ENVIRONMENTAL EPIGENETICS

Environmental Epigenetics, 2023, **9(1)**, 1–13
DOI: https://doi.org/10.1093/eep/dvad007
Advance Access Publication 30 November 2023
Research Article

Identifying unique exposure-specific transgenerational differentially DNA methylated region epimutations in the genome using hybrid deep learning prediction models

Pegah Mavaie[1], Lawrence Holder[1,‡,*] and Michael Skinner[2,‡,*]

[1]School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, USA, [2]School of Biological Sciences, Washington State University, Pullman, WA 99164-4236, USA
‡Senior Authors and Correspondence.
*Correspondence address. Center for Reproductive Biology, School of Biological Sciences, Washington State University, Pullman, WA 99164-4236, USA.
Tel: +509-335-1524; E-mail: skinner@wsu.edu
School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, USA.
Tel: +509-335-6138; E-mail: holder@wsu.edu

Abstract

Exposure to environmental toxicants can lead to epimutations in the genome and an increase in differential DNA methylated regions (DMRs) that have been linked to increased susceptibility to various diseases. However, the unique effect of particular toxicants on the genome in terms of leading to unique DMRs for the toxicants has been less studied. One hurdle to such studies is the low number of observed DMRs per toxicants. To address this hurdle, a previously validated hybrid deep-learning cross-exposure prediction model is trained per exposure and used to predict exposure-specific DMRs in the genome. Given these predicted exposure-specific DMRs, a set of unique DMRs per exposure can be identified. Analysis of these unique DMRs through visualization, DNA sequence motif matching, and gene association reveals known and unknown links between individual exposures and their unique effects on the genome. The results indicate the potential ability to define exposure-specific epigenetic markers in the genome and the potential relative impact of different exposures. Therefore, a computational approach to predict exposure-specific transgenerational epimutations was developed, which supported the exposure specificity of ancestral toxicant actions and provided epigenome information on the DMR sites predicted.

Keywords: epigenetics; transgenerational; DNA methylation; deep learning; genomics; toxicants; artificial intelligence; prediction

Introduction

Epigenetics studies the alterations to subsequent protein expression and gene expression that do not change the DNA sequence [1]. Epigenetics is defined as “molecular processes and factors around DNA that regulate genome activity, independent of DNA sequence, and are mitotically stable”. Epigenetic changes typically involve the induction, repression, or silencing of gene expression through epigenetic modifications such as DNA methylation, non-coding RNA (ncRNA), chromatin structure, and histone modifications [2].

One of the most studied epigenetic modifications of DNA is DNA methylation, but much remains to be learned about the underlying mechanisms. DNA methylation refers to the addition of a methyl group to the fifth carbon of primarily cytosine at a CpG nucleotide site [3]. This process can modify gene expression without changing the DNA sequence. In addition, studies show that DNA methylation influences the expression of genes and the regulation of protein binding [4]. These alterations in epigenetics develop gene expression patterns that can cause adverse clinical outcomes, such as allergies, obesity, schizophrenia, cancer, or Alzheimer’s disease, to name a few [5, 6].

Although the DNA sequence does not change with environmental effects, the governing methylation dramatically alters in response to the environment [5]. Environmental epigenetics is the main molecular mechanism that helps to promote phenotypic and physiological alterations [7, 8]. Various environmental factors such as nutrition, stress, or exposure to toxicants can alter the epigenome [9]. In addition, environmental factors early in development can permanently change the cellular molecular function, impacting later life diseases or phenotypes [7].

Examples of transgenerational inheritance are well studied in the literature. Many environmental toxicants have been shown to correspond to the transgenerational inheritance of increased disease susceptibility. For example, atrazine is a common herbicide in the USA and can cause the deterioration of multiple organs in animals [10]. Atrazine increases the risk of testis

Received 6 June 2023; revised 4 October 2023; accepted 28 November 2023

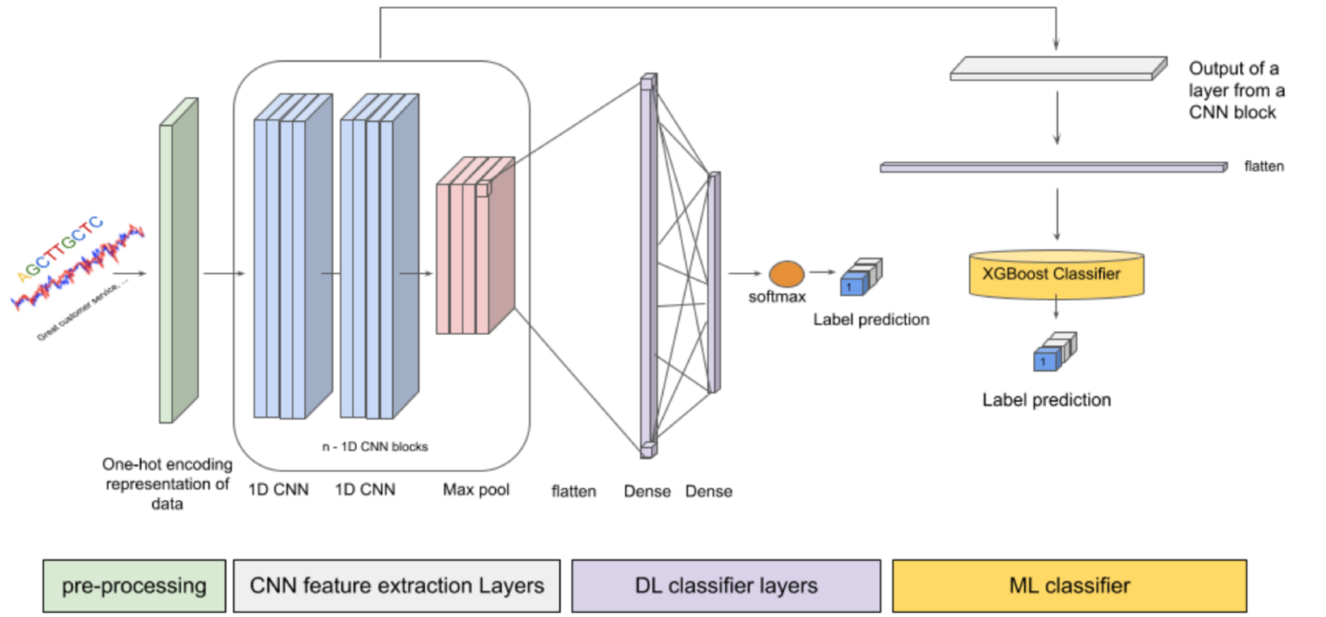


Figure 1: architecture of the hybrid DL-ML model. The model consists of two components: a deep neural network and a traditional ML classifier. The DMR sequence is input using a 5 × 1000 one-hot encoding, which is fed into two Convolutional Neural Network (CNN) blocks, each consisting of two 1D CNNs followed by a max pooling layer. The output of the last block is flattened, then passed to two dense layers, and then passed into a SoftMax layer that makes an internal prediction. After the deep neural network is trained, the output of the first CNN block is used as features to the ML classifier, in this case XGBoost. The XGBoost classifier makes the final prediction as to whether the input sequence is a DMR

disease, kidney disease, prostate disease, and an altered age at puberty [11]. Glyphosate is another commonly used herbicide in the USA that is capable of inducing the transgenerational inheritance of disease and germline (e.g. sperm) epimutations [12]. Pesticides increase the risk of developing neurodegenerative diseases, including Parkinson's disease, Alzheimer's disease, attention deficit hyperactivity disorder, and amyotrophic lateral sclerosis [13–15]. Dichloro-diphenyl-trichloroethane (DDT) is a risk factor for obesity transgenerationally and also induces increased rates of testis, ovary, and kidney pathologies [16, 17]. Various environmental toxicant exposures increase the risk of different diseases. Predicting regions of the genome susceptible to developing into transgenerational epimutations will improve the ability to diagnose and prevent these diseases.

Previous work [18] shows that a hybrid deep machine learning (DL-ML) model can accurately predict a DNA region's likelihood to be differentially methylated (DMR) as a result of ancestral exposure to nine environmental toxicants: atrazine [11], DDT [19], glyphosate [20], vinclozolin [21], pesticides permethrin and N, N-diethyl-meta-toluamide [22], dioxin [23], jet fuel [24], methoxychlor [25], and plastics bisphenol A and phthalates [26]. The hybrid DL-ML model (see Fig. 1) takes advantage of the deep learning network's ability to learn complex features from input DNA sequences, while the ML model overcomes the weakness of the DL model due to fewer training examples by using the DL features as input to a boosted random forest classifier. Using the hybrid DL-ML-based model helps identify DMRs across the whole genome beyond those revealed in the training samples.

However, learning a model to predict DMRs across all exposures can cause over-generalization [18]. One approach to address over-generalization is to determine a core set of predictions, which is the intersection of the predictions made by several trained

Table 1: method for finding the stopping point (SP) for each exposure. SP is computed as the minimal number of random subsets of the predicted DMRs, that when intersected together, result in the empty set. SP represents the number of models that must be training, and their DMR predictions intersected, to arrive at a core set of predicted DMRs that exclude noisy predictions due to variance in the models

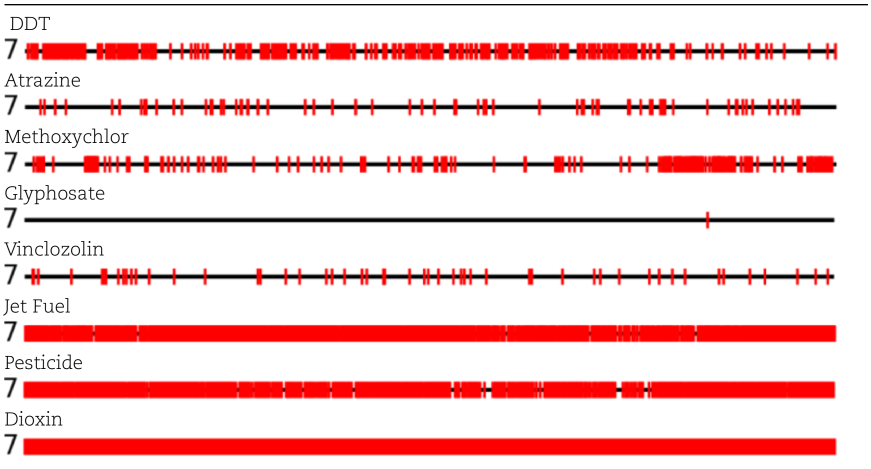
Finding the right number of models for exposure
1. N = # DMRs predicted by one trained hybrid model for exposure
2. R = all regions in genome
3. SP = 0
4. Repeat
a. R' = randomly choose N regions from all regions in genome
b. R = R intersect R'
c. SP = SP + 1
5. Until R is empty
6. Return stopping point.

models, each randomly initialized. The number of trained models necessary to generate the core set is computed as the stopping point (see Table 1). Also, many of the DMRs for the aforementioned nine exposures are unique. Therefore, another approach to address over-generalization is to learn individual models for each exposure. In addition, the mechanism by which epigenetic effects are realized may involve a preponderance of DMRs rather than a specific DMR signature, which would lead to an over-generalized model if focused on finding such an elusive signature. An exposure-specific model specialized to the exposure can identify common and unique predicted DMRs not revealed in the training data. Such a model also helps to recognize the toxicants to which an individual's ancestors were exposed and allows for early preventative treatment to avoid more long-term severe outcomes.

Table 2: the stopping point, the number of training DMRs, the average number of predicted DMRs in one model, the core set of DMRs, and the unique regions in each exposure for the training DMRs and the core set of predicted DMRs, all for chromosome 7. The same 6636 non-DMRs were used for training in each exposure

Exposure	S.P.	Training DMRs	Predicted DMRs	Core set DMRs	Unique training DMRs	Unique core DMRs
DDT	6	1543	14 370	3184	520	525
Atrazine	2	243	697	258	112	74
Methoxychlor	3	423	12 476	4474	222	258
Glyphosate	1	5	4	4	5	1
Vinclozolin	2	220	1375	978	70	58
Jet Fuel	27	1973	78 122	21 899	776	2282
Pesticide	15	1145	55 819	15 259	314	1069
Dioxin	79	2431	90 910	35 634	1264	12 760
Plastics	165	12 504	134 884	n/a	10 295	n/a

Table 3: location of the unique DMRs on chromosome 7 for each exposure



Results

Exposure-specific models were used to identify DMRs unique to each exposure and common across multiple exposures. These DMR sets were analyzed using four techniques: (i) visualize the location of the DMRs, (ii) identify transcription factor (TF) matches in the DMRs, (iii) identify genes associated with the DMRs, and (iv) identify common motifs in the DMRs. The results of this analysis for the whole genome are provided in the supplemental materials. Supplementary Tables S1–S6 show the number of unique DMRs in each chromosome for each exposure. Supplementary Figs S1–S22 visualize the location of the unique DMRs in each chromosome for each exposure. Supplementary Tables S7–S28 list the TF matches in each chromosome for each exposure. Supplementary Tables S29–S50 list the genes associated with the unique DMRs in each chromosome for each exposure. Supplementary Figs S23–S44 show the common motifs found in each chromosome for each exposure. Given the size of the analysis results for the whole genome, only results for chromosome 7 are shown here to demonstrate the analysis in a succinct form. Chromosome 7 was chosen somewhat arbitrarily but demonstrates the types of conclusions that can be drawn from results on other chromosomes.

Table 2 summarizes the data and results for each exposure for chromosome 7. Glyphosate and plastics exposures were not included in subsequent analysis due to their outlier properties. Table 3 shows the location of the unique DMRs in the other seven exposures for chromosome 7.

Motif Alignments for Unique DMRs

After composing the unique DMR set for each exposure, the TOMTOM tool is used to find the known motifs in the unique regions for each exposure [27]. Table 4 shows the matches found in the unique DMRs in each exposure for chromosome 7. Vinclozolin has only one motif alignment with its unique DMRs (L8GDR2_ACACA), and so is not included in the table for brevity. In the case of chromosome 7, each of these motifs had only one match to the unique DMRs. In other chromosomes, there were some cases of more than one match, but these cases were rare. The complete results for all chromosomes are included in Supplementary Tables S7–S28.

The results in Table 4 indicate several motifs that have known associations to the exposure. Atrazine is an herbicide that has been shown to have negative effects on amphibians, such as disrupting their endocrine systems and causing developmental abnormalities, cancer risk, and neurological problems [28]. Bd11a is a gene in amphibians that encodes a TF binding that regulates the genes and has a role in cancer progress [29]. It is possible that exposure to atrazine could affect the expression or activity of Bd11a or its binding to DNA. Another TF match with unique DMRs of atrazine is Mef2c. Mef2c is known to play critical roles in the development and function of multiple organs and tissues, including the heart, skeletal muscle, and brain [30].

With regard to dioxins, some studies have suggested that exposure to it may be associated with an increased risk of certain types of cancers [31], which may involve the dysregulation of genes controlled by TFs like Zpf384. Early growth response 3 is

Table 4: transcription factor matches found in the unique DMRs in each exposure for chromosome 7. The TOMTOM tool is used to find the known motifs in the unique regions for each exposure. Vinclozolin has only one motif alignment with its unique DMRs (L8GDR2_ACACA), and so is not included for brevity

DDT		Atrazine	Methoxychlor		Jet Fuel	Pesticide	Dioxin	
Zfp110	Cic	Zfp523	Klf6	Pou2f2	Prdm6	Bhlhe3	Zpf422	Srebf2
Mecom	Zzz3	Zfp354a	Zfp580	Zbtb37	Lin54	Nfib	Zpf287	Foxp2
Tcf7l2	Rbpj	Bd11a	Zfp641	Zfp90	Zfp189	Nfia	Zpf384	Zfp212
Mef2d	Cdc5l	Zfp513	Klf4	Glis3	Foxi1	Nfx1	Prdm6	RGD1304587
Irf8	Tbx4	Mef2c	Glis2	Rara	E2f7	Lin54	Prdm4	Rest
Gata1	Mga	Stat2	Klf3	Cdx4	Neurod1	Hmg20b	Zbtb26	Egr3
Trps1	Tbx5		Zfp449	Cdx2	rdm1	Mef2d	Pitx1	Zfp3
Gata2	Tbx1		Rreb1	Zfp382	Znf354b	Pou4f2	Zfp189	Nr5a1
Gata4	Tbx6		Sp4	Mynn	Zfp41	Scrt1	Zbtb12	Nr2f2
Gata6	Nfactc3			Dbx1	Gli3	Neurod1	Nr6a1	Nr4a1
Etv2	Ikzf3			Zfp410	Gli1	Yy1	Bcl6	Esrrg
Vdr	Zbb48			Zscan10	Gli2	Gli3	Zfp829	Esr1
Thra	Nr3c1			Zfp770	Rel	Klf1	Zfp513	Nr4a2
Thrb	Esrra			Zfp787	Ar	Klf9	Zfp410	Rarb
Zbtb12	Sox10			Nr2e3	Zfp24	Ebf1	Ctcf	Nr2e1
Smad4	Zfp283			Nhlh1	Zfp143	Zfp128	Zfp1	Rxra
Myrf	Hox6			Nr5a2	Ets1	Myrf	Thrb	Rxrb
Jund	Mxf1			Nr5a1	Tbx2	Sox10	Thra	Rarg
Mzf1	Onecut2			Esr1	Sox14	Zfp524	Zfp281	Ppard
Jun	Foxp2			Esrrg	Sox9	Znf454	Klf5	Nr2f1
Atf7				Rxrb	Sox13	Nhlh1	Zfp467	Nr2e3
				Sox2	Sox6	Ascl1	Klf16	Spi1
						Tbx20	Klf10	Nfkb2
						Zbtb26	Klf11	Gl3
						Dpf3	Klf14	Nfe2
						Rfx5	Klf12	Nwurod1
						Rreb1	Sp3	Bhlha15
						Elf	Znf354b	Stoh1
						Et1	E2f8	Runx1
						Ets2	E2f7	Mycn
						Zscan10	Foxk2	Foxa1

a TF that plays a role in the regulation of gene expression in response to various stimuli, such as growth factors, cytokines, and environmental toxins. Dioxins, which are highly toxic environmental pollutants, have been shown to activate early growth response 3 in some studies [32]. Additional motifs shown in Table 4 may suggest previously unknown effects of the exposures on the genome.

Genes Overlapping Unique DMRs

Table 5 shows the overlapping genes associated with the unique DMRs in each exposure for chromosome 7. DDT, atrazine, and vinclozolin do not have any overlapping genes. Only a sample of the genes overlapping dioxin is shown in this table for brevity. A complete list of all overlapping genes for all chromosomes is included in Supplementary Tables S29–S50.

Previous studies show that there are several connections among the associated overlapping genes and the exposures. As an example, anti-Müllerian hormone is an important regulator of folliculogenesis in the ovary and can be dysregulated by dioxin [33].

Table 5: overlapping genes associated with the unique DMRs in each exposure for chromosome 7. Rat gene locations were obtained from the Rat Genome Database (https://rgd.mcw.edu) and aligned with the predicted unique DMRs. None were found in the unique DMRs for DDT, atrazine, and vinclozolin

Dioxin		Jet fuel	Methoxychlor	Pesticide	
Atxn10	Acvrl1	Cand1	Gzmm	Phlda1	Best3
Baz2a	Adcy6	Dbx2	Npff	Pphln1	Cnn2
Bik	Adm2	Gtsf2	Sppl2b	R3hdm2	Cyp2b1
Bin2	Akap8l	Ilvbl	Spryd3	Rapgef3	Endou
Btg1	Amh	Mtss1	Tssk5	Tac3	Fzr1
Card10	Apof	Olr1045-ps	Zfp707	Tmem117	Kif21a
Ccn4	Apol11a	Olr1073		Tspan31	Map2k2
Cct2	Apol3	Ptprq		Zc3h10	Pdxp
Cdc34	Apol9a	RGD1560979		Zfp7	Pglyrp2
Cdc42ep1	Arc	RGD1561871		Znf7	
Cdk17	Arfgap3	RGD1565356			
Cdpf1	Arhgap45	Scyl2			
Celf5	Arhgap8	Tafa2			
Celsr1	Arhgef25	Them6			
Cenpm	Arsa	Tmem65			
Cfap54	Asap1	Tph2			
Chadl	Asic1				
Cradd	Cry1				

Most Repeated Motifs in the Unique DMRs

Figures 2–8 show the top five most repeated motifs in the exposure-specific DMRs for chromosome 7. Results for the whole genome are included in the Supplementary Figs S23–S44. The focus here is on motifs that are unique to one exposure. While DMRs require the presence of CpGs, the motifs discovered here are less likely to contain CpGs, since they are not unique to a particular exposure. The 1 kb DMRs may contain motifs that do not overlap with the CpGs within the DMR.

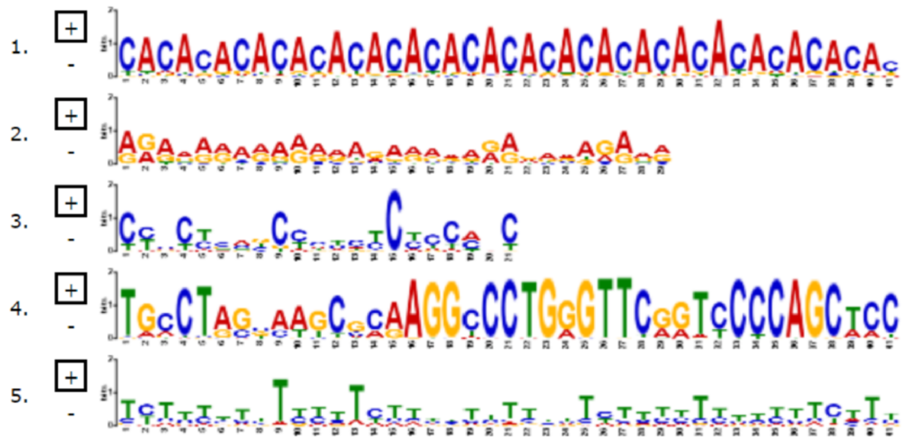


Figure 2: top-five most repeated motifs in the unique DMRs for DDT in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

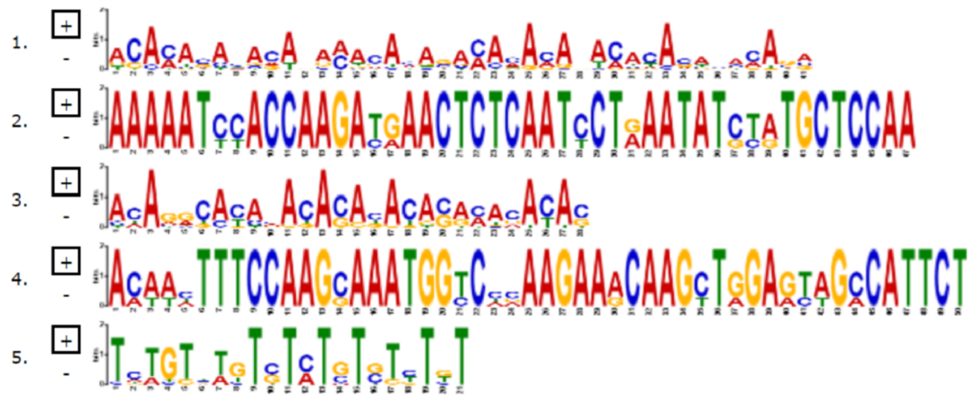


Figure 3: top-five most repeated motifs in the unique DMRs for vinclozolin in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

Figure 4: top-five most repeated motifs in the unique DMRs for pesticide in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

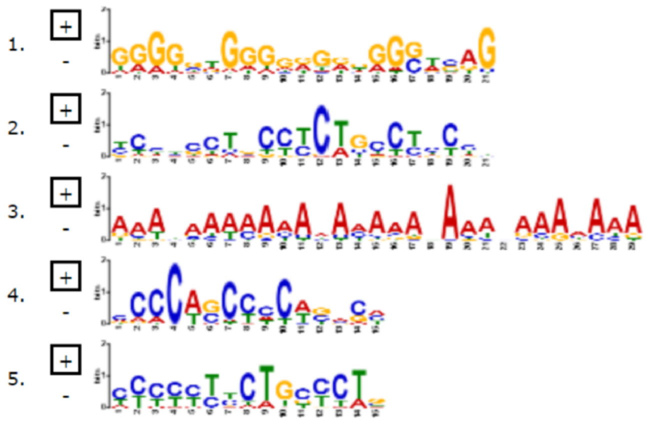


Figure 5: top-five most repeated motifs in the unique DMRs for methoxychlor in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

The motif visualizations indicate some patterns specific to certain exposures. For example, the common motifs in the DMRs unique to DDT (Fig. 2) show a predominance of the smaller ACA motif, which is associated with DNA-binding in the malaria parasite targeted by the pesticide DDT [34]. The common motifs in the DMRs unique to jet fuel (Fig. 6) show a predominance of the smaller GTG motif, which is associated with increased DNA-binding of TCF4 [35], and jet fuel (naphthalene) has been observed to inhibit the TCF4 binding [36]. The common motifs in the DMRs unique to atrazine (Fig. 7) show a predominance of the smaller TCT motif, which is associated with transcription of protein gene promoters [37], and atrazine has been observed to impact the transcription and regulatory processes [38].

Common DMRs Across all the Exposures

The above analyses were performed on the common DMRs across all exposures. Table 6 shows the number of DMRs common to at least N exposures. The analysis focused on the DMRs that were common among at least five (N = 5) exposures.

Table 7 shows the locations in the whole genome of the DMRs common to at least five exposures. Not surprisingly, the DMRs

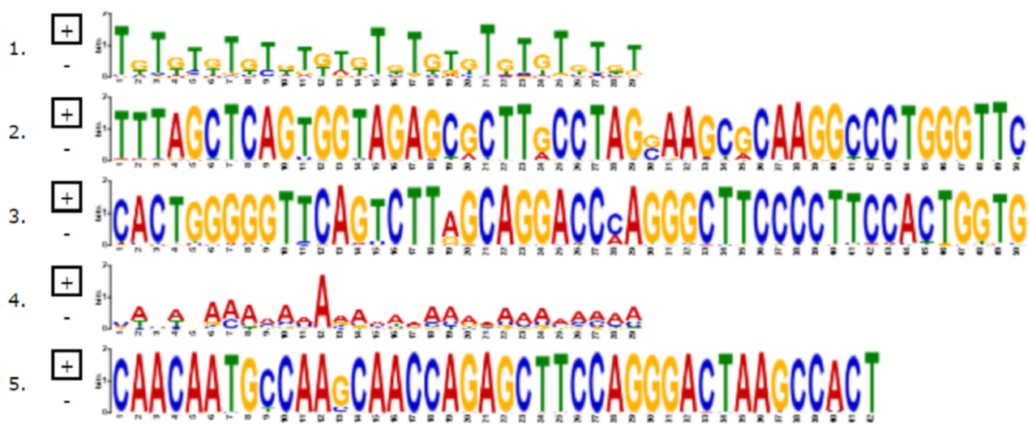


Figure 6: top-five most repeated motifs in the unique DMRs for jet fuel in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

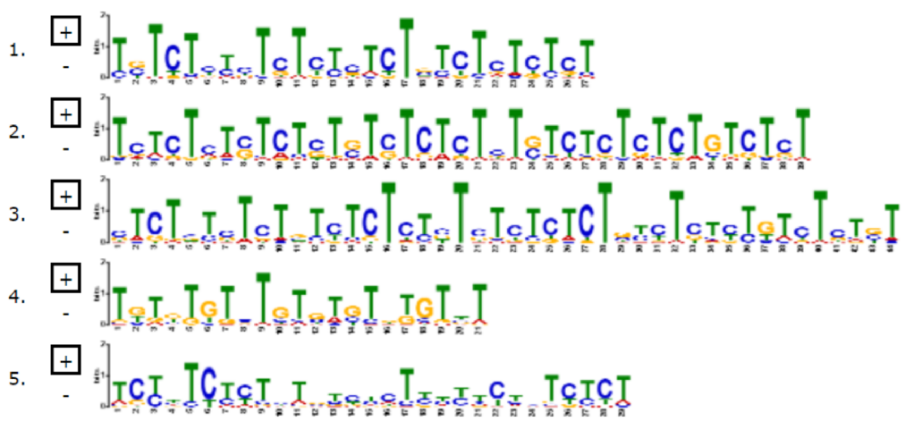


Figure 7: top-five most repeated motifs in the unique DMRs for atrazine in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

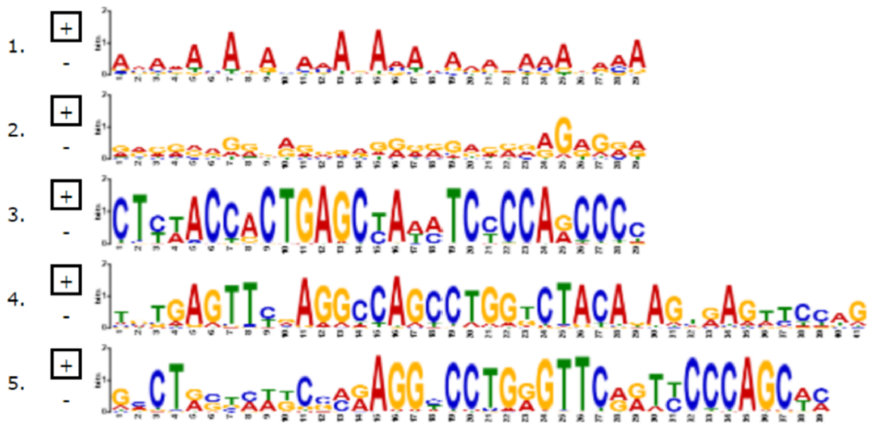


Figure 8: top-five most repeated motifs in the unique DMRs for dioxin in chromosome 7. The motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

Table 6: the number of core DMRs common to at least N different exposures for each chromosome. A total of seven exposures are used for this analysis; glyphosate and plastics are excluded. None of the core DMRs are common to all seven (N = 7) exposures on any chromosome. Some core DMRs are present in six (N = 6) different exposures, and as expected the number of common core DMRs increases as the constraint on the number of common exposures declines. Note that the N exposures that each core DMR has in common do not need to be the same N exposures, but any N of the seven exposures

	# Core DMRs common to N exposures						
Chr	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7
1	37 504	26 982	18 883	4989	372	11	0
2	29 388	22 191	15 659	4315	285	4	0
3	24 534	17 511	12 018	2717	176	5	0
4	23 642	16 616	11 175	2897	226	10	0
5	22 512	17 591	12 315	3385	199	6	0
6	19 812	13 654	9008	2450	143	7	0
7	19 487	15 410	10 921	2665	132	5	0
8	20 241	14 038	9056	1850	109	7	0
9	16 106	12 457	8230	1931	162	8	0
10	18 806	13 521	9135	1639	113	2	0
11	10 973	8608	5750	1440	115	3	0
12	10 717	8119	6859	990	55	6	0
13	13 947	11 022	7718	2018	155	6	0
14	13 883	11 739	8242	1846	165	5	0
15	13 445	11 132	7314	1678	112	7	0
16	11 733	9798	6532	1521	91	4	0
17	12 355	10 514	7774	1467	99	4	0
18	11 576	8942	5940	1467	152	1	0
19	10 121	7583	5300	1106	70	4	0
20	8177	7602	6511	1183	70	1	0
X	12 470	8625	6772	3229	241	2	0
Y	314	254	292	194	9	0	0
Total	361 743	273 909	191 404	46 977	3251	108	0

are uniformly distributed within chromosomes and across the whole genome. However, higher concentrations as well as significant gaps can be observed in some chromosomes. Table 8 shows the known motifs found in the common DMRs in each chromosome, and Table 9 shows the overlapping genes associated with the common DMRs. Figures 9–12 show the top three most repeated motifs among common DMRs for each chromosome. These results indicate potential common mechanisms by which most toxicants affect the genome. Observations can be contrasted to those in previous work [18] that identify motifs in the features extracted from the DL network. Feature motifs do not necessarily represent common patterns in DMRs, but can also represent patterns in non-DMRs that are useful to discriminate them from DMRs.

Discussion

A hybrid DL-ML approach that has previously shown success at predicting DMRs [18] was used to identify core sets of DMRs per exposure and then unique DMRs within these core sets. Analysis shows that there are unique DMRs associated with each exposure, and the exposure-specific models are a better solution to identify these unique DMRs.

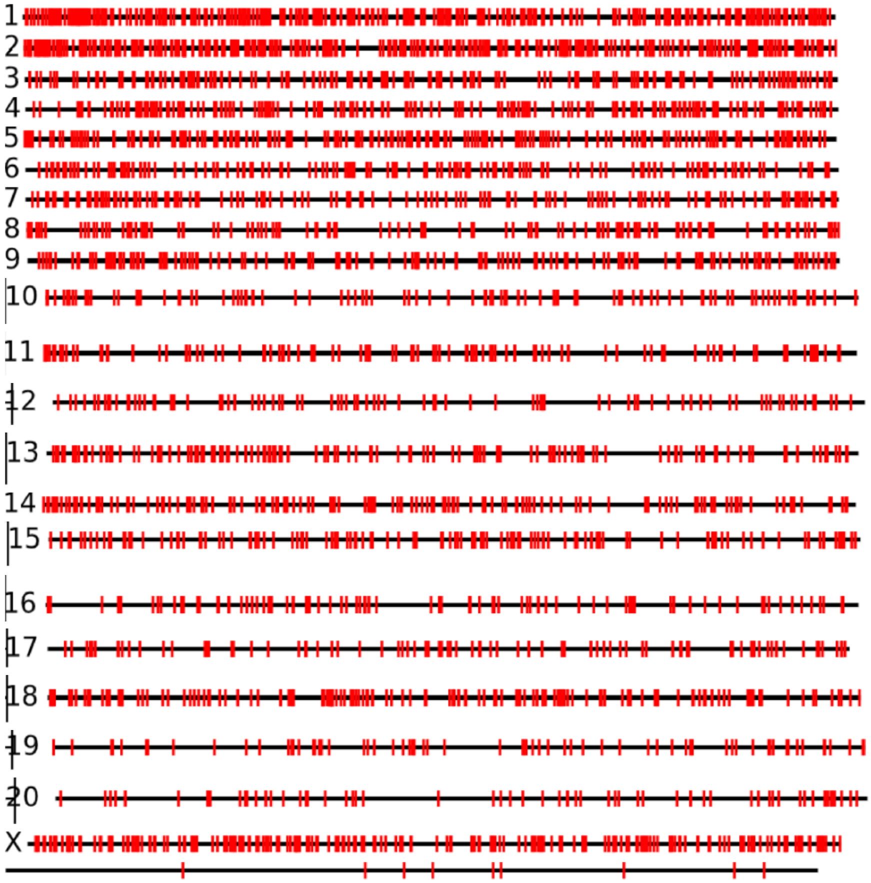
Results in previous work show that the hybrid model has high accuracy on the data constructed from nine different exposures [18]. However, training only one model on DMRs from all nine exposures results in high variance and large numbers of predicted DMRs The actual number of DMRs is likely fewer than the number predicted. This is addressed by intersecting the predictions of several models to identify a core set of DMRs that are predicted by every model.

This paper focuses primarily on analyzing the unique DMRs in each exposure. The unique DMR prediction in the whole genome is used to find biologically relevant features through visualization of DMR locations, motif analysis, and gene associations. This can indicate the unique effects of each toxicant on the formation of different DMRs. Analysis of the common DMRs across most exposures is also presented. The presence of predicted exposure-specific DMRs suggests such DMRs could be used to assess exposures within individuals and populations. The presence of such transgenerational exposure specific biomarkers may allow in the future the ability to determine ancestral exposure and how that may impact an individual's health in the future. Further research on exposure epigenome predictions could be used as a diagnostic tool in the areas of toxicology and medicine.

The DL-ML approach represents a new direction in the analysis of genomic data. The presence of genomic phenomena is often based on a quantitative analysis of laboratory results, e.g. in the case of this study, a DNA region is labeled as a DMR based on a threshold on the experimentally determined probability that the region is differentially methylated. The choice of this threshold

Table 7: the locations of the common DMRs (common to N = 5 exposures) on each chromosome in the whole genome

Chr	Visualization
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
X	
Y	



can significantly vary the number of regions labeled as DMR. Using machine learning, a set of high-confidence DMRs can be used for training the ML models, which can then make predictions about DMRs elsewhere in the genome. More analysis is needed to confirm that the ML-based predictions are more accurate, but if so, this approach reduces the need to precisely tune the confidence threshold, allows a more nuanced selection of DMRs rather than using a single threshold, and can identify DMRs that would not meet even minimally restrictive thresholds due to inconsistencies in the experimental process. While other ML approaches may be used for this purpose, the hybrid DL-ML approach is uniquely suited for two reasons. First, using the DL network to learn and extract features relieves the analyst from the burden of handcrafting features for ML. Second, using a non-DL classifier for the final DMR prediction avoids the typical need for large datasets when using a DL classifier alone. Thus, the hybrid DL-ML approach is uniquely positioned to succeed at this new approach to ML-based analysis of genomic data.

The approach described in this paper is focused on predicting exposure specific DMRs vs all non-DMRs in each model. However, one possible future direction is to view the problem as a one-vs-rest learning task by revising the definition of the negative samples. The models can still be trained with DMRs in each exposure as the positive samples, but with the DMRs in other exposures as the negative samples. In this case, the models would predict unique exposure specific DMRs directly. Another future direction is to apply a similar approach to the analysis of disease-specific DMRs. Models can be trained on DMRs associated with each disease vs non-DMRs or the DMRs from other diseases. Similar to the current approach, a core set of predicted DMRs can be identified for each disease, and then the DMRs unique to each disease and common to all diseased can be isolated and analyzed. Several observations suggest the environment has a significant impact on disease etiology [9]. Identifying exposure-specific and disease-specific DMRs can lead to a diagnostic tool for predicting susceptibility to certain diseases based on epigenetic mutations from ancestral exposures. However, more data are needed from human studies and from alternative analysis methods to validate the clinical viability of the approach. Future studies are needed to incorporate the use of computational approaches such as the hybrid deep learning to help facilitate future use of epimutations as biomarkers for exposure and disease. The procedure can be used on a variety of datasets, and so is not specific to DNA methylation or the analysis used. Observations demonstrate the hybrid deep learning approach can be used as a prediction tool for further epigenome studies.

Methods

The goal is to first identify a DNA region's susceptibility to develop an environmentally induced transgenerational alteration (i.e. a DMR) for each individual exposure based on a DL-ML model's

Table 8: transcription factor matches found in the common DMRs (common to N = 5 exposures) on each chromosome in the whole genome. The TOMTOM tool is used to find the known motifs in the common DMRs

Chr	Overlapping genes	Chr	Overlapping genes
1	Tbx20	12	Zbtb26
	Zfp287		Foxa3
	Sox10		Zfp287
	Hnf4a		Zfp182
2	Zfp105		Foxp2
	Zfp287	13	Prdm6
	Zfp879		Tbx20
	Prdm6	14	Zfp105
	Zfp105	15	Zfp105
3	Zbtb26		Foxg1
	Foxr1		FOXQ1_RAT
	Zfp287		Foxa3
	Sox10		Foxl2
4	Zfp105		Nr1d1
	Tbx20		Nr1d2
	Zfp105	16	Foxp2
5	Zbtb26		Msantd3
	Foxg1		Tbx20
	Foxl2		Zfp105
6	Sox10	17	Zbtb26
	Nr1d2		Prdm6
7	Zbtb26		Zfp105
	Sp3	18	Zbtb26
	Tbx20		
	Zfp287		Tbx20
	Klf9	20	Zfp24
	Klf4		Hdx
	Prdm6		Zfp287
	Zbtb26		Zfp182
8	Zfp28	X	Zfp422
9	Tbx20	Y	Zfp449
	Foxf1		
	Tbx20		
	Nr1d2		
10	Zfp105		
11	Rreb1		

Table 9: overlapping genes associated with the common DMRs (common to N = 5 exposures) on each chromosome in the whole genome. Rat gene locations were obtained from the Rat Genome Database (https://rgd.mcw.edu) and aligned with the common DMRs No known genes overlapped the common DMRs in chromosomes 4, 6, X, and Y

Chr	Overlapping genes	Chr	Overlapping genes
	Ascl3		Pcnp
	Ganab	11	Tra2b
	Irx1		Ache
	L3mbtl3		Stag3
	Syvn1		Vom2r-ps91
1	Trnas-gcu3	12	Vps37b
	Anp32e		Glrx2
	Bhlhe22	13	Nsl1
	Cct3	14	Noa1
	Khdc4		Kctd9
	Lysmd1	15	Mrpl57
	Plrg1		Ing1
	Ppid		Jund
2	Trnar-ucu3		Klf2
	Naif1		Mak16
	Snap23		Mpv17l2
3	Zfp341		Ncoa4
4	–	16	Sap30
	Trnas-aga1		Gmnn
5	Orc1	17	Msrb2
6	–		Chmp1b
	Dusp6		Mtmr1
	Hoxc12		Pcdhgb7
	Polr2e		Prdm6
7	Tnrc6b	18	Rps14
	Chrna5		Dhx38
	Fez1		Dus2
	Npat		Hook2
	Plekho2		Nip7
8	Trnar-acg2	19	Slc9a5
	Dazl	20	Pou5f1
9	Klhdc3	X	-
	Trnal-uag2	Y	-
10	Trnar-ucu4		

prediction. Then, the unique DMRs for each exposure can be identified and their existence suggests unique effects of individual exposures and potentially a means to detect ancestral exposure to the toxicants.

The overall method consists of several steps for each exposure dataset: (i) define positive and negative samples for the training process; (ii) train a hybrid DL-ML model to predict exposure-specific DMRs in the whole genome; (iii) find the proper number of models to address model variance and indicate how many models are required to identify a core set of predicted DMRs; (iv) train this number of hybrid DL-ML models and use these models to predict DMRs across the whole genome; (v) identify the core set of predicted DMRs, i.e. the DMRs predicted by all models; (vi) extract the unique DMRs in the core sets for each exposure; and (vii) search for known motifs, genes, and TFs associated with these unique DMRs.

The Skinner laboratory at Washington State University has produced several datasets based on the rat genome that identify the DMRs in the F3 generation after exposure of the F0 generation to one of nine toxicants: atrazine [11], DDT [19], glyphosate [20], vinclozolin [21], pesticides permethrin and N, N-diethyl-meta-toluamide [22], dioxin [23], jet fuel [24], methoxychlor [25], and plastics bisphenol A and phthalates [26]. Vinclozolin is used as both an agricultural fungicide and pesticide. Dioxin is a highly-toxic byproduct of the manufacture of chlorinated compounds, such as some herbicides, but also occurs naturally. Atrazine and glyphosate are commonly used herbicides. DDT is an insecticide that was used extensively in the 1950s and 1960s to combat insect-borne diseases such as malaria but has since been banned in the USA due to adverse health and environmental effects. Methoxychlor is an insecticide that was intended as a replacement for DDT, but was also banned in 2003 due to adverse health effects. Jet fuel (JP-8) is a hydrocarbon mixture used commonly by the military but has been found to be potentially toxic to the immune system, respiratory tract, and nervous system [39].

In these studies, the F0 generation consisted of gestating female rats divided into 'control' (no exposure) and 'exposure' (exposed to the toxicant) groups. The offspring of the F0 generation comprised the F1 generation. Males and females in the control or exposure groups of the F1 generation were bred to obtain the F2 generation. Then, the F2 generation rats were bred to obtain the F3 generation. The initial direct exposure of the gestating female F0 generation rats also exposes the developing F1 generation fetus and the germ cells within the F1 generation, resulting

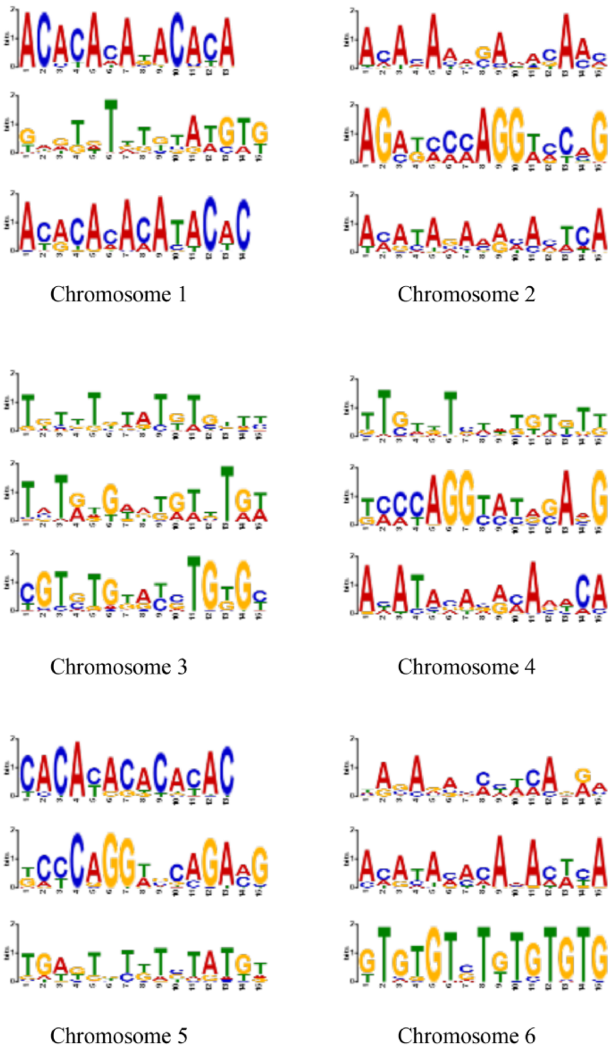


Figure 9: top three motifs found in the common DMRs (common to N = 5 exposures) for chromosomes 1-6. Motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to three. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

in a direct exposure to the F2 generation. Therefore, the F3 generation represents the first descendants with no direct exposure to the toxicant. Identification of DMRs of the DNA between the control and exposure lineage F3 generations indicates that the DMR was exposure-induced through epigenetic transgenerational inheritance [9].

The procedure for identifying DMRs in the transgenerational F3 generation involved a methylated DNA immunoprecipitation procedure followed by next-generation sequencing. The genome was divided into 1000bp regions, and DMRs with a specific pathology were identified. A *P* value was calculated for each of the 1000bp regions indicating the probability the region is not a DMR (non-DMR). Those regions whose *P* value $< 10^{-6}$ comprise the DMR set which constitutes the positive examples (DMRs) in the training examples used to train the hybrid DL-ML models. All molecular data have been deposited into the public database at NCBI under GEO #s: GSE113785 (vinclozolin), GSE114032 (DDT), GSE98683 (atrazine), GSE155922 (jet fuel), GSE157539 (dioxin), GSE158254 (pesticides), GSE158086 (methoxychlor), GSE163412 (plastics), and GSE152678 (glyphosate). In previous work [18], all the DMRs from all these datasets were used to train the model. In this work, a separate model is trained on each dataset using only the DMRs from that dataset.

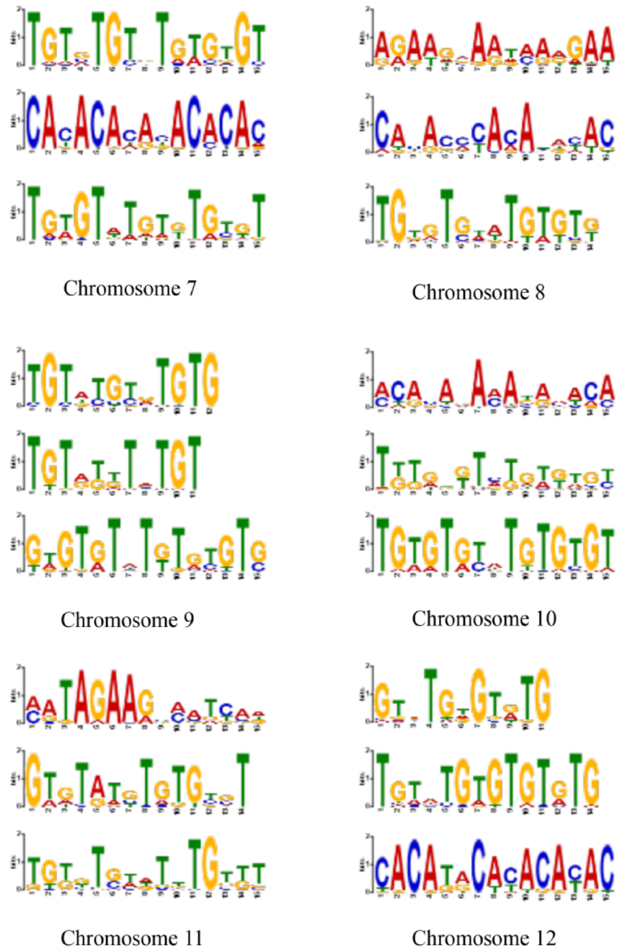


Figure 10: top three motifs found in the common DMRs (common to N = 5 exposures) for chromosomes 7-12. Motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to three. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value ≤ 0.05 were used

In these experiments, the number of DMRs meeting the *P* value threshold is a small fraction of the entire genome. However, regions that do not meet the *P* value threshold are not necessarily non-DMRs Thus, we seek a definition of a non-DMR that makes sense biologically and ideally is close to the number of DMRs to create a balanced training set for the learning model. Three constraints were considered for defining non-DMRs: (a) a region containing no CpGs, (b) a region which is a CpG-island (CpG-density > 10%), and (c) a region whose *P* value is greater than a specific threshold. The regions satisfying constraint (a) are non-DMRs because differential methylation is not possible without CpGs. The number of additional non-DMRs added by including constraints (b) and (c) was typically only 1-2% of the number of no CpG non-DMRs from constraint (a), but their addition as non-DMRs has a significant impact on whole-genome prediction. Therefore, regions satisfying constraints (a) and (b) were used as negative examples (non-DMRs) in the training set. The other constraint (c)

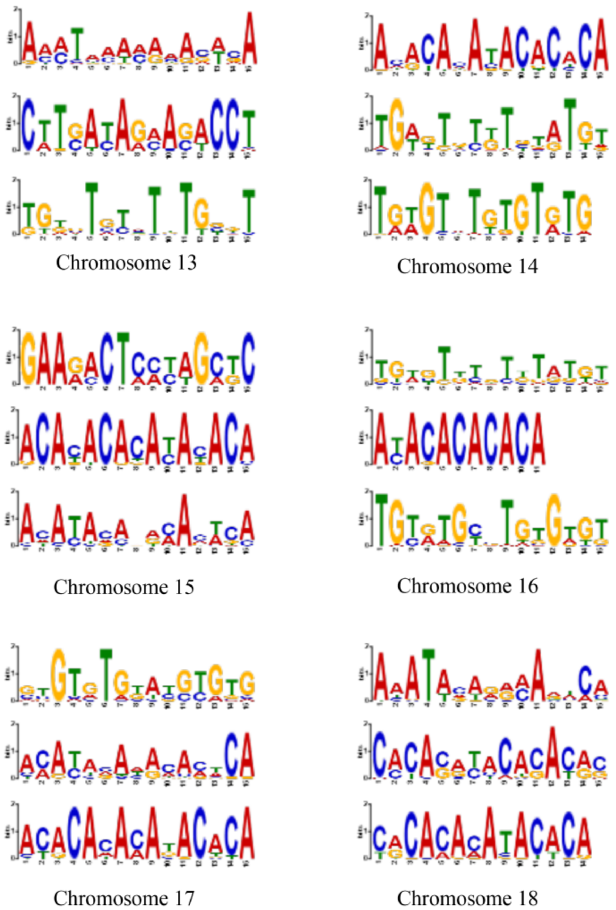


Figure 11: top three motifs found in the common DMRs (common to N = 5 exposures) for chromosomes 13-18. Motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to three. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value $\leq$ 0.05 were used

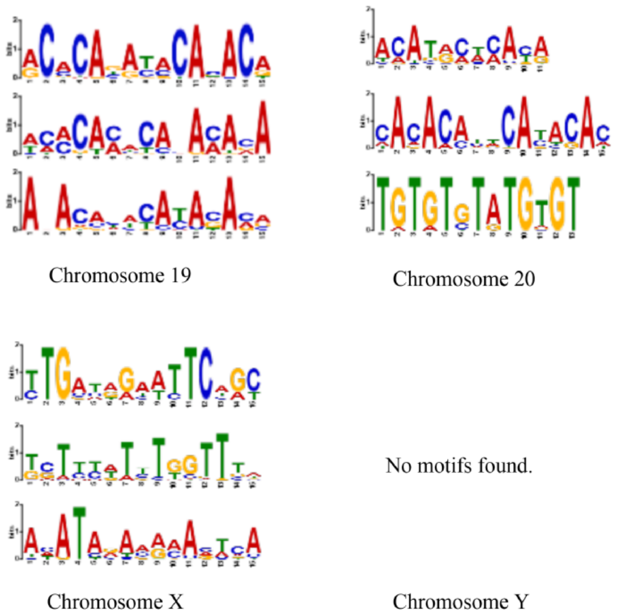


Figure 12: top three motifs found in the common DMRs (common to N = 5 exposures) for chromosomes 19, 20, and X. No motifs were found in chromosome Y. Motifs were identified using the MEME-ChIP discovery tool (https://meme-suite.org), using default web parameters, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" database for *R. norvegicus*, and the number of motifs to find was set to three. The MEME tool's default constraints on motif minimum width (6), maximum width (50), and E-value $\leq$ 0.05 were used

was considered for inclusion in the non-DMR samples but resulted in decreased performance.

The hybrid DL-ML model detailed in [18] takes a 1000bp region of the DNA sequence as input and produces a classification for the region as to whether it will be susceptible to environmental exposure as evidenced by differential methylation. The method is a hybrid model shown in Fig. 1 and consists of a DL network that is trained using the dataset and a traditional ML classifier that is also trained using the dataset, but with the input region re-expressed using features extracted from a layer of the deep learning network. The 1000bp DNA sequences are input to the DL network using a one-hot encoding, i.e. a 5×1000 array, where each column indicates which base-pair (A, C,G, T,N) is present. The network is trained using the training DMRs and non-DMRs. The training data are re-input to the trained network, and the outputs of the first convolutional layer are used as new extracted features to re-express each training example. The re-expressed training data are then used to train the XGBoost classifier. The prediction of the XGBoost classifier is used as the final prediction of DMR or non-DMR. The trained hybrid model is used to classify each region across the whole genome as to whether a region is susceptible to form a DMR in response to an ancestral environmental induced exposure. The hybrid DL-ML method has been successful at identifying DMRs not present in the training set [18]. The hybrid model has also been shown to outperform DL alone, ML alone, and alternative approaches to DMR prediction [18].

One issue with the hybrid approach is that the model's prediction has high variance. For example, two models trained on the same data can result in a significant difference in the set of DMRs predicted by the models. The variance is due to randomness in the training process, such as random initial weights and shuffling of training data. Even though one hybrid model predicts far fewer DMRs than all possible regions (based on the number of regions with at least one CpG), a model predicts nearly 20% of the genome as DMRs. There is a trade-off between two objectives for training the hybrid model, i.e. maintaining high model accuracy while avoiding overly general predictive models. To address this issue, multiple models are trained, and a core set of DMRs predicted by all models is identified. To find the proper number of trained models, a stopping point (SP) is defined, which indicates how many models are required to show a correlation among the core set of predicted DMRs. Given that a single model predicts N DMRs, if a set of N 1000bp regions were repeated selected at random from the genome, the SP is defined as the number of randomly selected sets of regions that would need to be intersected together for the intersection to be empty. If the same number of models are trained and their predicted DMRs intersected, then any DMRs remaining would have high certainty of being DMRs; these DMRs comprise the core set. The process used to determine SP for each exposure is shown in Table 1.

The next step is to define the core set of predicted DMRs as the intersection of the predicted DMRs from SP independently trained models. After generating the core set of DMRs for each exposure, the unique set of DMRs for each exposure can be determined. A unique DMR for an exposure is a region predicted as DMR in only that specific exposure. Once the unique DMRs for each exposure are identified, these DMRs are further analyzed by visualizing their locations on the genome, identifying known motifs among the DMRs, identifying genes associated with the DMRs, and identifying recurring motif patterns in the DMRs.

Table 2 summarizes the data and results for each exposure: the SP, the number of positive training samples in chromosome 7 (Training DMRs), the average number of predicted DMRs by a model (Predicted DMRs), the number of DMRs in the core set (intersection of DMRs predicted by SP models), and the number of unique regions in each exposure based on the training DMRs and based on the core DMRs as predicted by the whole-genome models. There were 6636 non-DMRs used for training in each exposure for chromosome 7. Due to the high number of training and predicted DMRs for the plastic exposure, identification of the core set of DMRs was prohibitive in time (training 165 models), and the core set is likely to be very large, which would tend to obscure unique DMRs in other exposures. Therefore, the plastic exposure DMRs were excluded from subsequent analyses. On the other extreme, there were only a small number of training DMRs, predicted DMRs, and unique DMRs for glyphosate. Table 2 shows only one unique core DMR for glyphosate on chromosome 7. For many chromosomes, there were zero DMRs for glyphosate. Therefore, the glyphosate exposure DMRs were also excluded from the analysis.

After composing the unique DMR set for each exposure, the TOMTOM tool is used to find the known motifs in the unique regions for each exposure [27]. Previous studies showed that methylated DNA fragments prevent the binding of TFs [1, 2]. As an example, CpGs are able to prevent binding TFs [1]. Identifying TF motif matches in unique DMRs can help in predicting the potential downstream effects of DNA methylation changes on gene expression and cellular processes. For example, if a TF binding site is differentially methylated in a cancer cell, it may affect the expression of downstream genes involved in tumor growth and progression. To find the TF binding specificity alignments, Catalog of Inferred Sequence Binding Preferences (CisBP) is used as the reference database (http://cisbp.ccbr.utoronto.ca/). CisBP is an online database of TF binding specificities. CisBP currently incorporates data from over 700 species covering more than 300 TF families, totaling more than 390 000 TFs (of which over 165 000 have at least one DNA binding motif). This method maps motifs across and within species, using DNA binding domain similarity thresholds [40].

The next analysis is to identify genes overlapping the DMRs unique to each exposure. Gene overlap occurs when a known gene shares the same region of a nucleotide sequence in a genome [41], where in this case the sequence is a 1000bp DMR unique to a particular exposure. Rat gene locations were obtained from the Rat Genome Database (https://rgd.mcw.edu). This experiment provides insights into the functional implications of DNA methylation changes. DMRs that overlap with genes are more likely to have functional consequences on gene expression and may be directly involved in disease development.

The next step in the analysis is to identify repeated motifs in each set of exposure specific DMRs. The top five repeated motifs in each set of exposure specific unique DMRs were identified using the MEME-ChIP discovery tool (https://meme-suite.org). The default parameters in the web-based interface were used for all runs, except the motifs were input from the "CIS-BP 2.00 Single Species DNA" for *Rattus norvegicus*, and the number of motifs to find was set to five. The MEME tool's default constraints on minimum width (6), maximum width (50), and E-value $\leq$ 0.05 were used. The MEME-ChIP tool searches for matches to a motif in both the forward primary sequence and the reverse complement sequence. But the motifs are visualized in the forward primary sequence order. These motifs can help to visualize distinct properties of the DMRs across different exposures. Computational methods for comparing motifs [27] may uncover more global patterns in the differences of motifs across different exposures.

The final step of the analysis is to apply the previous analysis steps to the common DMRs across all the exposures. Identifying the common DMRs across all the exposures can provide insights into the shared pathways and biological processes affected by different exposures. Table 6 shows the number of DMRs common to at least N exposures. None of the core DMRs are common to seven or more exposures. Since there were not any common DMRs across all the exposures, the DMRs that were common among at least five (N = 5) exposures were studied.

Author Contributions

Pegah Mavaie: conceptualization, formal analysis, investigation, validation, wrote original draft, and reviewed and edited manuscript.

Lawrence Holder: conceptualization, formal analysis, investigation, supervision, validation, writing, and reviewed and edited manuscript.

Michael Skinner: conceptualization, formal analysis, funding acquisition, investigation, supervision, validation, writing, and reviewed and edited the manuscript.

Data Availability

Data are uploaded as supplementary information.

Supplementary Data

Supplementary data is available at *EnvEpig* online.

Funding

This study was supported by John Templeton Foundation (50183 and 61174) (https://templeton.org/) grants to M.K.S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest statement. The authors have declared that no competing interests exist.

Ethics Statement. Not applicable.

References

1. Breton CV, Marutani AN. Air pollution and epigenetics: recent findings. *Curr Environ Health Rep* 2014;**1**:35–45.
2. Inbar-Feigenberg M, Choufani S, Butcher DT *et al.* Basic concepts of epigenetics. *Fertil Steril* 2013;**99**:607–15.
3. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. *J Mol Biol* 1987;**196**:261–82.
4. Cedar H. DNA methylation and gene activity. *Cell* 1988;**53**:3–4.
5. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* 2007;**8**:253–62.
6. Waddington CH. The epigenotype. 1942. *Int J Epidemiol* 2012;**41**:10–3.
7. Skinner MK, Nilsson EE. Role of environmentally induced epigenetic transgenerational inheritance in evolutionary biology: unified evolution theory. *Environ Epigenetics* 2021;**7**:dvab012, 1–12.

8. Kratz CP, Edelman DC, Wang Y *et al.* Genetic and epigenetic analysis of monozygotic twins discordant for testicular cancer. *Int J Mol Epidemiol Genet* 2014;**5**:135–9.
9. Nilsson E, Sadler-Riggleman I, Skinner MK. Environmentally induced epigenetic transgenerational inheritance of disease. *Environ Epigenetics* 2018;**4**:1–13, dvy016.
10. Sanchez OF, Lin L, Bryan CJ *et al.* Profiling epigenetic changes in human cell line induced by atrazine exposure. *Environ Pollut* 2020;**258**:1–11.
11. Thorson JLM, Beck D, Ben Maamar M *et al.* Epigenome-wide association study for atrazine induced transgenerational DNA methylation and histone retention sperm epigenetic biomarkers for disease. *PLoS One* 2020;**15**:1–29, e0239380.
12. Kubsad D, Nilsson EE, King SE *et al.* Assessment of glyphosate induced epigenetic transgenerational inheritance of pathologies and sperm epimutations: generational toxicology. *Sci Rep* 2019;**9**:1–17.
13. Martins R, Carruthers M. Testosterone as the missing link between pesticides, Alzheimer disease, and Parkinson disease. *JAMA Neurol* 2014;**71**:1189–90.
14. Paul KC, Chuang Y-H, Cockburn M *et al.* Organophosphate pesticide exposure and differential genome-wide DNA methylation. *Sci Total Environ* 2018;**645**:1135–43.
15. Yan D, Zhang Y, Liu L *et al.* Pesticide exposure and risk of Parkinson's disease: dose-response meta-analysis of observational studies. *Regul Toxicol Pharmacol* 2018;**96**:57–63.
16. Skinner MK, Manikkam M, Tracey R *et al.* Ancestral dichlorodiphenyltrichloroethane (DDT) exposure promotes epigenetic transgenerational inheritance of obesity. *BMC Med* 2013;**11**:228, 1–16.
17. Nilsson EE, Ben Maamar M, Skinner MK. Role of epigenetic transgenerational inheritance in generational toxicology. *Environ Epigenetics* 2022;**8**:dvac001, 1–9.
18. Mavaie P, Holder L, Beck D *et al.* Predicting environmentally responsive transgenerational differential DNA methylated regions (epimutations) in the genome using a hybrid deep-machine learning approach. *BMC Bioinform* 2021;**22**: 1–25.
19. King SE, McBirney M, Beck D *et al.* Sperm epimutation biomarkers of obesity and pathologies following DDT induced epigenetic transgenerational inheritance of disease. *Environ Epigenetics* 2019;**5**:1–15, dvz008.
20. Ben Maamar M, Beck D, Nilsson EE *et al.* Epigenome-wide association study for glyphosate induced transgenerational sperm DNA methylation and histone retention epigenetic biomarkers for disease. *Epigenetics* 2021;**16**:1150–67.
21. Nilsson E, King SE, McBirney M *et al.* Vinclozolin induced epigenetic transgenerational inheritance of pathologies and sperm epimutation biomarkers for specific diseases. *PLoS One* 2018;**13**:1–29, e0202662.
22. Thorson JLM, Beck D, Ben Maamar M *et al.* Epigenome-wide association study for pesticide (Permethrin and DEET) induced DNA methylation epimutation biomarkers for specific transgenerational disease. *Environ Health* 2020;**19**:1–9.
23. Ben Maamar M, Nilsson E, Thorson JLM *et al.* Transgenerational disease specific epigenetic sperm biomarkers after ancestral exposure to dioxin. *Environ Res* 2020;**192**:1–15.
24. Ben Maamar M, Nilsson E, Thorson JLM *et al.* Epigenome-wide association study for transgenerational disease sperm epimutation biomarkers following ancestral exposure to jet fuel hydrocarbons. *Reprod Toxicol* 2020;**98**:61–74.
25. Nilsson E, Thorson JLM, Ben Maamar M *et al.* Epigenome-Wide Association Study (EWAS) for potential transgenerational disease epigenetic biomarkers in sperm following ancestral exposure to the pesticide methoxychlor. *Environ Epigenetics* 2020;**6**:1–25, dvaa020.
26. Thorson JLM, Beck D, Ben Maamar M *et al.* Ancestral plastics exposure induces transgenerational disease-specific sperm epigenome-wide association biomarkers. *Environ Epigenetics* 2021;**7**:1–13, dvaa023.
27. Gupta S, Stamatoyannopoulos JA, Bailey TL *et al.* Quantifying similarity between motifs. *Genome Biol* 2007;**8**:1–9.
28. Boffetta P, Adami H-O, Berry C *et al.* Atrazine and cancer: a review of the epidemiologic evidence. *Eur J Cancer Prev* 2013;**22**:169–80.
29. Carrión DV, Kitagawa Y, Zhang J *et al.* The role of the Bub1 gene in aneuploidy and cancer progression. *Cancer Res* 2004;**64**:994–5.
30. Wang X, Zhou J, Shen M *et al.* Chlorpyrifos exposure induces lipid metabolism disorder at the physiological and transcriptomic levels in larval zebrafish. *Acta Biochim Biophys Sin (Shanghai)* 2019;**51**:890–9.
31. Cole P, Trichopoulos D, Pastides H *et al.* Dioxin and cancer: a critical review. *Regul Toxicol Pharmacol* 2003;**38**:378–88.
32. Hsu EL, Yoon D, Choi HH *et al.* A proposed mechanism for the protective effect of dioxin against breast cancer. *Toxicol Sci* 2007;**98**:436–44.
33. Bedenk J, Vrtacnik-Bokal E, Virant-Klun I. The role of anti-Mullerian hormone (AMH) in ovarian disease and infertility. *J Assist Reprod Genet* 2020;**37**:89–100.
34. Campbell TL, De Silva EK, Olszewski KL *et al.* Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog* 2010;**6**:1–15.
35. Khund-Sayeed S, He X, Holzberg T *et al.* 5-Hydroxymethylcytosine in E-box motifs ACAT|GTG and ACAC|GTG increases DNA-binding of the B-HLH transcription factor TCF4. *Integr Biol (Camb)* 2016;**8**:936–45.
36. Tian W, Han X, Yan M *et al.* Structure-based discovery of a novel inhibitor targeting the beta-catenin/Tcf4 interaction. *Biochemistry* 2012;**51**:724–31.
37. Parry TJ, Theisen JWM, Hsu J-Y *et al.* The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* 2010;**24**:2013–8.
38. Platero AI, García-Jaramillo M, Santero E *et al.* Transcriptional organization and regulatory elements of a Pseudomonas sp. strain ADP operon encoding a LysR-type regulator and a putative solute transport system. *J Bacteriol* 2012;**194**:6560–73.
39. Mattie DR, Sterner TR. Past, present and emerging toxicity issues for jet fuel. *Toxicol Appl Pharmacol* 2011;**254**:127–32.
40. Lambert SA, Yang AWH, Sasse A *et al.* Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet* 2019;**51**:981–9.
41. Schlub TE, Buchmann JP, Holmes EC. A simple method to detect candidate overlapping genes in viruses using single genome sequences. *Mol Biol Evol* 2018;**35**:2572–81.