifications exist. In future, this 'apegenome' remains to be discovered and its function further investigated.

Accession codes DNA sequencing data has been deposited in the NCBI GEO (Gene Expression Omnibus) database with the deposition ID GSE74184. The UHPLC-MS/MS data has been deposited in the MetaboLights database with the deposition ID MTBLS276.

Acknowledgements M.J.K. was supported by the Long-Term Human Frontiers Fellowship (LT000149/2010-L), the Medical Research Council grant (G1001690), and by the Isaac Newton Trust Fellowship (RG76588). The work was sponsored by the Biotechnology and Biological Sciences Research Council grant BB/M022994/1 (J.B.G. and M.J.K.). The Gurdon laboratory is funded by the grant 101050/Z/13/Z (J.B.G.) from the Wellcome Trust, and is supported by the Gurdon Institute core grants, namely by the Wellcome Trust Core Grant (092096/Z/10/Z) and by the Cancer Research UK Grant (C6946/A14492). C.R.B. and G.E.A. are funded by the Wellcome Trust Core Grant. We are grateful to D. Simpson and R. Jones-Green for preparing *X. laevis* eggs and oocytes, F. Miller for providing us with *M. musculus* tissue, T. Dyl for *X. laevis* eggs and *D. rerio* samples, and to Gurdon laboratory members for their critical comments. We thank U. Ruether for providing us with *M. musculus* kidney DNA (Entwicklungs- und Molekularbiologie der Tiere, Heinrich Heine Universitaet Duesseldorf, Germany). We also thank J. Ahringer, S. Jackson, A. Bannister and T. Kouzarides for critical input and advice, M. Sciacovelli and E. Gaude for suggestions.

Author Contributions A.S.H.C. performed all UHPLC-MS/MS analyses. M.J.K. conceived the study, designed and performed all experiments, analyzed the data, supervised all research and wrote the paper. C.R.B. and G.E.A. performed the bioinformatic analyses, developed ideas and helped to generate figures and to write the paper. C.F. advised on and supervised all UHPLC-MS/MS analyses and both A.S.H.C. and C.F. helped to design UHPLC-MS/MS studies and to write the paper. J.B.G. assisted with writing the paper, and supervised all research. Correspondence should be addressed to M.J.K. (m.koziol@gurdon.cam.ac.uk).

References

1. Hotchkiss, R.D. The quantitative separation of purines, pyrimidines and nucleosides by paper chromatography. *J Biol Chem* **175**, 315–332 (1948)
2. Gruenbaum, Y., Naveh-Many, T., Cedar, H., Razin, A. Sequence specificity of methylation in higher plant DNA. *Nature* **292**, 860–862 (1981)
3. Boyes, J. & Bird, A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64**, 1123–1134 (1996)
4. Jones, P.A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* **3**, 415–428 (2002)
5. Cantara, W.A., *et al*. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–201 (2011)
6. Luria, S.E. & Human, M.L. A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol.* **64**, 557–569 (1952)
7. Bertani, G. & Weigle, J.J. Host controlled variation in bacterial viruses. *J. Bacteriol.* **65**, 113–121 (1953)
8. Lu, M., Campbell, J.L., Boye, E., Klecker, N. SeqA: a negative modulator of replication initiation in E.coli. *Cell* **77**, 413–426 (1994)
9. Laengle-Rouault, F., Maenhaut-Michel, G., Radman, M. GATC sequence and mismatch repair in Escherichia coli. *Embo J.* **5**, 2009–2013 (1986)
10. Braaten, B.B., Nou, X., Kaltenbach, L.S., Low, D.A. Methylation patterns in pap regulatory DNA control pyelonephritis-associated pili phase variation in E. coli. *Cell* **76**, 577–588 (1994)

11. Roberts, D., Hoopes, B.C., McClure, W.R., Klecker, N. IS10 transposition is regulated by DNA adenine methylation. *Cell* **43**, 117–130 (1985)
12. Kay, P.H., *et al.* Evidence for adenine methylation within the mouse myogenic gene Myo-D1. *Gene* **151**, 89–95 (1994)
13. Reyes, E.M., Camacho-Arroyo, I., Nava, G., Cerbon, M.A. Differential methylation in steroid 5 alpha-reductase isozyme genes in epididymis, testis, and liver of the adult rat. *J. Androl.* **18**, 372–377 (1997)
14. Vanyushin, B.F., Tkacheva, S.G., Belozersky, A.N. Rare bases in animal DNA. *Nature* **225**, 948–949 (1970)
15. Lawley, P.D., Crathorn, A.R., Shah, S.A., Smith, B.A. Biomethylation of deoxyribonucleic acid in cultured human tumour cells (HeLa). Methylated bases other than 5-methylcytosine not detected. *Biochem. J.* **128**, 133–138 (1972)
16. Gunthert, U., Schweiger, M., Stupp, M., Doerfler, W. DNA methylation in adenovirus, adenovirus-transformed cells, and host cells. *PNAS* **73**, 3923–3927 (1976)
17. Zhang, G., *et al.* N6-Methyladenine DNA Modification in Drosophila. *Cell* 161, 1–14 (2015)
18. Fu, Y., *et al.* N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. *Cell* 161, 1–14 (2015)
19. Greer, L.E., *et al.* DNA Methylation on N6-Adenine in C. elegans. *Cell* 161, 1–11 (2015)

20. Munns, T. W., Liszewski, M.K., Sims, H.F. Characterization of antibodies specific for N6-methyladenosine and for 7-methylguanosine. *Biochemistry* **16**, 2163–2168 (1977).

21. Marinus, M.G. & Morris, N.R. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J. Bacteriol.* **114**, 1143–1150 (1973)

22. May, M.S. & Hattman, S. Analysis of bacteriophage deoxyribonucleic acid sequences methylated by host- and R-factor-controlled enzymes. *J. Bacteriol.* **123**, 768–770 (1975)

23. Kobayashi, H., *et al.* Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks. *PLoS Genetics* **8**, e1002440 (2012)

24. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010)

25. Geier, G. & Modrich, P. Recognition sequence of the dam methylase of Escherichia coli K12 and mode of cleavage of DpnI endonuclease. *J. Biol. Chem.* **254**, 1408–1413 (1979)

26. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on ISMB, AAAI Press,* pp. 28–36 (1994)

27. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011)

28. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.* 50, 7008–7012 (2011)

29. Ford, E., Nikopoulou, C., Kokkalis, A., Thanos, D. A method for generating highly multiplexed ChIP-seq libraries. *BMC Res. Notes* **7**, 312 (2014)

Figure Legends

Figure 1. Identification of dA^{6m} in the genome of higher eukaryotes. a, Illustration of dA^{6m} identification. dA^{6m} was identified using dot blots, UHPLC-MS/MS and dA^{6m} Ab DIP sequencing (DIP-Seq). **b**, Dot blot with dA^{6m} Ab on DNA templates. **c-d**, Representative (1 out of 4) UHPLC-MS/MS extracted ion chromatogram (EIC) (left) and fragmentation spectrum (right) monitoring presence of dA^{6m} in genomic DNA from *X. laevis*. * indicates parent ion, AU=arbitrary units, m/z=mass to charge ratio, n=4 biological replicates (tissues from different animals). **e**, Percentage of dA^{6m} versus total deoxyadenosines in DNA from different samples following dA^{6m} Ab DIP enrichment. Error bars, s.e.m. n=4 tissues from different animals or from independent bacterial cell cultures, **P≤0.005, *P≤0.05, two-sided t-test. **f**, Percentage of dA^{6m} versus total deoxyadenosines in DNA from different samples without dA^{6m} Ab DIP enrichment. Error bars, s.e.m., n=4 tissues from different animals or from independent bacterial cell cultures, **P≤0.005, *P≤0.05, two-sided t-test, but when one sample was zero the one-sided t-test was applied. **g**, dA^{6m} dot-blot of DNA from different higher eukaryotic sources, n=3 technical replicates.

Figure 2. Genome wide identification of dA^{6m} marks in *X. laevis* fat, oviduct, testes and *M. musculus* kidney. a, dA^{6m} peak signal tracks. One gene region for each tissue type is shown. The y-axis represents the amount of immunoprecipitated DNA at each position normalized by the total number of reads. RefSeq gene annotations are shown. n=2 for each *X. laevis* tissue and n=3 for *M. musculus*, biological replicates from different animals, AU=arbitrary units. **b**, Heat map of Pearson correlation coefficient values from comparisons between *X. laevis* and *M. musculus* dA^{6m} DIP-Seq and input-Seq samples. Correlation was calculated pairwise for all samples, excluding windows where both samples in the pair had zero depth. n=2 for each *X. laevis* tissue and n=3 for *M. musculus*, biological replicates from different animals, $^{**}P<1\times10^{-16}$, Pearson Correlation test on mapped reads. **c**, Scatter plots comparing individual samples pairwise for overlapping peaks. The enrichment score ranging from 0–8 is plotted on both axes. The color in each plot reflects the correlation value (Pearson correlation coefficient), which is also shown in the top right corner of each plot. n=2 for each *X. laevis* tissue and n=3 for *M. musculus*, biological replicates from different animals, $^{**}P<1\times10^{-16}$, two sided t-test on dA^{6m} peaks. **d**, Number of unique and overlapping dA^{6m} peaks identified in *X. laevis* and in *M. musculus* tissues. n=2 for each *X. laevis* tissue and n=3 for *M. musculus*, biological replicates from different animals, $^{**}P<1\times10^{-16}$, χ2-test on dA^{6m} peaks.

Figure 3. Genome wide distribution of dA^{6m} in the vicinity of genes. a, Distribution of dA^{6m} peaks around TSS, from 20kb 5' to 20kb 3', identified in *X. laevis* fat, oviduct, testes and in *M. musculus* kidney. One biological replicate from one animal is shown in each graph. **b**, Density of dC^{5m} versus unmethylated dC in distinct areas of the *M. musculus* testes genome. Only the regions shown in dark grey are statistically significant. One biological replicate from one animal is shown, **P<0.007, binomial test on dA^{6m} peaks. **c**, Density of dA^{6m} versus unmethylated dA in distinct areas of the genome of *X. laevis* testes. Only the regions shown in dark grey are statistically significant. One biological replicate from one animal is shown, *P<0.03, binomial test on dA^{6m} peaks.

Figure 4. Genome wide distribution of dA^{6m} peaks in *X. laevis* testes samples as determined with three different dA^{6m} recognizing antibodies in DIP-Seq and comparison to input-Seq or IgG-Seq controls. a, dA^{6m} peak signal tracks. Tracks obtained from dA^{6m} Ab DIP-Seq, dA^{6m} Ab* DIP-Seq, dA^{6m} Ab** DIP-Seq and control input-Seq and IgG-Seq are shown. One gene region for each biological replicate is shown. The y-axis of each profile represents the amount of reads at each position normalized by the total number of reads in a given dataset. RefSeq gene annotations are shown. n=2, biological replicates from different animals, AU=arbitrary units. **b**, Heat map of Pearson correlation coefficient values from comparisons between *X. laevis* testes dA^{6m} Ab DIP-Seq, dA^{6m} Ab* DIP-Seq, dA^{6m} Ab** DIP-Seq, IgG DIP-Seq and input-Seq samples. Correlation was calculated pairwise for all samples, excluding windows where both samples in the pair had zero depth. n=2, biological replicates from different animals, **$P<1\times10^{-16}$, Pearson Correlation test on mapped reads. **c**, Scatter plots comparing individual samples pairwise for overlapping peaks. The enrichment score ranging from 0–8 is plotted on both axes. The color in each plot reflects the correlation value (Pearson correlation coefficient), which is also shown in the top right corner of each plot. n=2, biological replicates from different animals, **$P<1\times10^{-16}$, two sided t-test on dA^{6m} peaks.

Figure 5. dA^{6m} motif identification in bacteria and *X. laevis*. **a**, Abundance of 4bp motif in dA^{6m} peaks. 256 potential 4bp motifs, each representing one column along x-axis, were ranked for their abundance in Dam+ and Dam– bacterial dA^{6m} peaks. The Dam recognition motif GATC is shown in red. n=1 biological replicate from one bacterial culture for Dam+ and Dam- bacteria, Spearman rank correlation coefficient=0.94 between Dam+ and Dam– bacteria, Spearman's **$\rho=2.2\times10^{-16}$. **b**, Ratio between Dam+ and Dam– bacteria enriched 4bp motifs. Only motifs that are at least 5% enriched are illustrated. **c**, Bacterial dA^{6m} motif identified by MEME. In Dam+ bacteria, the motif GATC has been identified (73 out of 73 gene regions identified by MEME). n=1 biological replicate from one bacterial culture for Dam+ and Dam- bacteria, AU=arbitrary units. **d**, *X. laevis* dA^{6m} motif identification by MEME. Overlaps between biological replicates from different animals were used for analysis. Same tissue n=2 biological replicates from different animals, different tissue overlaps n=6 biological replicates for different animals, **E-value$<1.2\times10^{-8}$, statistics by MEME, AU=arbitrary units.

Supplementary Table 1. Summary of samples sequenced.

Source	Sample / biological replicates	Mapped Reads (Mil)
X. laevis testes	IP dA^{6m}-rep 1	7.6
	input-rep 1	16.9
	IP dA^{6m} Ab*-rep 1	10.5
	IP dA^{6m} Ab**-rep 1	10.4
	IP IgG-rep 1	0.4
	IP dA^{6m}-rep 2	7.8
	input-rep 2	18.9
	IP dA^{6m} Ab*-rep 2	9.9
	IP dA^{6m} Ab**-rep 2	9.9
	IP IgG-rep 2	0.2
X. laevis fat	IP dA^{6m}-rep 1	17.2
	input-rep 1	28.3
	IP dA^{6m}-rep 2	22.8
	input-rep 2	31.8
X. laevis oviduct	IP dA^{6m}-rep 1	23.5
	input-rep 1	40.6
	IP dA^{6m}-rep 2	28.6
	input-rep 2	33.3
M. musculus kidney	IP dA^{6m}-rep 1	21.1
	input-rep 1	27.8
	IP dA^{6m}-rep 2	22.3
	input-rep 2	26.3
	IP dA^{6m}-rep 3	36.8
	input-rep 3	28.1
E. coli Dam+	IP dA^{6m}-rep 1	1.2
	input-rep 1	0.2
E. coli Dam-	IP dA^{6m}-rep 1	0.8
	input-rep 1	0.4

Supplementary Table 1. Summary of samples sequenced. *X. laevis*, *M. musculus* and *E. coli* bacterial genomes were sequenced using a various number of replicates. IP samples represent Ab pulldown of sonicated, sheared genomic DNA. The corresponding input or IgG sample within one biological replicate serves as a control for the IP experiment.

Supplementary Table 3. Percentage of regions with dA^{6m} peak.

Presence of dA^{6m}	*X. laevis* Testes	*X. laevis* Fat	*X. laevis* Oviduct	*M. musculus* Kidney	*Total*
Gene region with at least one dA^{6m} peak (Region: 1kb 5' to 1kb 3' of gene including introns)	7.6-12.8%	12.5-20.6%	7.5-9.4%	6.7-9.2%	6.7-20.6%
Genes with at least one dA^{6m} peak in exons	0.2-0.5%	0.5-0.6%	0.3%	0.1-0.2%	0.1-0.6%
Genes with at least one dA^{6m} peak 5kb upstream (Region: 0-5kb 5' to gene)	2.6-3.9%	4.1-7.6%	2.3-2.8%	1.1-1.8%	1.1-7.6%
Genes with at least one dA^{6m} peak 1kb upstream (Region: 0-1kb 5' to gene)	0.4-0.7%	0.7-1.3%	0.4-0.5%	0.1%	0.1-1.3%
Genes with at least one dA^{6m} peak 1kb downstream (Region: 0-1kb 3' to gene)	0.4-0.7%	0.6-1.1%	0.3-0.4%	0.1-0.3%	0.1-1.1%
Genes with at least one dA^{6m} peak in introns	6.6-10.9%	10.7-17.6%	6.5-8.3%	6.4-8.6%	6.4-17.6%
Genes with one dA^{6m} peak in introns	4.0-8.3%	8.3-12.3%	5.5-6.8%	4.7-6.0%	4.0-12.3%
Genes with at least two dA^{6m} peaks in introns	1.1-2.6%	2.4-5.2%	1.0-1.5%	1.7-2.7%	1.0-5.2%

Supplementary Table 3. Percentage of regions with dA^{6m} peak. The percentage of dA^{6m} peaks in different regions in and near the vicinity of genes is shown for different biological samples. The range of the percentage covers the percentage obtained from the biological replicates.

Supplementary Table 4. Summary of GO term analysis.

	GO terms	Observed/Expected Hits (Overrepresented: >1; Underrepresented: <1)	Adjusted P-value
X. laevis testes	Nucleic acid binding	Underrepresented	1.67×10^{-8}
(2 replicate overlap)	Nucleobase, nucleoside, nucleotide & nucleic acid metabolic process	Underrepresented	7.81×10^{-7}
	RNA binding protein	Underrepresented	7.92×10^{-7}
	Binding	Underrepresented	5.42×10^{-6}
	Metabolic process	Underrepresented	1.02×10^{-5}
	Cell communication	Overrepresented	4.57×10^{-5}
	primary metabolic process	Underrepresented	5.26×10^{-5}
	RNA binding	Underrepresented	1.36×10^{-4}
	Transcription	Underrepresented	3.65×10^{-4}
	Signal transduction	Overrepresented	3.65×10^{-4}
	DNA binding	Underrepresented	2.00×10^{-3}
	Transcription from RNA polymerase II promoter	Underrepresented	2.00×10^{-3}
X. laevis fat	Nucleic acid binding	Underrepresented	4.18×10^{-18}
(2 replicate overlap)	Nucleobase, nucleoside, nucleotide & nucleic acid metabolic process	Underrepresented	2.24×10^{-13}
	Binding	Underrepresented	9.30×10^{-11}
	Metabolic process	Underrepresented	9.30×10^{-11}
	Primary metabolic process	Underrepresented	1.69×10^{-10}
	RNA binding protein	Underrepresented	1.69×10^{-10}
	DNA binding	Underrepresented	9.95×10^{-10}
	Transcription	Underrepresented	1.63×10^{-8}
	Transcription from RNA polymerase II promoter	Underrepresented	9.58×10^{-8}
	Transcription factor activity	Underrepresented	1.73×10^{-7}
	Transcription regulator activity	Underrepresented	1.73×10^{-7}
	Transcription factor	Underrepresented	1.73×10^{-7}
X. laevis oviduct	Nucleic acid binding	Underrepresented	2.23×10^{-8}
(2 replicate overlap)	Nucleobase, nucleoside, nucleotide & nucleic acid metabolic process	Underrepresented	4.54×10^{-8}
	Metabolic process	Underrepresented	1.24×10^{-7}
	Primary metabolic process	Underrepresented	2.03×10^{-7}
	RNA binding protein	Underrepresented	1.57×17^{-5}
	Voltage-gated calcium channel activity	Overrepresented	1.67×10^{-5}
	Calcium channel	Overrepresented	1.67×10^{-5}
	Voltage-gated calcium channel	Overrepresented	1.67×10^{-5}
	Transcription	Underrepresented	1.92×10^{-5}
	Cell communication	Overrepresented	4.36×10^{-5}
	Protein complex	Underrepresented	6.11×10^{-5}
	Transcription from RNA polymerase II promoter	Underrepresented	6.27×10^{-5}
M. musculus kidney	Ion channel activity	Overrepresented	6.86×10^{-6}
(3 replicate overlap)	Plasma membrane	Overrepresented	9.73×10^{6}
	Postsynaptic membrane	Overrepresented	1.12×10^{-5}
	Membrane	Overrepresented	1.85×10^{-5}
	Synapse	Overrepresented	2.40×10^{-3}
	Cell adhesion	Overrepresented	4.24×10^{-4}
	ATP binding	Overrepresented	5.12×10^{-4}
	Synaptosome	Overrepresented	9.62×10^{-4}
	Intrinsic to external side of plasma membrane	Overrepresented	1.44×10^{-3}
	Ventricular cardiac myofibril development	Overrepresented	1.44×10^{-3}
	Cell junction	Overrepresented	1.57×10^{-3}
	Extracellular-glutamate-gated ion channel activity	Overrepresented	1.79×10^{-3}

Supplementary Table 4. GO terms associated with dA^{6m} peaks in *X. laevis*. Association between dA^{6m} peaks and biological pathways. Adjusted P-value are show, GO statistical test. Top 12 hits with lowest adjusted P-values are shown for each tissue biological replicate overlap (tissues from different animals).

Online Methods

Genomic DNA isolation. All *X. laevis* and *M. musculus* tissues were directly isolated from sacrificed vertebrates. This was done following all provisions and ethical regulations of the Animal (Scientific Procedures) Act 1986, while having licenses and approval from the Home Office and the Local Ethical Committee (AWERB). No statistical method was used to predetermine sample size. The experiments were not randomized and were not performed with blinding to the conditions of the experiments. After homogenization of the samples (Precellys 24) and addition of phenol chloroform, and the aqueous phase was precipitated with isopropanol and sodium acetate. After 2 washes with 70% EtOH, the DNA was digested with RNAse A for at least 16hrs at 37°C. The DNA was subsequently treated with Proteinase K and purified using the DNeasy Blood & Tissue Kit (Qiagen). To ensure removal of any RNAse, the DNA was again digested with RNAse A and Proteinase K, and repeatedly extracted with the DNeasy Blood & Tissue Kit. The concentration of the genomic DNA was measured using the Qubit double stranded High Sensitivity assay kit.

DNA immunoprecipitation (DIP). DNA immunoprecipitation was prepared using the protocol from Dominissini *et al.*, with the following variations[30]: After the isolation of genomic DNA, at least 20µg DNA was fractionated into 100–200bp fragments using the bioruptor (Diagenode). About 1µg or the fractionated DNA was put aside as an input control. The rest of the fragmented genomic DNA was resuspended in a 1ml final reaction volume containing 10mM Tris-HCl, 150mM NaCl, 0.1% (v/v) Igepal CA-630

and 1.5ug/ul BSA. To this, at least 2.5µg of the dA^{6m} Ab (Synaptic Systems GmbH, m6A antibody, Cat. No. 202003) was added. Its function was validated as described in Figure 1 and Supplementary Figure 1. Alternatively, the following antibodies were used: dA^{6m} Ab* (Synaptic Systems GmbH, m6A antibody, Cat. No. 202011), dA^{6m} Ab** (Synaptic Systems GmbH, m6A antibody, Cat. No. 202111), IgG (Abcam, Cat. No. ab171870). Validation of these antibodies is provided on the manufacturer's website, and has been supplemented by our findings in Figure 4. Please note, species specific validation of these antibodies is not required as the antibodies were only exposed to DNA, and their target is dA^{6m}, which is identical between different species. After an overnight incubation at 4°C on a rotor, 100µl of prewashed protein A magnetic beads were added for 2 hrs at 4°C. Next, the supernatant was removed while Ab bound to the beads was retained using a magnet. After 5 washes with the washing buffer (10mM Tris-HCl, 150mM NaCl, 0.1% (v/v) Igepal CA-630), the DNA bound Ab fraction was eluted. For almost all applications, the DNA was eluted from the Ab and beads with 150µl of elution buffer that contained 10mM Tris-HCl, 150mM NaCl, 0.1% (v/v) Igepal CA-630, 6.7mM methylated adenosine triphophate). The mixture was then incubated for 1hr with continuous shaking at 37°C. The supernatant was removed, and then another 150µl of elution buffer was added to remove any unbound remaining fraction. Next, the supernatants were combined, precipitated, and used for subsequent analysis.

Dot blot. The desired amount of genomic DNA, in most cases 25ng per sample, was diluted in 100µl of 0.5mM EDTA (pH8.0), 7.4% formaldehyde and 6xSSC. Next, the samples were incubated for 30min at 60°C, and then kept on ice. A Nylon + membrane

was soaked in distilled water, and then in 10xSSC. The membrane was transferred into a pre-cleaned dot blot filtration apparatus, and was placed on top of 3 Whatman Paper sheets. While the apparatus was under vacuum pressure, first, the membrane was rehydrated with 10xSSC, and then the DNA containing samples were applied into individual wells. After the samples were fully absorbed in the membrane, the wells were washed with 10xSSC. The apparatus was dismantled, and the membrane was then removed while the apparatus was still under vacuum. After drying in air for about 10min, the membrane was crosslinked at 302nm with UV and blocked for 1hr in 5% nonfat dry milk and 0.1% PBST (0.1% Tween-20 in 1xPBS, pH7.4). Subsequently, the antibody dA^{6m} (Synaptic Systems GmbH) was diluted to 1:1000 in 0.1% PBST and incubated overnight at 4°C. Following 3 washes with 0.1% PBST, a fluorescent secondary antibody was applied for 30min at room temperature (Alexa Fluor Ab, Invitrogen). After further 3 washes with 0.1% PBST, the fluorescent signal was visualized and quantified. All samples that were processed by dot blots were done in triplicate (technical replicates), as well as biological replicates. Original images of blots used presented in the main figures can be found in Supplementary Data Set 1.

Dot blot competition experiments. The genomic DNA was applied and cross-linked to the membrane as described above. However, the dA^{6m} Ab, used at the same dilution of 1:1000, was pre-incubated with different competitors and different competitor concentrations for 3hrs before being applied onto the sample dotted onto the membrane. Dot blots have been cut in order to expose the individual samples to different competitor

concentrations, but all samples were processed in parallel. The subsequent steps are the same as for the normal dot blot procedure described above.

Image analysis. Antibody stained blots were visualized with the LI-COR Odyssey CLx. The images were acquired and quantified with the Image Studio Ver 4.0 software.

DNA oligos. Synthetic oligos were used on dot blots and pulldown studies. The sequence of the 25bp DNA oligo with no dA^{6m} is 5' AGTCGTTCATCTAGTTGCGGTGTAC 3'. The sequence of the 25bp DNA oligo with dA^{6m} is 5' AGTCGTTCATCT(dA^{6m}) GTTGCGGTGTAC 3'. The sequence of the 110bp DNA oligo with no dA^{6m} is 5' TGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCT GGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGG CGATGCC 3'. The sequence of the 110bp DNA oligo with dA^{6m} is 5' TGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCG (dA^{6m}) GCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGA GGGCGATGCC 3'.

Strain genotypes. The Dam– bacteria are a K12 strain, and in addition to lacking the Dam methylase, they are also deficient in deoxycytosine methylation (Dcm–). These strains were obtained from NEB (dam–/dcm– Competent *E.coli*), and have been authenticated by UHPLC-MS/MS and dot blots (Figure 1, Supplementary Figure 2). The

genotype is provided on the manufacturer's website. The Dam+ bacteria are a DH10B strain (Invitrogen). The genotype is provided on the manufacturer's website, and the strain has been authenticated by UHPLC-MS/MS and dot blots (Figure 1, Supplementary Figure 2). *X. laevis* fat and oviduct samples were obtained from female adults. *X. laevis* testes were isolated from adult males. All *X. laevis* were pigmented, and purchased from eNASCO. *M. musculus* kidneys were obtained directly from adult wild type males, with the strain C57B6. All mouse cell lines used in the dot blot experiments come originally from the C57B6 strain, have been identified and tested for mycoplasma contamination by Q-PCR. The human cell line 293T has been obtained from ATCC (ATCC CRL-3216), has been identified and tested for mycoplasma contamination by Q-PCR.

Sample preparation for UHPLC-MS/MS analysis. AGenomic DNA to be analyzed by UHPLC-MS/MS was diluted in a volume of 250µl water. Such samples were then denatured by heating them at 100°C for 5min and immediately placing them on ice. 20µl of 20mM $ZnSO_4$ and 10µl of the nuclease P1 (200units/ml in 30mM sodium acetate, pH5.3) were added, in order to digest any DNA strands into individual nucleotides. After an overnight incubation at 50°C, 180µl of water, and 1µl of bacterial alkaline phosphatase (BAP, 150U/ul) were added. After a 24hr incubation at 37°C, 1ul of the BAP was added again, and the samples were incubated at 65°C for another hour. Next, 30µl of 0.5M Tris-HCl (pH7.9) was added, and the phosphatase reaction was continued for another hour at 37°C. Next, 400µl of water was added, together with silicic acid that filled the 1.7ml Eppendorf tube to about 200µl. After incubation for 15min with occasional vortexing, the sample mixture was transferred onto 0.45µm cellulose acetate filters. After

centrifugation the silicic acid granules were removed. The flow through was then analyzed by UHPLC-MS/MS.

LC-MS/MS analysis. Analysis of global levels of dA and dA^{6m} was performed on a Q Exactive Orbitrap mass spectrometer coupled to a Dionex UltiMate 3000 Rapid Separation LC fitted with an Acquity UHPLC HSS T3 column (100 x 2.1 mm, 1.8 µm particle size). The mobile phase consisted of 0.1% aqueous formic acid (solvent A) and 0.1% formic acid in acetonitrile (solvent B) at a flow rate of 300 µl/min. Calibration curves were generated using serial dilutions of synthetic standards for deoxyadenosine (dA, Sigma) and N6-methyl-2'-deoxyadenosine (dA^{6m}, Sigma). The mass spectrometer was set in a positive ion mode and operated in parallel reaction monitoring. Ions of masses 252.11 (dA) and 266.12 (dA^{6m}) were fragmented and full scans were acquired for the base fragments 136.0618 and 150.0774 ± 5ppm (adenine and methyladenine, respectively). The EIC of the base fragment was used for quantification. Accurate mass of the corresponding base-fragment was extracted using the XCalibur Qual Browser and XCalibur Quan Browser software (Thermo Scientific), and used for quantification. Quantification was performed by comparison with the corresponding standard curve obtained from the pure nucleoside standards running with the same batch of samples. The level of dA^{6m} present in the sample was expressed as a percentage of total adenosine content (methylated and non-methylated), calculated according to the following equation: $(\%)\ dA^{6m} = 100 \times dA^{6m} / [dA+dA^{6m}]$. Differences in dA^{6m} percent abundance were considered significant when $P \leq 0.05$.

Illumina sequencing library preparation. High throughput sequencing libraries were prepared with different genomic DNA samples following the protocol described by Ford *et al.*[29]. All libraries were sequenced using Illumina HiSeq 2000 / 2500, single end, 50bp. At least two biological replicates (tissues from different animals) were performed for each experiment, except for bacterial control samples, where only one replicate was carried out. Each experiment consisted of one dA^{6m} DIP and its corresponding input and occasionally also IgG sample (Supplementary Table 1).

Bioinformatics analysis. Genome alignment for frog, mouse and bacteria genomes: Fastq files were filtered for low quality reads (<Q20) and low quality bases were trimmed from the ends of the reads (<Q20). Bwa 0.6.2 was used to align the resulting reads to the appropriate genome[31]. Frog data was aligned to the filtered version of the *X. laevis* 7.1 Genome[32]. Mouse data was aligned to UCSC mm9 except where specified[33]. Bacteria data was aligned to *E. coli* K-12 strain MG1655 (ref. 33).

Annotation and gene set enrichment: *X. laevis* and *M. musculus* sequences were annotated using InterProScan to provide both InterPro Domains and Panther ontology terms[34,35]. Descriptions for the remaining NCBI sourced sequences were downloaded from the NCBI. Gene set enrichment was obtained using Panther GO Slim terms (7.2) with topGO (Bioconductor package version 2.6.0. http://www.bioconductor.org/packages/release/bioc/html/topGo.html).

Peak calling: PCR duplicates were removed from the aligned datasets and peaks were called using SICER, comparing dA^{6m} pull-down over input or IgG (parameterisation: redundancy threshold = 1, window size = 200, fragment size = 200, effective genome fraction = 0.74 (0.89 *X. laevis*), gap size = 400, FDR = 0.05)[36].

X. laevis genome filtering: Due to the lower quality of this genome assembly, the following filtering steps were performed to increase the accuracy of both mapping and motif analysis. Repeat masker was used to remove any residual repeats (RepeatModeler Open-1.0. 2008-2010 <http://www.repeatmasker.org>). Sequences of low quality (represented as lower case) were masked. Uninformative sequences were removed using DUST and homopolymers of more than 4 bases were removed[37].

Replicate overlaps: The overlaps between replicate peaks were detected using the R bioconductor library "Genomic Features->findOverlaps" (ref. 38). For statistical purposes, one set of peaks in each pairwise comparison was randomly redistributed (shuffled) around the genome. This was repeated 100 times and the mean number of intersected peaks was taken. These shuffled peaks were then compared to the non-shuffled pair resulting in the number of overlaps. This number was then compared to the number of overlaps generated from the original (both pairs un-shuffled) set of peaks using a $\chi 2$-square test. In all cases, the resulting P-value was less than 10^{-16}.

Motif analysis with MEME: Motifs were called using MEME on the sequences below the peaks[26]. For *E.coli* peaks with at least 2 times enrichment were analyzed, while for larger genomes, namely *M. musculus* and *X. laevis,* the top 300 enriched peaks were analyzed by MEME. For *M. musculus* the data was mapped to a filtered version of mm9 where

known uninformative sequences were removed using Repeat Masker (RepeatModeler Open-1.0. 2008-2010 <http://www.repeatmasker.org>) for repeats and DUST for regions of low complexity[37].

Homopolymers of more than 4 bases were also removed from the genome. For *X. laevis* the filtered genome as described above was used. MEME was run on the sequences of the peaks from overlapping replicates or tissues, and on the sequences of the peaks obtained from individual replicates. Statistically significant (E-values < 0.05) 8bp motifs were only found in the overlapping data between replicates and tissues, and only in *X. laevis*.

MEME was run using the ZOOPS model generating different motifs[26]. Motifs were then called using shifted peaks with the same size and distribution to build a background probability of occurrence for the observed motif.

4bp motif analysis: The frequency of all possible four base pair combinations was calculated for sequences of regions under peaks and not under peaks. For all combinations, differences between these two frequencies were then ranked by the respective ratios.

TSS plots: The distance to each TSS from the midpoint of all reads within 20 kb of that TSS was measured. The distances were pooled over all TSS locations and plotted in a histogram with 200 bp bins.

Scatter plots: For all pairwise intersections of peaks, a scatterplot was generated where enrichment values were plotted for each pair of overlapping peaks. Pearson correlations

were also calculated over all enrichment values and t-tests were performed, yielding values less than 10^{-16} in all cases.

Conservation analysis: Conservation scores were downloaded from UCSC phyloP conservation tracks for mouse (mm9)[24]. For each base position under a peak the P score was taken and the overall score was calculated. For each set of peak calls, a conservation score distribution was generated by extracting the phyloP for each position in the peak ranges (n>50 in all cases, so Central Limit Theorem applies). This was compared between peaks shifted by 10kb and the original peaks and the original peaks were shown by a t-test to have a significantly higher level of conservation (P-value < 2.2×10^{-16} in all cases).

Determination of dA^{6m} peak distribution in gene regions: The midpoint of peak locations was classified into regions of the genome including exonic, intronic, intergenic and around the TSS and transcriptional termination sites. The data was normalized to abundance in genome. The peak counts in these regions were modeled by a Poisson distribution assuming, under the null hypothesis, that the incidence rate in each was equal to that of the whole genome average. The probability of the observed counts, given this distribution, was calculated for each region to ascertain whether peak rates were significantly different to the whole genomic background.

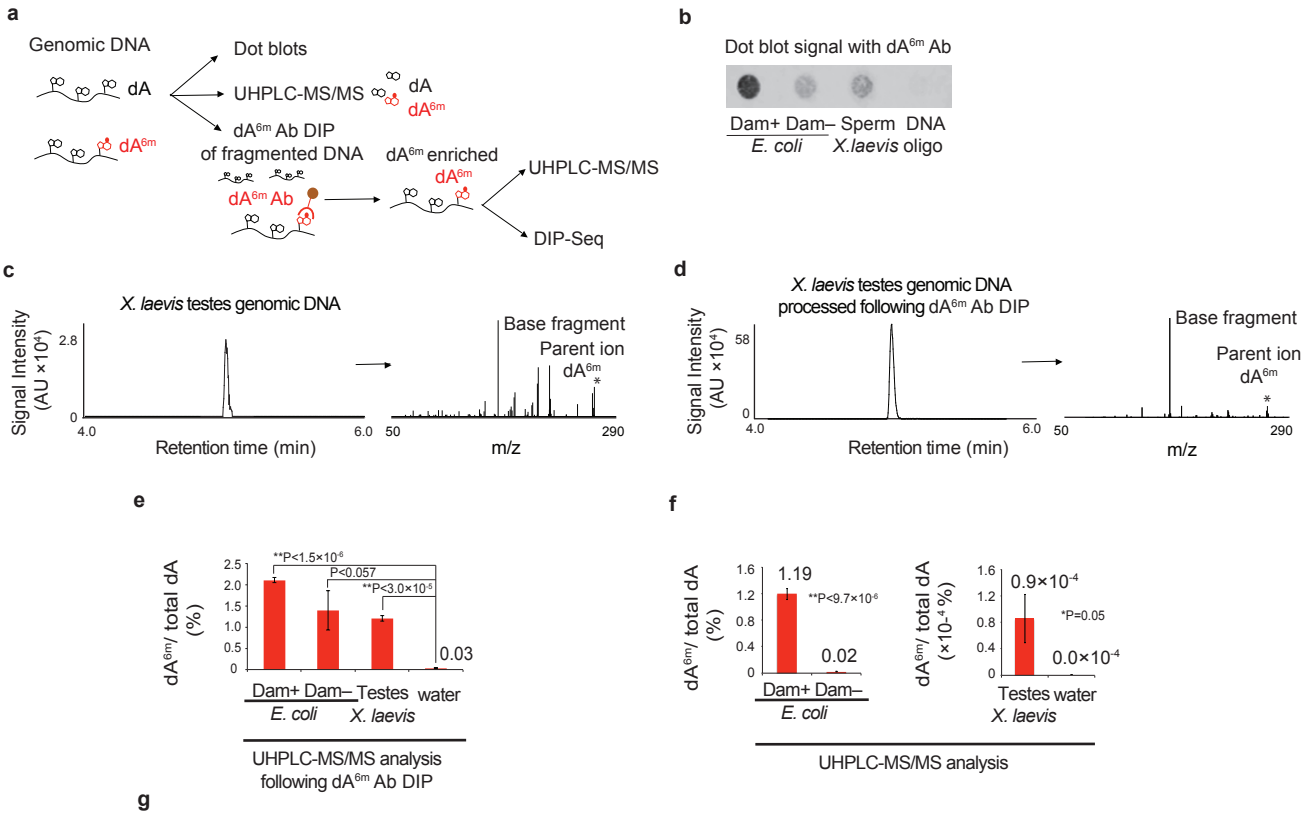
Statistical analysis (of dot blots, mass spec and Ab pulldown yield quantifications). Statistical differences, P-values, were calculated using the two-sided t-test for paired samples. For calculations of P-values between samples in which one sample had only

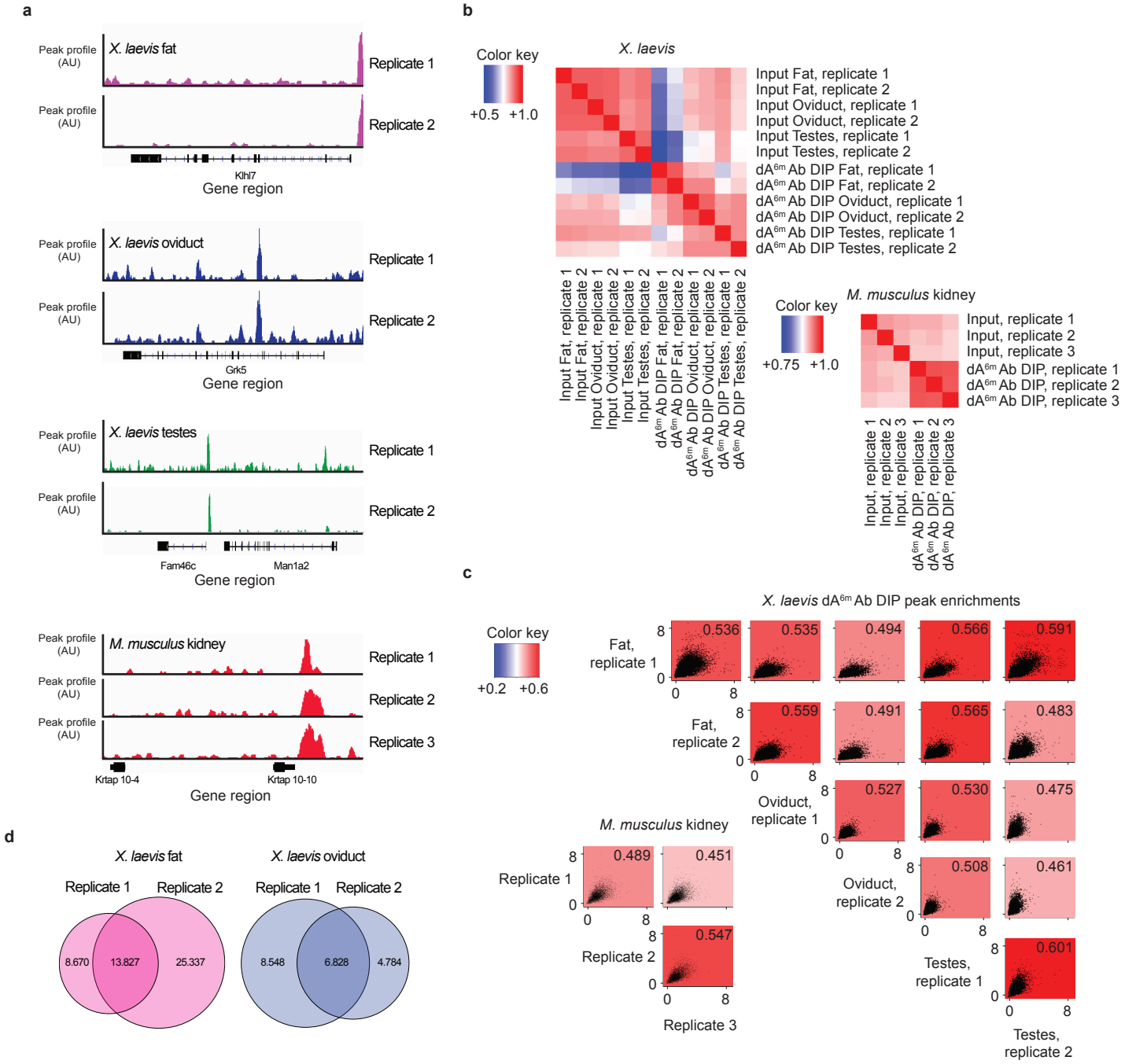
zero as values, the one-sided t-test was applied. All experiments were carried out with 3 technical and biological replicates, indicated by n. All P-values ≤ 0.05 are formatted as *P, while all P-values ≤ 0.005 are indicated as **P.

Venn diagrams. Venn diagrams were drawn with the help of eulerAPE v3 (ref. 39).

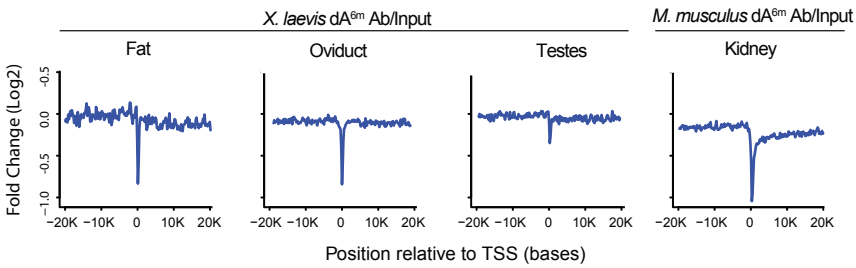
References for Online Methods

30. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., Rechavi, G. Transcriptiome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nature Protocols* **8**, 176–189 (2013)
31. Li, H., Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009)
32. Bowes, J.B., *et al.* Xenbase: a Xenopus biology and genomics resource. *Nucleic Acids Research* **36**, 761–767 (2007)
33. Kent, W.J., *et al.* The human genome browser at UCSC. *Genome Research* **12**, 996–1006 (2002)
34. Zdobnov, E.M., Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001)
35. Thomas, P.D., *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research* **13**, 2129–2141 (2003)
36. Zang, C., *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009)
37. Morgulis, A., *et al.* A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006)
38. Lawrence, M., *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol,* **9**, e1003118 (2013)
39. Micallef, L., Rodgers, P. eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses. *PLoS ONE* **9**, e101717 (2014)

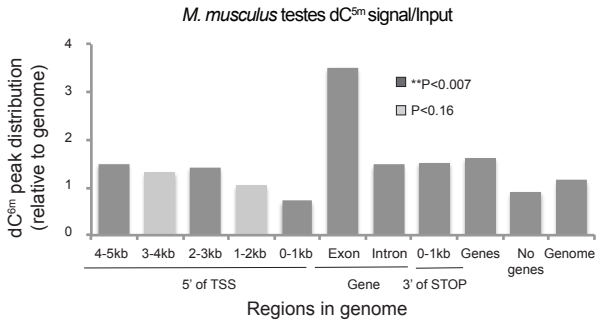

a
Genomic DNA
dA
dA6m
Dot blots
UHPLC-MS/MS
dA
dA6m
dA6m Ab DIP
of fragmented DNA
dA6m Ab
dA6m enriched
dA6m
UHPLC-MS/MS
DIP-Seq
b
Dot blot signal with dA6m Ab
Dam+ Dam– Sperm DNA
E. coli X.laevis oligo
c
X. laevis testes genomic DNA
Signal Intensity (AU ×10⁴)
2.8
0
4.0
6.0
Retention time (min)
Base fragment
Parent ion
dA6m
50
290
m/z
d
X. laevis testes genomic DNA processed following dA6m Ab DIP
Signal Intensity (AU ×10⁴)
58
0
4.0
6.0
Retention time (min)
Base fragment
Parent ion
dA6m
50
290
m/z
e
dA6m/ total dA (%)
**P<1.5×10⁻⁶
P<0.057
**P<3.0×10⁻⁵
2.5
2.0
1.5
1.0
0.5
0
0.03
Dam+ Dam– Testes water
E. coli X. laevis
UHPLC-MS/MS analysis following dA6m Ab DIP
f
dA6m/ total dA (%)
1.6
1.2
0.8
0.4
0
1.19
**P<9.7×10⁻⁶
0.02
Dam+ Dam–
E. coli
dA6m/ total dA (×10⁻⁴ %)
0.9×10⁻⁴
*P=0.05
0.0×10⁻⁴
Testes water
X. laevis
UHPLC-MS/MS analysis
g
Primary tissues
Tissue culture
X. laevis
M. musculus
D. rerio
M. musculus
H. sapiens
Sperm
Fat
Oviduct
Skin
Muscle
Testes
Fat
Testes
Total
ES
MEF
MEF p53-/-
MEF p53-/- Dnmt1N/N
293T
DNA oligo


a
Peak profile (AU)
X. laevis fat
Replicate 1
Replicate 2
Klhl7
Gene region
X. laevis oviduct
Grk5
X. laevis testes
Fam46c
Man1a2
M. musculus kidney
Replicate 3
Krtap 10-4
Krtap 10-10
b
Color key
+0.5 +1.0
X. laevis
Input Fat, replicate 1
Input Fat, replicate 2
Input Oviduct, replicate 1
Input Oviduct, replicate 2
Input Testes, replicate 1
Input Testes, replicate 2
dA6m Ab DIP Fat, replicate 1
dA6m Ab DIP Fat, replicate 2
dA6m Ab DIP Oviduct, replicate 1
dA6m Ab DIP Oviduct, replicate 2
dA6m Ab DIP Testes, replicate 1
dA6m Ab DIP Testes, replicate 2
+0.75 +1.0
M. musculus kidney
Input, replicate 1
Input, replicate 2
Input, replicate 3
dA6m Ab DIP, replicate 1
dA6m Ab DIP, replicate 2
dA6m Ab DIP, replicate 3
c
+0.2 +0.6
X. laevis dA6m Ab DIP peak enrichments
Fat, replicate 1
0.536
0.535
0.494
0.566
0.591
Fat, replicate 2
0.559
0.491
0.565
0.483
Oviduct, replicate 1
0.527
0.530
0.475
M. musculus kidney
0.489
0.451
Replicate 1
Replicate 2
0.547
Replicate 3
Oviduct, replicate 2
0.508
0.461
Testes, replicate 1
0.601
Testes, replicate 2
d
X. laevis fat
X. laevis oviduct
8.670
13.827
25.337
8.548
6.828
4.784
X. laevis testes
3.897
8.094
15.383
X. laevis different tissues
Fat (overalps)
Oviduct (overlaps)
Testes (overlaps)
7.207
1.734
1.051
630
3.413
2.578
1.473
M. musculus kidney
5.320
1.596
1.640
2.726
9.702
2.803
7.130

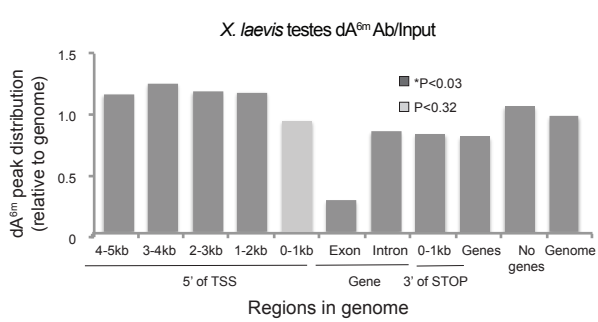
a

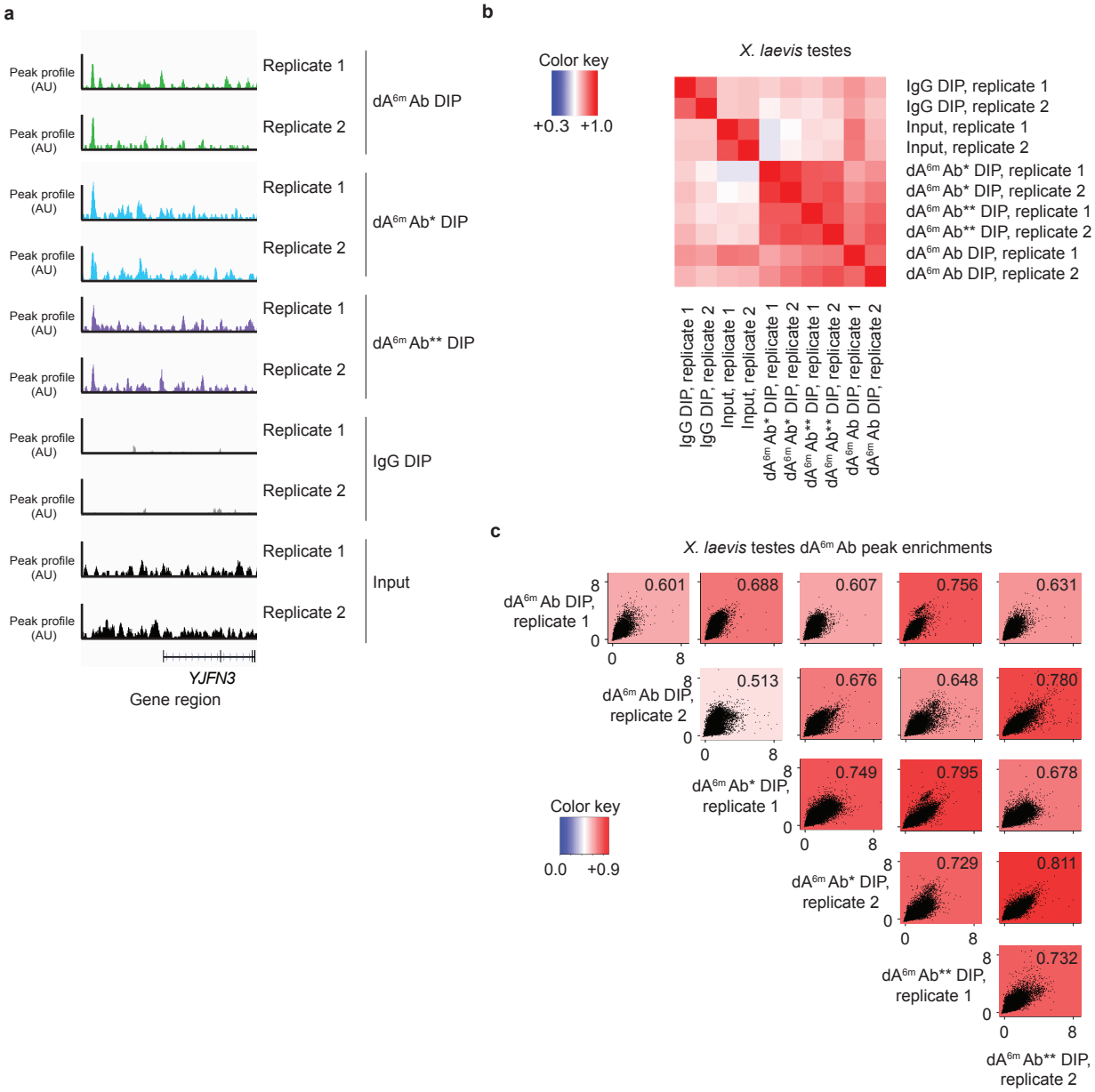

X. laevis dA6m Ab/Input
M. musculus dA6m Ab/Input
Fat
Oviduct
Testes
Kidney
Fold Change (Log2)
-0.5
0.0
-0.5
-1.0
−20K −10K 0 10K 20K
Position relative to TSS (bases)

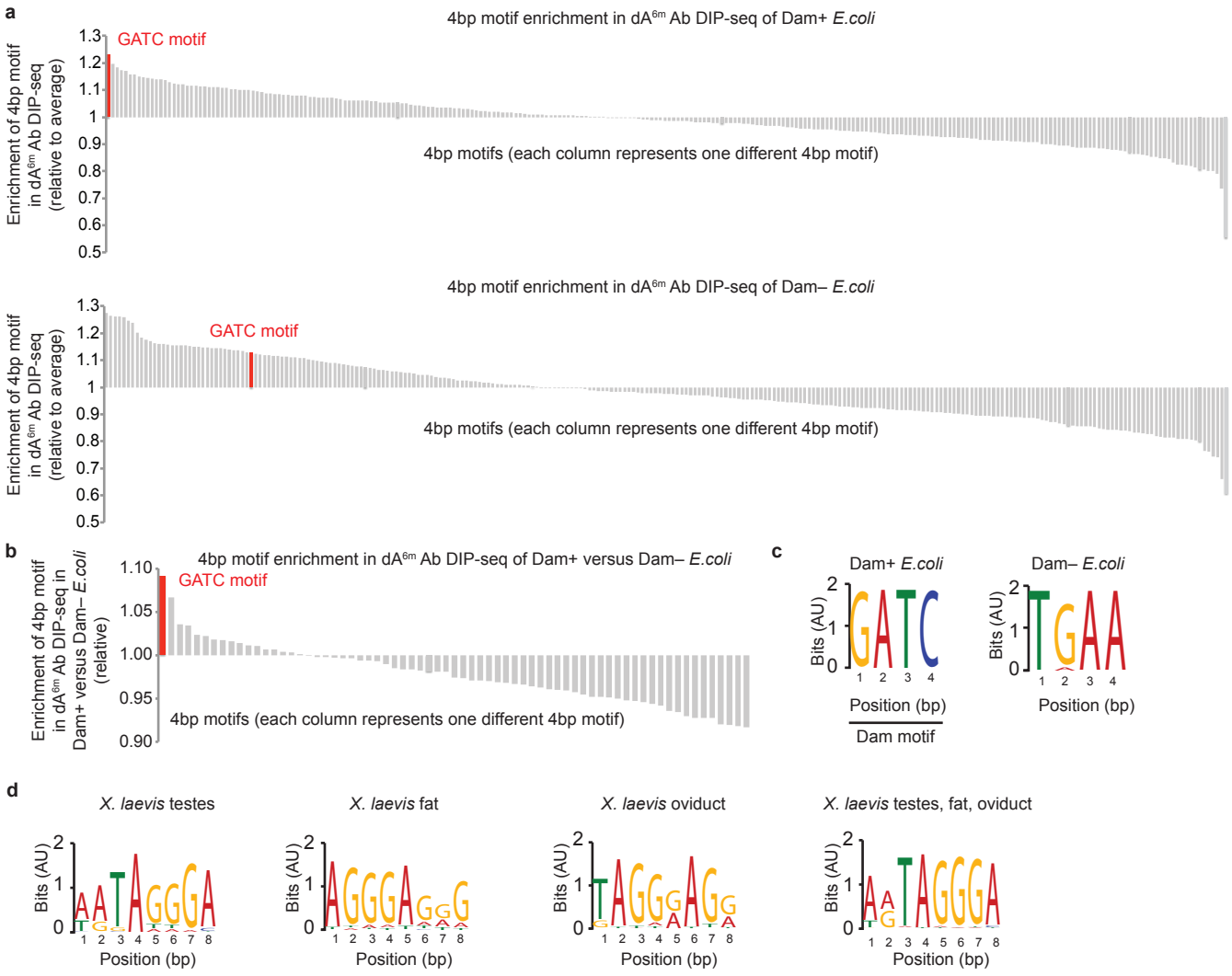
b


M. musculus testes dC5m signal/Input
dC5m peak distribution (relative to genome)
4
3
2
1
0
**P<0.007
P<0.16
4-5kb 3-4kb 2-3kb 1-2kb 0-1kb Exon Intron 0-1kb Genes No genes Genome
5' of TSS
Gene
3' of STOP
Regions in genome

c


X. laevis testes dA6m Ab/Input
dA6m peak distribution (relative to genome)
1.5
1.0
0.5
0
*P<0.03
P<0.32
4-5kb 3-4kb 2-3kb 1-2kb 0-1kb Exon Intron 0-1kb Genes No genes Genome
5' of TSS
Gene
3' of STOP
Regions in genome


a
Peak profile (AU)
Replicate 1
Replicate 2
dA6m Ab DIP
dA6m Ab* DIP
dA6m Ab** DIP
IgG DIP
Input
YJFN3
Gene region
b
X. laevis testes
Color key
+0.3 +1.0
IgG DIP, replicate 1
IgG DIP, replicate 2
Input, replicate 1
Input, replicate 2
dA6m Ab* DIP, replicate 1
dA6m Ab* DIP, replicate 2
dA6m Ab** DIP, replicate 1
dA6m Ab** DIP, replicate 2
dA6m Ab DIP, replicate 1
dA6m Ab DIP, replicate 2
c
X. laevis testes dA6m Ab peak enrichments
Color key
0.0 +0.9
0.601
0.688
0.607
0.756
0.631
0.513
0.676
0.648
0.780
0.749
0.795
0.678
0.729
0.811
0.732


a
4bp motif enrichment in dA6m Ab DIP-seq of Dam+ E.coli
GATC motif
Enrichment of 4bp motif in dA6m Ab DIP-seq (relative to average)
4bp motifs (each column represents one different 4bp motif)
4bp motif enrichment in dA6m Ab DIP-seq of Dam– E.coli
GATC motif
Enrichment of 4bp motif in dA6m Ab DIP-seq (relative to average)
4bp motifs (each column represents one different 4bp motif)
b
4bp motif enrichment in dA6m Ab DIP-seq of Dam+ versus Dam– E.coli
GATC motif
Enrichment of 4bp motif in dA6m Ab DIP-seq in Dam+ versus Dam– E.coli (relative)
4bp motifs (each column represents one different 4bp motif)
c
Dam+ E.coli
Dam– E.coli
Bits (AU)
Position (bp)
Dam motif
d
X. laevis testes
X. laevis fat
X. laevis oviduct
X. laevis testes, fat, oviduct
Bits (AU)
Position (bp)