# Elucidating disease-associated mechanisms triggered by pollutants via the epigenetic landscape using large-scale ChIP-Seq data

Zhaonan Zou[1], Yuka Yoshimura[1], Yoshihiro Yamanishi[2] and Shinya Oki[1,*]

[1] Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan
[2] Department of Complex Systems Science, Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan

* To whom correspondence should be addressed. Tel: +81-75-366-7479; Email: oki.shinya.3w@kyoto-u.ac.jp

## ABSTRACT

Despite well-documented effects on human health, the action modes of environmental pollutants are incompletely understood. Transcriptome-based approaches are widely used to predict associations between chemicals and disorders. However, the molecular cues regulating gene expression remain unclear. To elucidate the action modes of pollutants, we proposed a data-mining approach, termed "DAR-ChIPEA," combining epigenome (ATAC-Seq) and large-scale public ChIP-Seq data (human, $n$ = 15,155; mouse, $n$ = 13,156) to identify transcription factors (TFs) that are enriched not only in gene-adjacent domains but also across differentially accessible genomic regions, thereby integratively regulating gene expression upon pollutant exposure. The resultant pollutant–TF matrices are then cross-referenced to a repository of TF–disorder associations to account for pollutant modes of action. For example, TFs that regulate Th1/2 cell homeostasis are integral in the pathophysiology of tributyltin-induced allergic disorders; fine particulates ($PM_{2.5}$) inhibit the binding of C/EBPs, Rela, and Spi1 to the genome, thereby perturbing normal blood cell differentiation and leading to immune dysfunction; and lead induces fatty liver by disrupting the normal regulation of lipid metabolism by altering hepatic circadian rhythms. Thus, our approach has the potential to reveal pivotal TFs that mediate adverse effects of pollutants, thereby facilitating the development of strategies to mitigate environmental pollution damage.

## INTRODUCTION

Short- and/or long-term exposure to environmental pollutants has been associated with assorted adverse human health outcomes, including increased respiratory symptoms, heart or lung disease, and even early mortality (1–4). Toxicant pollutants in particular are implicated in carcinogenesis and other serious health effects, such as reproductive disorders and congenital anomalies (5–7). To elucidate the modes of action (MoAs) by which chemicals, including environmental contaminants, elicit such toxic responses, computational strategies utilizing biomedical big data analysis have attracted the attention of researchers in recent decades (8–10). For example, with the understanding that bioactive chemicals, such as pharmaceuticals and pollutants, affect human physiology by either respectively maintaining or disrupting normal gene expression, comparative analyses of transcriptional profiles before and after chemical exposure, followed by gene ontology (GO) and pathway enrichment analyses, have been performed to elucidate chemical MoAs (11,12). However, although GO and pathway analyses can comprehensively extract functional characteristics from profiles of chemically induced differentially expressed genes (DEGs), considerable effort is required to identify specific primary literature that demonstrates the association between input DEGs and associated GOs or pathways. In addition, graph analytics (network analysis) is also frequently employed to estimate "molecular hubs" within chemically responsive gene regulatory networks, despite limitations in terms of the biological interpretability of the prediction outcomes (13).

Bioactive chemicals, including a proportion of pollutants, can target a wide variety of molecules, such as membrane transport proteins, ion channels, and metabolic enzymes. In particular, research on steroid receptors, which are activated by female (estrogens) and male (androgens) hormones, is actively pursued owing to the reported effects of endocrine-disrupting chemicals, also known as environmental hormones, on reproduction and development (14,15). In addition, heavy metals, including mercury and cadmium, which have been associated with pollution-related disorders such as Minamata disease and Itai-itai disease, may induce epigenetic changes as a component of their disease mechanisms (16,17). Furthermore, both intranuclear steroid receptors and enzymes that regulate epigenomic states, such as DNA methylation and histone modifications, can function as transcription factors (TFs) to directly control gene expression. Accordingly, certain pharmaceuticals such as estrogen receptor inhibitors and histone deacetylase inhibitors have been shown to exert

anticancer effects by directly regulating TF function (18,19). Together, these observations suggest that TFs that collectively regulate the on/off switches of genes whose expression is significantly altered by exposure to environmental chemicals may function as direct targets or mediators of chemical MoAs.

To identify TFs whose binding is enriched to a given set of genes or genomic regions, binding motif-based methods such as HOMER and MEME are widely used (20,21). This approach predicts TF binding by extracting shared short nucleotide sequences, commonly referred to as "motifs", from multiple input genomic regions. These motifs are then compared to known TF–DNA binding motifs by comparative analysis. As an alternative method for identifying key TFs in chemical MoAs, we have recently developed an approach that is independent of TF–DNA binding motifs, which we termed the DEG-ChIPEA method (22). This method is based on the ChIP-Atlas database, which extracts actual TF binding sites on the reference genome by curating and integrating tens of thousands of previously published ChIP-Seq experimental data examining TF binding to genomic DNA (23). The DEG-ChIPEA method involves inputting chemically induced DEGs and performing an enrichment analysis based on all ChIP-Seq experiments obtained from ChIP-Atlas (i.e., ChIP-Seq-based enrichment analysis; ChIPEA) to identify TFs whose binding is significantly enriched in the vicinity of the input DEGs. These identified TFs can be considered as promising candidate pivotal factors mediating the expression of therapeutic or adverse chemical effects. However, despite the generally high accuracy of DEG-ChIPEA, it retains the potential limitation that only TFs in the immediate vicinity of the transcription start site (transcription start site ± 5 kb) of the input chemically induced genes are analyzed. Thus, potential contributions to gene expression by TFs that bind to distal transcriptional regulatory regions, such as enhancers and topologically associating domain boundaries located far from the gene locus, are not taken into account.

In this study, we aimed to gain further insight into the MoAs of chemicals, particularly environmental pollutants, with regard to health effects. To achieve this, we propose a novel approach, termed "DAR-ChIPEA", that considers TF binding events throughout the genome, including both proximal and distal transcriptional regulatory regions of genes, in contrast to the limited purview of the DEG-ChIPEA method. The DAR-ChIPEA method involves extracting differentially accessible regions (DARs), which are genomic regions with significantly different

chromatin accessibility prior to and following chemical exposure, using data from environmental chemical perturbation epigenome experiments, such as ATAC-Seq datasets. These DARs are then used as input for ChIPEA to generate a chemical–TF association matrix, which allows the identification of TFs whose binding is significantly enriched in the input DARs. In addition, we constructed a chemical–TF–disorder triadic association by incorporating known TF–disease associations into the resulting chemical–TF matrix. Our approach outperformed methods based on TF–DNA binding motifs, as well as the previously developed DEG-ChIPEA method (22), which focuses only on TFs that bind in the vicinity of chemically induced DEGs.

## MATERIAL AND METHODS

### Reagents

Not applicable.

### Biological Resources

Not applicable.

### Web Sites/Data Base Referencing

The URLs and dates of access or download of websites and databases used in the study are as follows:

DBKERO (https://kero.hgc.jp/; accessed on March 30, 2022) (24), TaRGET (https://targetepigenomics.org/; accessed on March 30, 2022) (25), National Center for Biotechnology Information Sequence Read Archive (NCBI SRA; https://www.ncbi.nlm.nih.gov/sra), ChIP-Atlas (https://chip-atlas.org/; accessed on March 30, 2022) (23), DisGeNET (v7.0; https://www.disgenet.org/static/disgenet_ap1/files/downloads/curated_gene_disease_associations.tsv.gz; accessed on June 4, 2020) (26), The Comparative Toxicogenomics Database (CTD; chemical–disease interactions dataset; http://ctdbase.org/reports/CTD_chemicals_diseases.tsv.gz; accessed on June 4, 2020) (27), DEG-ChIPEA and DAR-ChIPEA sample codes (https://github.com/zouzhaonan/dar_moa/),

HOCOMOCO (v11; https://hocomoco11.autosome.org/final_bundle/hocomoco11/full/MOUSE/mono/HOCOMOCOv11_full_MOUSE_mono_meme_format.meme; downloaded on May 3, 2022) (28), DEG-MotifEA and DAR-MotifEA sample codes (https://github.com/zouzhaonan/dar_moa/), and Mouse Genome Informatics database (https://www.informatics.jax.org/; accessed on April 15, 2023) (29).

The software versions and analytical protocols used in the study are as follows:

SRA Toolkit (v3.0.0; https://github.com/ncbi/sra-tools), FASTX-Toolkit (v0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/), HISAT2 (v2.1.0) (30), featureCounts (v2.0.3) (31), R (v4.1.3; https://www.r-project.org/), edgeR (v3.40.2) (32), BEDtools (v2.23.0) (33), MEME Suite (v5.4.1) (21), ROCR (v1.0-11; http://ipa-tys.github.io/ROCR/), and Genomic Regions Enrichment of Annotations Tool (GREAT; v4.0.4) (34).

**Chemical- or Pollutant-Perturbed Transcriptome and Epigenome Data**

To identify genomic regions with significant differences in chromatin accessibility (DARs) in response to chemical or pollutant exposure as determined through epigenome perturbation experiments, we used ATAC-Seq datasets retrieved from DBKERO ($n$ = 2,078; DRA006903–DRA006930) and TaRGET ($n$ = 383; GSE146508) databases (24,25). In addition, RNA-Seq datasets corresponding to the ATAC-Seq datasets were made available in these databases and used to identify DEGs induced by chemical or pollutant exposure as input to the baseline analyses (Supplementary Figure S1; DBKERO, $n$ = 2,012, DRA006875–DRA006902; TaRGET, $n$ = 468, GSE146508). DBKERO accumulates the perturbation data of 99 chemicals in human cell lines, whereas TaRGET curates *in vivo* exposure data for seven pollutants of blood and liver samples from mice. All data used in this analysis were archived in the NCBI SRA and are associated with accession IDs as listed in Supplementary Tables S1–S4.

**Genome-Wide TF Binding and Chromatin Accessibility Experimental Data from ChIP-Atlas**

Genome-wide TF binding site information was obtained from the ChIP-Atlas database (23), which comprehensively integrates nearly all publicly available ChIP-Seq data. Peak call data in BED format (MACS2 (35); $Q$-value $<1 \times 10^{-10}$) were used in this study. Data from ChIP-Seq

experiments examining binding of 1,796 and 866 TFs on the human (GRCh38/hg38; hereafter referred to as hg38) and mouse genomes (GRCm38/mm10; hereafter referred to as mm10), respectively, were used in this study (hg38, $n$ = 15,155; mm10, $n$ = 13,156). In addition, the information related to accessible chromatin used in this analysis to identify DARs induced by chemical or pollutant exposure was also obtained from ChIP-Atlas, which also provides peak call data for ATAC-Seq.

**TF–Disease Associations**

The association data between TFs and diseases were obtained from DisGeNET (v 7.0) (26), which is a database of genetic and genomic information associated with human diseases. DisGeNET integrates this information from other expert-curated repositories and PubMed literature. Only manually curated data that were labeled with referenced PubMed IDs were retrieved, resulting in a total of 12,857 associations between TFs and diseases, involving 1,004 TFs and 3,205 diseases.

**Known Chemical–Disease Associations as Standard Data**

Manually curated chemical–disease association data were acquired from the CTD (27). In total, we incorporated 28,821 chemical–disease associations involving 69 chemicals (DBKERO) and seven pollutants (TaRGET), and 2,150 diseases in this study.

**Identification of Chemical- or Pollutant-Induced DEGs**

After obtaining the SRA accession IDs of the chemical or pollutant perturbation RNA-Seq data from DBKERO and TaRGET, the raw sequencing data archived in NCBI SRA were downloaded and decoded into FASTQ format using the "fastq-dump" command from the SRA Toolkit (v3.0.0). Quality control was then performed using the "fastq_quality_filter" command from the FASTX-Toolkit (v0.0.13). Alignment to the reference genomes (DBKERO, hg38; TaRGET, mm10) was performed using HISAT2 (v2.1.0) (30), and expression quantification was performed using featureCounts (v2.0.3) (31). All commands were run with default settings and parameters. DEGs were detected using R Package "edgeR" (v3.40.2) (32). The statistical significance used for identification of DEGs is detailed in the Statistical Analyses section.

**Identification of Chemical- or Pollutant-Induced DARs**

Referring to the SRA accession IDs for the chemical- or pollutant-perturbed ATAC-Seq data obtained from DBKERO and TaRGET, the peak call data (in BED format) for the corresponding ATAC-Seq experiments were acquired from ChIP-Atlas. Peak calling was performed using MACS2 (35) in ChIP-Atlas; only peaks with a *Q*-value of $<1 \times 10^{-10}$ were considered as statistically significant accessible chromatin regions. Next, common open chromatin regions across all replicates for each chemical or pollutant perturbation condition were extracted using BEDtools (v2.23.0) (33). Differential analysis was then performed to identify DARs induced by the chemicals or pollutants. Note that the term "differential analysis" in this context refers to the identification of genomic regions with no overlap between groups with and without chemical or pollutant exposure. In this study, DARs that exhibited relaxed or condensed chromatin upon chemical or pollutant exposure are referred to as "opened" or "closed" DARs, respectively.

**TF Enrichment Analysis Fully Incorporating Public ChIP-Seq Experiments (ChIPEA)**

Details of the ChIPEA procedure using the application programming interface are provided at https://github.com/inutano/chip-atlas/wiki/Perform-Enrichment-Analysis-programmatically (22,36). To exploit the extensive ChIP-Seq data in ChIP-Atlas (23), we performed ChIPEA to profile TFs whose binding sites were enriched around genomic regions or genes of interest. In particular, using the genomic coordinates of opened and closed DARs, or the gene symbols of genes up- and down-regulated by a query chemical or pollutant as input, we first counted the number of intersects between the input data and peak-call data (MACS2; *Q*-value $<1 \times 10^{-10}$) of all TF-related experiments archived in the ChIP-Atlas database, using the "intersectBed" command of BEDTools (v2.23.0) (33). Note that 5 kb upstream and downstream of the transcription start site were added for each gene when chemically induced DEGs were queried. Enrichment scores ($-\log_{10}P$) were then calculated using the two-tailed Fisher's exact probability test, with the null hypothesis that equivalent proportions of the two data sets (opened and closed DARs, or up- and down-regulated genes) overlap with the ChIP-Seq peak-call data; fold enrichment values were concurrently obtained. If a chemical– or pollutant–TF association was indicated by multiple ChIP-Seq experiments, only the highest enrichment score was adopted. Parameters used in application

programming interface-based ChIPEA were as follows: "genome=hg38 (DBKERO) or mm10 (TaRGET); antigenClass=TFs and others; cellClass=All cell types; threshold=100." Sample codes for DEG-ChIPEA and DAR-ChIPEA are provided at https://github.com/zouzhaonan/dar_moa/.

**TF Enrichment Analysis Focusing on TF–DNA Binding Motifs (MotifEA)**

In this method, TF binding to chemical- or pollutant-induced DARs and DEGs was estimated using MotifEA with STREME and TOMTOM in the MEME Suite (v5.4.1) (21), using default settings termed DAR-MotifEA and DEG-MotifEA, respectively. STREME was first run to identify 1,000 sequence motifs from queried multiple genomic regions, i.e., chemical- or pollutant-induced DARs or DEGs. Subsequently, the TFs with enriched binding to the input genomic regions or genes were estimated by matching the identified motifs with known TF binding motifs obtained from HOCOMOCO v11 (28) using TOMTOM. Note that 5 kb upstream and downstream of the transcription start site were added for each gene when chemically induced DEGs were queried. Sample codes for DEG-MotifEA and DAR-MotifEA are provided at https://github.com/zouzhaonan/dar_moa/.

**Establishment of Chemical– or Pollutant–TF–Disorder Triadic Associations**

Using TFs enriched in chemical- or pollutant-induced DEGs and DARs as mediators, the TF–disease associations retrieved from DisGeNET (26) were integrated with the chemical– or pollutant–TF matrices obtained as outcomes of MotifEA or ChIPEA, thereby establishing chemical– or pollutant–TF–disorder triadic associations. The enrichment scores obtained via ChIPEA or MotifEA were continuously used as a quantitative measure to assess the degree of newly constructed chemical– or pollutant–disorder associations.

**Calculation of Global Area under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision Recall Curve (AUPR) Scores**

After validating their authenticity by referencing known chemical–disease associations from CTD (27), all predicted chemical– or pollutant–disorder associations were reorganized into a single matrix with dimensions of $m \times n$, where $m$ and $n$ represent the total number of chemical or pollutant perturbation conditions and the number of disorders, respectively. This matrix contained enrichment scores along with "known association" and "unknown

association" tags. The maximum enrichment score within each column (disorder) was then extracted and stored as a vector of *n* elements. ROC (which plots the true positive rate against the false positive rate) and PR (which plots the precision [positive predictive value] against the recall [sensitivity]) curves were then generated using this vector (R Package "ROCR" v1.0-11). The performance results were summarized into a global AUROC score, where 1 represents perfect classification and 0.5 represents random classification, and AUPR score, where 1 represents perfect inference and the proportion of positive examples in the standard data corresponds to random inference.

### Mouse Phenotype Ontology Enrichment Analysis

GO enrichment analysis on opened DARs induced by pollutant exposure was performed using GREAT (v4.0.4) (34). Default parameters were used in this analysis; namely, "Species assembly: mm10; Association rule: Basal+extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension, curated regulatory domains included." Ontologies categorized as "Mouse Phenotype" were then extracted. Parent category information for individual ontology terms was obtained with reference to the Mouse Genome Informatics database (29).

### Statistical Analyses

Statistical analysis was performed using R (v4.1.3). Unpaired two-tailed Wilcoxon rank-sum tests were used to compare AUROC and AUPR scores between two groups. Fisher's exact test was used for 2 × 2 contingency analysis. Sample sizes and *P*-values are shown in the figures and *P*-value < 0.05 was considered as statistically significant. NS denotes no significance. Chemical- and pollutant-induced DEGs were identified using the R package edgeR (v3.40.2), and a *P*-value <0.05 (DBKERO) or an adjusted *P*-value <0.1 (TaRGET) with absolute value of $\log_2$ fold change >1 was recognized as statistically significant.

## RESULTS AND DISCUSSION

### Study Design: Elucidation of Chemical MoAs Focusing on Chemically Reorganized Chromatin Accessibility

To expand upon our prior approach (DEG-ChIPEA) (22), here, we present a novel analytical approach using chemically induced DARs as input to cover the full spectrum of TF binding to both proximal and distal transcription regulatory regions (e.g. promotors and enhancers; Figure 1A). Our approach identifies TFs that integratively regulate chemical- or pollutant-induced DARs by comprehensively analyzing tens of thousands of publicly available ChIP-Seq data obtained from the ChIP-Atlas database (Figure 1B) (23). In addition, we integrated the predicted chemical–TF associations with disease-associated TF data obtained from the DisGeNET database (26). The resulting chemical–TF–disorder triadic associations were then statistically validated using a known chemical–disease association dataset from CTD as standard data (Supplementary Figure S1) (27).

### Validation of the Proposed DAR-ChIPEA Approach to Predict TF-Driven Actions for *In Vitro* Chemical Perturbation

Because TaRGET (25) contains information on only a few types of pollutants, it is not sufficient to statistically evaluate the reliability of DAR-ChIPEA. Therefore, we first evaluated the prediction accuracy of our method using the DBKERO database (24), which contains data on hundreds of chemical perturbations to cell cultures, providing correlated RNA-Seq and ATAC-Seq datasets. After retrieving pre-processed peak call datasets of ATAC-Seq from cell lines with and without chemical treatment in the DBKERO database from ChIP-Atlas, we performed a comparative analysis to detect DARs opened or closed by chemical perturbation (Supplementary Figure S2). Next, we fully exploited almost all publicly available human TF ChIP-Seq data in ChIP-Atlas to identify TFs enriched for binding to the extracted DARs using ChIPEA (DAR-ChIPEA). This allowed us to generate chemical–TF associations for each drug treatment condition. Furthermore, to evaluate the pivotality of TFs in drug MoAs, chemical–TF–disorder associations were then constructed by assigning TF-related disease information obtained from DisGeNET to the resulting chemical–TF association matrix. For comparison, in the chemical–TF association generation step, we also used chemically induced DEGs as input to the ChIPEA (DEG-ChIPEA), as previously reported (Supplementary Figure S2) (22). For the same purpose, a more conventional TF enrichment analysis method focusing on TF–DNA binding motifs (MotifEA) was also performed with DEGs or DARs as input (hereafter referred to as DEG-MotifEA and DAR-MotifEA; Figure 2A).

We observed that the statistically significant maximum enrichment scores ($-\log_{10}P$-value) obtained from the proposed DAR-ChIPEA method were globally higher than those from the baseline DEG-ChIPEA method (Figure 2B; mean maximum enrichment score = 10.25 and 8.139, respectively; $P = 3.013 \times 10^{-4}$; two-tailed Wilcoxon rank-sum test), indicating that DAR-ChIPEA has a wider dynamic range than DEG-ChIPEA. Thus, DAR-ChIPEA offers a significant advantage in discriminating associated and non-associated TFs with regard to individual chemicals.

We then examined whether the chemical–TF–disorder triadic associations derived from DAR-ChIPEA that exhibited higher statistical significance were congruent with existing knowledge. The distribution of AUROC and AUPR scores for each dosing condition, generated using manually curated chemical–disease association data obtained from CTD as reference data, (1,457, 1,711, 1,621, and 1,800 conditions for DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and DAR-ChIPEA, respectively) was depicted using bee swarm plots (Figure 2C, D; Supplementary Table S5). The mean AUROC and AUPR scores of the DAR-ChIPEA approach across all dosage conditions were 0.7287 and 0.4164, respectively. These scores were significantly higher than those of the baseline methods (DEG-MotifEA: mean AUROC 0.6736; mean AUPR 0.3488, DAR-MotifEA: mean AUROC 0.6769; mean AUPR 0.3641, and DEG-ChIPEA: mean AUROC 0.7060; mean AUPR 0.3862) ($P<0.05$ for all; two-tailed Wilcoxon rank-sum test).

In addition, because the maximum enrichment scores varied drastically depending on the given dosing conditions, we were concerned that the discrimination of items was limited when comparing TFs within a dosing condition with low maximum enrichment scores. Therefore, we also calculated "global" statistics using an inter-chemical merged enrichment score vector to emphasize the significance of the actual values of the enrichment scores (detailed in the Materials and Methods section). The global AUROC and AUPR scores of the proposed approach were 0.7894 and 0.6137, respectively, significantly outperforming those of the baseline methods ($P<0.05$; two-tailed Wilcoxon rank-sum test; Figure 2C, D). These results suggest that the ChIPEA method using chemically induced DARs as input is a powerful approach for identifying pivotal TFs, elucidating key factors for specific chemical MoAs, and predicting disorders or phenotypes associated with chemical treatments. The use of this approach may ultimately lead to increased predictive power compared to those of the existing DEG-ChIPEA method and MotifEA-based computational methods.

**Application of the DAR-ChIPEA Approach to Predict TF-Driven MoAs upon Pollutant Administration *In Vivo***

We next employed the DAR-ChIPEA method to predict the disorders caused by pollutants. After retrieving ATAC-Seq peak call datasets obtained with and without pollutant exposure from the TaRGET database available in ChIP-Atlas, we identified opened and closed DARs induced by various pollutants as described in the Materials and Methods section (Supplementary Figure S2). We used publicly available mouse TF ChIP-Seq data in ChIP-Atlas to generate pollutant–TF associations using the proposed DAR-ChIPEA approach. For comparison, we simultaneously performed analyses using the baseline methods, namely, DEG-ChIPEA, DEG-MotifEA, and DAR-MotifEA (Figure 3A). The resulting pollutant–TF matrices were then used to generate pollutant–TF–disorder associations, focusing on pivotal TFs, with reference to TF–disease associations from the DisGeNET database. We evaluated the accuracy of the predicted pollutant–disorder associations using ROC and PR curves, using known chemical-disease association datasets from CTD serving as a standard.

The maximum enrichment scores attained using the proposed DAR-ChIPEA approach were globally higher than those achieved with the baseline DEG-ChIPEA method, similar to the results obtained using *in vitro* input data, further supporting the wider dynamic range for discriminating associated TFs for individual pollutants (Figure 3B; mean maximum enrichment score = 76.77 and 11.57, respectively; $P = 8.421 \times 10^{-8}$; two-tailed Wilcoxon rank-sum test).

The distribution of AUROC and AUPR scores for each exposure condition (40, 28, 40, and 21 conditions for DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and DAR-ChIPEA, respectively) was visualized using bee swarm plots (Figure 3C, D; Supplementary Table S6). The mean AUROC of the proposed DAR-ChIPEA approach across exposure conditions was 0.6748, significantly higher than those of the baseline methods (DEG-MotifEA, 0.6071; DAR-MotifEA, 0.6147; DEG-ChIPEA, 0.6575; $P<0.05$; two-tailed Wilcoxon rank-sum test). However, the DAR-ChIPEA AUPR score did not significantly ($P>0.05$) exceed those of the baseline methods (DEG-MotifEA, 0.7656; DAR-MotifEA, 0.7908; DEG-ChIPEA, 0.7890; proposed: DAR-ChIPEA, 0.7929). The global AUROC and global AUPR scores of DAR-ChIPEA were 0.7492 and 0.8491, respectively. These results suggest that the DAR-ChIPEA method performed satisfactorily in

terms of AUROC; the absence of significant difference regarding AUPR may be attributed to the limited number of datasets analyzed.

**Biological Interpretation of the Predicted Pollutant–TF–Disorder Associations**

In the proposed DAR-ChIPEA method, high AUROC scores were obtained for tributyltin (TBT) (blood samples), $PM_{2.5}$ (blood samples), and lead (liver samples) (AUROC = 0.7028, 0.7030, and 0.7349, respectively; Supplementary Table S4, S6; highlighted with bold red font), which motivated us to evaluate the results of putative pivotal TFs involved in the MoAs of these pollutants in further detail (Figure 4; Supplementary Tables S7–S9). The organotin compound TBT is widely utilized in various industrial applications such as pesticides, paints, preservatives, and heat stabilizers for plastics. Nevertheless, TBT has been associated with a potentially increased risk of a number of lifestyle-related diseases, including obesity, fatty liver, cardiovascular disease, and abnormal glucose metabolism (37–40). We initially conducted ontology enrichment analysis to discern any common nature of genes located proximal to opened DARs of blood samples upon TBT exposure using mammalian phenotype (MP) ontology. Notably, we found that the enriched gene sets were prominently manifested in phenotypic features of the immune system and hematopoietic system. To further understand the transcriptional regulatory mechanisms involved the response to TBT exposure, we then performed DAR-ChIPEA. The results showed that the opening of specific DARs susceptible to TBT exposure could enhance the enriched binding of Pparg, a master regulator of adipogenesis (41). This is firmly consistent with the *in vitro* findings that TBT stimulates lipid accumulation by activating Pparg (42).

In addition, TBT can induce apoptosis in diverse tissues and cell categories, such as human peripheral blood lymphocytes and human amnion cells, and high-dose exposure to TBT has been reported to induce T lymphocyte apoptosis and thymic atrophy (43–45). Furthermore, TBT suppresses the differentiation of naive $CD4^+$ cells into Th1 cells, while stimulating the differentiation into Th2 cells, thereby causing allergic disorders through the overproduction of Th2 cells, although the MoAs remain incompletely understood (46). Our results showed that TBT exposure led to the enrichment of TFs such as Rela, Stat1, and Rbpj in the opened DARs, consistent with the pro-inflammatory nature of TBT. Conversely, the binding of Bcl11b and Tbx21 was markedly enriched in the closed DARs. BCL11B has apoptosis-inhibitory properties and is involved in maintaining T cell differentiation, as supported by the reduced

numbers of T cells observed in *Bcl11b*-deficient mice (47,48). Notably, TBX21 also plays a key role in Th1 cell differentiation, with *Tbx21*-deficient naive Th cells exhibiting distorted Th2 cell polarization, a phenotype somewhat similar to that observed upon TBT exposure (49,50). These findings suggest that Bcl11b and Tbx21 may contribute to TBT-induced disruption of the Th1/Th2 balance. Overall, these results imply that the DAR-ChIPEA approach is valuable for identifying key regulators involved in the physiological mechanisms of TBT exposure underlying the associated lifestyle-related diseases and allergic diseases.

$PM_{2.5}$ is fine particulate matter suspended in the atmosphere with a diameter of 2.5 µm or less, which has the ability to transport pollutants in the environment. Exposure to $PM_{2.5}$ is associated with respiratory diseases such as asthma, bronchitis, and pneumonia, as well as heart diseases such as arrhythmias and atherosclerosis, and significantly increases mortality from lung cancer (51–56). Moreover, $PM_{2.5}$ can worsen bacterial infections such as pneumonia caused by *Streptococcus pneumoniae* by impairing the function of macrophages in the lungs (57). Oxidative and inflammatory damage and increased cellular autophagy and apoptosis have been implicated in the induction of various diseases by $PM_{2.5}$; however, the underlying MoAs remain to be elucidated. Notably, the results of the MP ontology enrichment analysis for $PM_{2.5}$ exhibited strong similarity to those of TBT, highlighting the prominent enrichment of phenotypes related to the immune and hematopoietic systems following $PM_{2.5}$ exposure, consistent with previous research (58). Furthermore, the results of DAR-ChIPEA revealed that the binding of blood cell differentiation-related factors such as Cebpe, Cebpb, Cebpa, Rela, and Spi1 were enriched in the DARs of blood samples closed by $PM_{2.5}$ exposure, i.e., showed binding in the unexposed samples. Among the enriched TFs, the C/EBP family members Cebpe and Cebpb are essential factors for differentiation and functional maintenance of granulocytes and the monocyte–macrophage axis, respectively, whereas Cebpa contributes to the regulation of growth arrest and differentiation of various cell types including progenitor cells in bone marrow (59–61). In addition, Rela is important for erythropoiesis and Spi1 is important for early differentiation of B cells and bone marrow cells (62,63). Taken together, these results suggest that $PM_{2.5}$ inhibits the normal binding of these TFs to the genome, suppressing the differentiation and functional maturation of blood cells and leading to immune dysfunction.

Lead (Pb) has historically been used in a variety of applications, including paints and cosmetic dyes, and remains a common component of water pipes, solder, and other materials. Pb poisoning is typically a chronic condition that may not present with acute symptoms. Regardless of the absence of acute symptoms, Pb poisoning ultimately exhibits irreversible effects such as cognitive impairment, peripheral neuropathy, and progressive renal dysfunction (64–66). Chronic Pb exposure may also induce fatty liver disease, which may be associated with changes in the gut microbiota (67). Consistent with these observations, MP ontology enrichment analysis on liver samples exposed to Pb corroborated the significant enrichment of genes associated with the regulation of liver physiology, as well as lipid homeostasis and metabolism, in the proximity of DARs induced by Pb administration. DAR-ChIPEA further revealed that Nr1d1 (also known as Rev-erbα), an important member of the feedback loop regulating the circadian clock composed of Clock and Arntl (also known as Bmal1) (68), was the most enriched TF in Pb-induced opened DARs. In addition, Ep300 and CREB-binding protein (Crebbp), transcriptional coactivators and members of the Clock/Bmal1 activator complex, were also found to be enriched in the opened DARs. These findings suggest that circadian rhythms in the liver may be disrupted by Pb exposure. Moreover, clock genes have been shown to act on the promoter and introns of adipogenic Ppara (69,70). In the liver, Ppara enhances lipid metabolism and is generally considered to have an inhibitory effect on hepatic steatosis, although some reports suggest that Ppara may promote the pathogenesis of non-alcoholic steatohepatitis (71,72). These findings together with those from MP ontology analysis suggest that Pb induces fatty liver not only through the gut microbiota pathway, but also by disrupting the normal regulation of lipid metabolism by circadian rhythms in the liver. Thus, this serves as a representative illustration of how the proposed DAR-ChIPEA approach can contribute to the formulation of reasonable hypotheses for elucidating the MoAs of environmental chemicals.

**Overall Characteristics and Future Prospects of the DAR-ChIPEA**

When dealing with health hazards caused by air pollution or industrial wastewater, it can be difficult to isolate a single pollution-causing molecule from such complex context. This, it can be difficult to apply approaches that focus on the detailed molecular structure of chemicals to analyze their MoAs, such as docking simulation and supervised learning using molecular structure as features (8–10,73). In this respect, the enrichment analysis-based approaches

presented in this study are advantageous in that they use omics data as input and thus are better suited to analyze complex mixtures. In parallel with the processes used for causative molecule isolation, in the proposed DAR-ChIPEA method, chemically induced DAR information is obtained from experiments involving exposure to polluted air and water and passed through the proposed analysis pipeline, allowing for rapid and rough estimation of the key TFs involved in the MoAs of pollutants. This approach is expected to serve as an efficient complementary approach for investigating the causes of clinically-relevant pollution damage.

We found that the ChIPEA approach, an enrichment analysis method based on actual ChIP-Seq experiments, was more accurate than MotifEA, a numerical analysis method based on DNA binding motifs of TFs, in estimating chemical–disorder associations via key TFs. The moderate-performance MotifEA likely occurred, in part, because the presence of a particular binding motif does not necessarily indicate the actual binding of a TF, and conversely, it is not uncommon for a TF to bind to genomic regions that are distinct from known motifs. Furthermore, binding motifs can vary greatly depending on the tissue or cell type. In comparison, ChIPEA is highly reliable for studying the interaction between TFs and genomic DNA because it extracts TF binding sites in the genome based on a large number of ChIP-Seq experiments reported in the literature in a motif-independent manner (22,36). However, it should be acknowledged that the number of ChIP-Seq experiments currently accumulated varies considerably across different tissues and cell types, and thus ChIPEA may be less effective in situations where the available data are insufficient to generate statistically significant results.

Notably, the pivotal TFs found using our proposed approach may also serve as potential targets for drug discovery aimed at mitigating health hazards posed by environmental pollutants. Therefore, our approach is not only useful for elucidating the MoAs of environmental chemicals, but also for screening candidate compounds for potential pharmaceuticals. Owing to its ability to achieve adequate depth of analysis with as few as 500 cells, ATAC-Seq provides the possibility to perform large-scale analyses of various compounds using small-well plates (74). RNA-Seq may also be employed simultaneously for evaluating the pharmacological effects of the input compounds. In addition to the DBKERO database used in this study, other databases are also available that focus on predicting the

medicinal effects and toxicity of various chemicals by analyzing changes in gene signatures upon chemical perturbation. For example, the LINCS L1000 project and Open TG-GATEs have already obtained extensive transcriptome data on tens of thousands of drugs applied to various cell lines and laboratory animals (75,76). To further clarify the molecular cues that mediate gene expression changes in response to chemical administration, ATAC-Seq data should be incorporated into the compound screening pipeline in these projects, thereby facilitating the effective application of the proposed DAR-ChIPEA method. Furthermore, because ChIPEA is already publicly available through the ChIP-Atlas website in both a user-friendly graphical user interface and a command-based application programming interface, it could be easily applied in various situations to elucidate drug MoAs. Realization of this potential application could help identify TFs that are critical for the expression of drug effects, allowing the formulation of strategies to minimize unexpected risk, especially for compounds with potential adverse effects. In addition, the identification of novel targets will allow the repositioning of drugs that have successfully passed safety testing as therapeutic agents for other diseases. Thus, the proposed DAR-ChIPEA approach could also contribute to lower drug discovery costs and more efficient drug repositioning in the future.

## DATA AND CODE AVAILABILITY

We used freely available datasets; specific version numbers or dates of access or download are indicated as appropriate elsewhere in the Materials and Methods section. Sample codes and all raw results of the TF enrichment analyses (i.e., DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and DAR-ChIPEA) were deposited at https://github.com/zouzhaonan/dar_moa/.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

*Titles for HTML*:

Supplementary Figures S1, S2

Supplementary Tables S1–S9

## AUTHOR CONTRIBUTIONS

Shinya Oki: Conceptualization, Supervision, Data curation, Funding acquisition, Writing–original draft, Writing–review & editing. Zhaonan Zou: Formal analysis, Funding acquisition, Writing–original draft, Writing–review & editing. Yuka Yoshimura: Formal analysis. Yoshihiro Yamanishi: Validation, Writing–original draft, Writing–review & editing.


## ACKNOWLEDGEMENTS

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics, Japan. We thank Prof. Shuh Narumiya and the members of the Department of Drug Discovery Medicine at Kyoto University for providing us with valuable advice.

## FUNDING

This work was supported by Grants-in-Aid for Scientific Research (KAKENHI), Japan Society for the Promotion of Science [22J15229 to Z.Z., 23KF0048 to Z.Z and S.O., 22H02819, 23H04954 to S.O.]; Japan Science and Technology Agency [SPRING JPMJSP2110 to Z.Z., PRESTO JPMJPR1942, NBDC JPMJND2202, and ERATO JPMJER2101 to S.O.]; Japan Agency for Medical Research and Development [PRIME JP22gm6710001, BINDS JP22ama121017 to S.O.], and Kyoto University [MIP and KMS-FUND to Z.Z.]. Funding for open access charge: Japan Science and Technology Agency [JPMJND2202].


## CONFLICT OF INTEREST

The authors declare no conflict of interest. The funders had no role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

## REFERENCES

1. Lee,S.W., Yon,D.K., James,C.C., Lee,S., Koh,H.Y., Sheen,Y.H., Oh,J.W., Han,M.Y. and Sugihara,G. (2019) Short-term effects of multiple outdoor environmental factors on risk of asthma exacerbations: age-stratified time-series analysis. J. Allergy Clin. Immunol., 144, 1542–1550.e1.
2. Kaufman,J.D., Adar,S.D., Barr,R.G., Budoff,M., Burke,G.L., Curl,C.L., Daviglus,M.L., Diez Roux,A.V., Gassett,A.J., Jacobs,D.R.,Jr. et al. (2016) Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. Lancet, 388, 696–704.
3. Yang,B.Y., Qian,Z., Howard,S.W., Vaughn,M.G., Fan,S.J., Liu,K.K. and Dong,G.H. (2018) Global association between ambient air pollution and blood pressure: a systematic review and meta-analysis. Environ. Pollut., 235, 576–588.
4. Zhu,R.X., Nie,X.H., Chen,Y.H., Chen,J., Wu,S.W. and Zhao,L.H. (2020) Relationship between particulate matter ($PM_{2.5}$) and hospitalizations and mortality of chronic obstructive pulmonary disease patients: a meta-analysis. Am. J. Med. Sci., 359, 354–364.
5. Wong,C.M., Tsang,H., Lai,H.K., Thomas,G.N., Lam,K.B., Chan,K.P., Zheng,Q., Ayres,J.G., Lee,S.Y., Lam,T.H. et al. (2016) Cancer mortality risks from long-term exposure to ambient fine particle. Cancer Epidemiol. Biomarkers Prev., 25, 839–845.
6. Gianicolo,E.A.L., Bruni,A., Rosati,E., Sabina,S., Guarino,R., Padolecchia,G., Leo,C., Vigotti,M.A., Andreassi,M.G. and Latini,G. (2012) Congenital anomalies among live births in a polluted area. A ten-year retrospective study. BMC Pregnancy Childbirth, 12, 165.
7. Mahalingaiah,S., Hart,J.E., Laden,F., Farland,L.V., Hewlett,M.M., Chavarro,J., Aschengrau,A. and Missmer,S.A. (2016) Adult air pollution exposure and risk of infertility in the Nurses' Health Study II. Hum. Reprod., 31, 638–647.
8. Bosc,N., Atkinson,F., Felix,E., Gaulton,A., Hersey,A. and Leach,A.R. (2019) Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J. Cheminform., 11, 4.
9. Jacob,L. and Vert,J.P. (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics, 24, 2149–2156.
10. Kolb,P., Ferreira,R.S., Irwin,J.J. and Shoichet,B.K. (2009) Docking and chemoinformatic screens for new ligands and targets. Curr. Opin. Biotechnol., 20, 429–436.
11. Pham,T.H., Qiu,Y., Liu,J., Zimmer,S., O'Neill,E., Xie,L. and Zhang,P. (2022) Chemical-induced

gene expression ranking and its application to pancreatic cancer drug repurposing. Patterns (N. Y.), 3, 100441.

12. Pilarczyk,M., Fazel-Najafabadi,M., Kouril,M., Shamsaei,B., Vasiliauskas,J., Niu,W., Mahi,N., Zhang,L., Clark,N.A., Ren,Y. et al. (2022) Connecting omics signatures and revealing biological mechanisms with iLINCS. Nat. Commun., 13, 4678.
13. Lee,C.W., Kim,S.M., Sa,S., Hong,M., Nam,S.M. and Han,H.W. (2023) Relationship between drug targets and drug-signature networks: a network-based genome-wide landscape. BMC Med. Genomics, 16, 17.
14. GBD DALYs and HALE Collaborators (2018) Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet, 392, 1859–1922.
15. Quenby,S., Gallos,I.D., Dhillon-Smith,R.K., Podesek,M., Stephenson,M.D., Fisher,J., Brosens,J.J., Brewin,J., Ramhorst,R., Lucas,E.S. et al. (2021) Miscarriage matters: the epidemiological, physical, psychological, and economic costs of early pregnancy loss. Lancet, 397, 1658–1667.
16. Go,S., Kurita,H., Matsumoto,K., Hatano,M., Inden,M. and Hozumi,I. (2018) Methylmercury causes epigenetic suppression of the tyrosine hydroxylase gene in an *in vitro* neuronal differentiation model. Biochem. Biophys. Res. Commun., 502, 435–441.
17. Huang,D., Zhang,Y., Qi,Y., Chen,C. and Ji,W. (2008) Global DNA hypomethylation, rather than reactive oxygen species (ROS), a potential facilitator of cadmium-stimulated K562 cell proliferation. Toxicol. Lett., 179, 43–47.
18. van Tilburg,C.M., Milde,T., Witt,R., Ecker,J., Hielscher,T., Seitz,A., Schenk,J.P., Buhl,J.L., Riehl,D., Frühwald,M.C. et al. (2019) Phase I/II intra-patient dose escalation study of vorinostat in children with relapsed solid tumor, lymphoma, or leukemia. Clin. Epigenetics, 11, 188.
19. Bardia,A., Kaklamani,V., Wilks,S., Weise,A., Richards,D., Harb,W., Osborne,C., Wesolowski,R., Karuturi,M., Conkling,P. et al. (2021) Phase I study of elacestrant (RAD1901), a novel selective estrogen receptor degrader, in ER-Positive, HER2-negative advanced breast cancer. J. Clin. Oncol., 39, 1360–1370.
20. Duttke,S.H., Chang,M.W., Heinz,S. and Benner,C. (2019) Identification and dynamic quantification of regulatory elements using total RNA. Genome Res., 29, 1836–1846.

21. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. Nucleic Acids Res., 43, W39–W49.
22. Zou,Z., Iwata,M., Yamanishi,Y. and Oki,S. (2022) Epigenetic landscape of drug responses revealed through large-scale ChIP-seq data analyses. BMC Bioinformatics, 23, 51.
23. Zou,Z., Ohta,T., Miura,F. and Oki,S. (2022) ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res., 50, W175–W182.
24. Suzuki,A., Kawano,S., Mitsuyama,T., Suyama,M., Kanai,Y., Shirahige,K., Sasaki,H., Tokunaga,K., Tsuchihara,K., Sugano,S. et al. (2018) DBTSS/DBKERO for integrated analysis of transcriptional regulation. Nucleic Acids Res., 46, D229–D238.
25. Wang,T., Pehrsson,E.C., Purushotham,D., Li,D., Zhuo,X., Zhang,B., Lawson,H.A., Province,M.A., Krapp,C., Lan,Y. et al. (2018) The NIEHS TaRGET II Consortium and environmental epigenomics. Nat. Biotechnol., 36, 225–227.
26. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res., 48, D845–D855.
27. Davis,A.P., Wiegers,T.C., Johnson,R.J., Sciaky,D., Wiegers,J. and Mattingly,C.J. (2023) Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res., 51, D1257–D1262.
28. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N.,Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. et al. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res., 46, D252–D259.
29. Blake,J.A., Baldarelli,R., Kadin,J.A., Richardson,J.E., Smith,C.L., Bult,C.J. and Mouse Genome Database Group (2021) Mouse Genome Database (MGD): knowledgebase for mouse-human comparative biology. Nucleic Acids Res., 49, D981–D987.
30. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol., 37, 907–915.
31. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30, 923–930.
32. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for

differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140.

33. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842.
34. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of *cis*-regulatory regions. Nat. Biotechnol., 28, 495–501.
35. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9, R137.
36. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep., 19, e46255.
37. Xu,J., Ou,K., Chen,C., Li,B., Guo,J., Zuo,Z. and Wang,C. (2019) Tributyltin exposure disturbs hepatic glucose metabolism in male mice. Toxicology, 425, 152242.
38. Chamorro-Garcia,R., Diaz-Castillo,C., Shoucri,B.M., Käch,H., Leavitt,R., Shioda,T. and Blumberg,B. (2017) Ancestral perinatal obesogen exposure results in a transgenerational thrifty phenotype in mice. Nat. Commun., 8, 2012.
39. Zuo,Z., Chen,S., Wu,T., Zhang,J., Su,Y., Chen,Y. and Wang,C. (2011) Tributyltin causes obesity and hepatic steatosis in male mice. Environ. Toxicol., 26, 79–85.
40. Mendes,A.B.A., Motta,N.A.V., Lima,G.F., Autran,L.J., Brazão,S.C., Magliano,D.C., Sepúlveda-Fragoso,V., Scaramello,C.B.V., Graceli,J.B., Miranda-Alves,L. et al. (2022) Evaluation of the effects produced by subacute tributyltin administration on vascular reactivity of male Wistar rats. Toxicology, 465, 153067.
41. Lehrke,M. and Lazar,M.A. (2005) The many faces of PPARgamma. Cell, 123, 993–999.
42. Li,X., Ycaza,J. and Blumberg,B. (2011) The environmental obesogen tributyltin chloride acts via peroxisome proliferator activated receptor gamma to induce adipogenesis in murine 3T3-L1 preadipocytes. J. Steroid Biochem. Mol. Biol., 127, 9–15.
43. Ueno,S., Kashimoto,T., Susa,N., Asai,T., Kawaguchi,S., Takeda-Homma,S., Terada,Y. and Sugiyama,M. (2009) Reduction in peripheral lymphocytes and thymus atrophy induced by organotin compounds *in vivo*. J. Vet. Med. Sci., 71, 1041–1048.
44. Zhang,Y., Liang,J., Sun,L., Guo,Z. and Xu,L. (2013) Inhibition of PP2A and the consequent

activation of JNK/c-Jun are involved in tributyltin-induced apoptosis in human amnionic cells. Environ. Toxicol., 28, 390–400.

45. Tada-Oikawa,S., Murata,M. and Kato,T. (2010) [Preferential induction of apoptosis in regulatory T cells by tributyltin: possible involvement in the exacerbation of allergic diseases]. Nihon Eiseigaku Zasshi, 65, 530–535.
46. Kato,T., Uchikawa,R., Yamada,M., Arizono,N., Oikawa,S., Kawanishi,S., Nishio,A., Nakase,H. and Kuribayashi,K. (2004) Environmental pollutant tributyltin promotes Th2 polarization and exacerbates airway inflammation. Eur. J. Immunol., 34, 1312–1321.
47. Grabarczyk,P., Przybylski,G.K., Depke,M., Völker,U., Bahr,J., Assmus,K., Bröker,B.M., Walther,R. and Schmidt,C.A. (2007) Inhibition of BCL11B expression leads to apoptosis of malignant but not normal mature T cells. Oncogene, 26, 3797–3810.
48. Ikawa,T., Hirose,S., Masuda,K., Kakugawa,K., Satoh,R., Shibano-Satoh,A., Kominami,R., Katsura,Y. and Kawamoto,H. (2010) An essential developmental checkpoint for production of the T cell lineage. Science, 329, 93–96.
49. Szabo,S.J., Kim,S.T., Costa,G.L., Zhang,X., Fathman,C.G. and Glimcher,L.H. (2000) A novel transcription factor, T-bet, directs Th1 lineage commitment. Cell, 100, 655–669.
50. Zhu,J.,Jankovic,D.,Oler,A.J.,Wei,G.,Sharma,S.,Hu,G.,Guo,L.,Yagi,R.,Yamane,H., Punkosdy,G. et al. (2012) The transcription factor T-bet is induced by multiple pathways and prevents an endogenous Th2 cell program during Th1 cell responses. Immunity, 37, 660–673.
51. Kim,Y., Park,E.H., Ng,C.F.S., Chung,Y., Hashimoto,K., Tashiro,K., Hasunuma,H., Doi,M., Tamura,K., Moriuchi,H. et al. (2021) Respiratory function declines in children with asthma associated with chemical species of fine particulate matter ($PM_{2.5}$) in Nagasaki, Japan. Environ. Health, 20, 110.
52. Chen,P.C., Mou,C.H., Chen,C.W., Hsieh,D.P.H., Tsai,S.P., Wei,C.C. and Sung, F.C. (2022) Roles of ambient temperature and $PM_{2.5}$ on childhood acute bronchitis and bronchiolitis from viral infection. Viruses, 14, 1392.
53. Shi,W., Liu,C., Annesi-Maesano,I., Norback,D., Deng,Q., Huang,C., Qian,H., Zhang,X., Sun,Y., Wang,T. et al. (2021) Ambient $PM_{2.5}$ and its chemical constituents on lifetime-ever pneumonia in Chinese children: a multi-center study. Environ. Int., 146, 106176.
54. He,F., Yanosky,J.D., Fernandez-Mendoza,J., Chinchilli,V.M., Al-Shaar,L., Vgontzas,A.N., Bixler,E.O. and Liao,D. (2022) Acute impact of fine particulate air pollution on cardiac arrhythmias in a population-based sample of adolescents: the Penn State Child Cohort. J.

Am. Heart Assoc., 11, e026370.

55. Chaulin,A.M. and Sergeev,A.K. (2022) The role of fine particles (PM 2.5) in the genesis of atherosclerosis and myocardial damage: emphasis on clinical and epidemiological data, and pathophysiological mechanisms. Cardiol. Res., 13, 268–282.
56. Huang,F., Pan,B., Wu,J., Chen,E. and Chen,L. (2017) Relationship between exposure to PM2.5 and lung cancer incidence and mortality: a meta-analysis. Oncotarget, 8, 43322–43331.
57. Chen,Y.W., Huang,M.Z., Chen,C.L., Kuo,C.Y., Yang,C.Y., Chiang-Ni,C., Chen,Y.Y.M., Hsieh,C.M., Wu,H.Y., Kuo,M.L. et al. (2020) $PM_{2.5}$ impairs macrophage functions to exacerbate pneumococcus-induced pulmonary pathogenesis. Part. Fibre Toxicol., 17, 37.
58. Ge,J., Yang,H., Lu,X., Wang,S., Zhao,Y., Huang,J., Xi,Z., Zhang,L. and Li,R. (2020) Combined exposure to formaldehyde and $PM_{2.5}$: Hematopoietic toxicity and molecular mechanism in mice. Environ. Int., 144, 106050.
59. Yamanaka,R., Barlow,C., Lekstrom-Himes,J., Castilla,L.H., Liu,P.P., Eckhaus,M., Decker,T., Wynshaw-Boris,A. and Xanthopoulos,K.G. (1997) Impaired granulopoiesis, myelodysplasia, and early lethality in CCAAT/enhancer binding protein epsilon-deficient mice. Proc. Natl. Acad. Sci. U. S. A., 94, 13187–13192.
60. Tanaka,T., Akira,S., Yoshida,K., Umemoto,M., Yoneda,Y., Shirafuji,N., Fujiwara,H., Suematsu,S., Yoshida,N. and Kishimoto,T. (1995) Targeted disruption of the NF-IL6 gene discloses its essential role in bacteria killing and tumor cytotoxicity by macrophages. Cell, 80, 353–361.
61. Reckzeh,K., Bereshchenko,O., Mead,A., Rehn,M., Kharazi,S., Jacobsen,S.E., Nerlov,C. and Cammenga,J. (2012) Molecular and cellular effects of oncogene cooperation in a genetically accurate AML mouse model. Leukemia, 26, 1527–1536.
62. Grossmann,M., Metcalf,D., Merryfull,J., Beg,A., Baltimore,D. and Gerondakis,S. (1999) The combined absence of the transcription factors Rel and RelA leads to multiple hemopoietic cell defects. Proc. Natl. Acad. Sci. U. S. A., 96, 11848–11853.
63. McKercher,S.R., Torbett,B.E., Anderson,K.L., Henkel,G.W., Vestal,D.J., Baribault,H., Klemsz,M., Feeney,A.J., Wu,G.E., Paige,C.J. et al. (1996) Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. EMBO J., 15, 5647–5658.
64. Lee,H., Lee,M.W., Warren,J.R. and Ferrie,J. (2022) Childhood lead exposure is associated with lower cognitive functioning at older ages. Sci. Adv., 8, eabn5164.

65. Rubens,O., Logina,I., Kravale,I., Eglîte,M. and Donaghy,M. (2001) Peripheral neuropathy in chronic occupational inorganic lead exposure: a clinical and electrophysiological study. J. Neurol. Neurosurg. Psychiatry, 71, 200–204.
66. Lin,J.L., Tan,D.T., Hsu,K.H. and Yu,C.C. (2001) Environmental lead exposure and progressive renal insufficiency. Arch. Intern. Med., 161, 264–271.
67. Wan,H., Wang,Y., Zhang,H., Zhang,K., Chen,Y., Chen,C., Zhang,W., Xia,F., Wang,N. and Lu,Y. (2022) Chronic lead exposure induces fatty liver disease associated with the variations of gut microbiota. Ecotoxicol. Environ. Saf., 232, 113257.
68. Allada,R. and Bass,J. (2021) Circadian mechanisms in medicine. N. Engl. J. Med., 384, 550–561.
69. Oishi,K., Shirai,H. and Ishida,N. (2005) CLOCK is involved in the circadian transactivation of peroxisome-proliferator-activated receptor alpha (PPARalpha) in mice. Biochem. J., 386, 575–581.
70. Inoue,I., Shinoda,Y., Ikeda,M., Hayashi,K., Kanazawa,K., Nomura,M., Matsunaga,T., Xu,H., Kawai,S., Awata,T. et al. (2005) CLOCK/BMAL1 is involved in lipid metabolism via transactivation of the peroxisome proliferator-activated receptor (PPAR) response element. J. Atheroscler. Thromb., 12, 169–174.
71. Kasano-Camones,C.I., Takizawa,M., Ohshima,N., Saito,C., Iwasaki,W., Nakagawa,Y., Fujitani,Y., Yoshida,R., Saito,Y., Izumi,T. et al. (2023) PPARalpha activation partially drives NAFLD development in liver-specific Hnf4a-null mice. J. Biochem., 173, 393–411.
72. Klaunig,J.E., Babich,M.A., Baetcke,K.P., Cook,J.C., Corton,J.C., David,R.M., DeLuca,J.G., Lai,D.Y., McKee,R.H., Peters,J.M. et al. (2003) PPARalpha agonist-induced rodent tumors: modes of action and human relevance. Crit. Rev. Toxicol., 33, 655–780.
73. Stanton,D.T. (2012) QSAR and QSPR model interpretation using partial least squares (PLS) analysis. Curr. Comput. Aided Drug Des., 8, 107–127.
74. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods, 10, 1213–1218.
75. Igarashi,Y., Nakatsu,N., Yamashita,T., Ono,A., Ohno,Y., Urushidani,T. and Yamada,H. (2015) Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res., 43, D921–D927.
76. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F.,

Tubelli,A.A., Asiedu,J.K. et al. (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 171, 1437–1452 e17.

**FIGURE LEGENDS**

Figure 1. Overview of the proposed DAR-ChIPEA approach. (**A**) Schematic diagram illustrating the difference between our previously reported DEG-ChIPEA method (22) and the proposed DAR-ChIPEA approach. The DEG-ChIPEA method is restricted to analyzing TF binding to the vicinity of gene loci. In contrast, the DAR-ChIPEA approach comprehensively covers TF binding to accessible chromatin regions across the genome including proximal (e.g. promoters) and distal (e.g. enhancers) regulatory elements. The circles surrounded by dotted lines represent the TFs binding to distal regulatory elements that are not detected in the DEG-ChIPEA method. (**B**) To demonstrate the feasibility of DAR-ChIPEA, we identified the TFs enriched at pollutant-induced DARs by analyzing large-scale ChIP-Seq data. Overlaps were evaluated between the DARs (green box) and peak-call data (black lines) of TF-related ChIP-Seq experiments archived in ChIP-Atlas.

Figure 2. Validation of the proposed DAR-ChIPEA approach to predict TF-driven MoAs of *in vitro* chemical perturbation. (**A**) Schematic diagram for TF enrichment analyses using *in vitro* chemical perturbation datasets. DEGs or DARs upon chemical perturbation were detected using data from RNA-Seq or ATAC-Seq, respectively, followed by enrichment analyses to profile TF binding. Methods used for TF enrichment analysis are shown in blue; the proposed method (DAR-ChIPEA) is highlighted in bold. (**B**) Ranking plot of maximum enrichment scores ($-\log_{10}P$-value) for each dosing condition obtained using the DAR-ChIPEA and DEG-ChIPEA methods. (**C**, **D**) Distribution of AUROC (**C**) and AUPR (**D**) scores for each chemical–TF–disorder triadic association predicted using DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and ChIPEA. Orange dots represent mean scores; global scores are noted below the plots. Differences between the methods are presented as *P*-values (two-tailed Wilcoxon rank-sum test) above the bee swarm plots.

Figure 3. Application of the DAR-ChIPEA approach to predict TF-driven modalities of pollutant administration *in vivo*. (**A**) Schematic diagram for TF enrichment analyses using *in vivo* pollutant exposure datasets. DEGs or DARs upon pollutant exposure were detected

using RNA-Seq or ATAC-Seq, respectively, followed by enrichment analyses to profile TF binding. Methods used for TF enrichment analysis are shown in blue, with the proposed DAR-ChIPEA method highlighted in bold. (**B**) Ranking plot of maximum enrichment scores ($-\log_{10}P$-value) for each exposure condition as obtained using the DAR-ChIPEA and DEG-ChIPEA methods. (**C**, **D**) Distribution of AUROC (**C**) and AUPR (**D**) scores for each pollutant–TF–disorder triadic association predicted using DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and ChIPEA. Orange dots represent mean scores; global scores are noted beneath the bee swarm plots. Differences between the methods are presented as *P*-values (two-tailed Wilcoxon rank-sum test) above the bee swarm plots.

Figure 4. Biological interpretations of the predicted pollutant–TF–disorder associations for three representative pollutants. Mouse phenotype ontologies enriched for the genes proximal to pollutant-induced opened DARs are shown in the left panels. Bars representing individual ontology terms are colored based on their parent categories, which were obtained from the Mouse Genome Informatics database. Volcano plots illustrating the results of the proposed DAR-ChIPEA approach are shown in the right panels. Dots indicate individual TFs and are colored based on cell type classes obtained from ChIP-Atlas. Positive and negative values of $\log_2$ fold enrichment (X-axis) indicate TFs enriched in relaxed and condensed chromatin induced by pollutant exposure, respectively. The enrichment score ($-\log_{10}P$-value; Y-axis) indicates the statistical significance of ChIPEA.

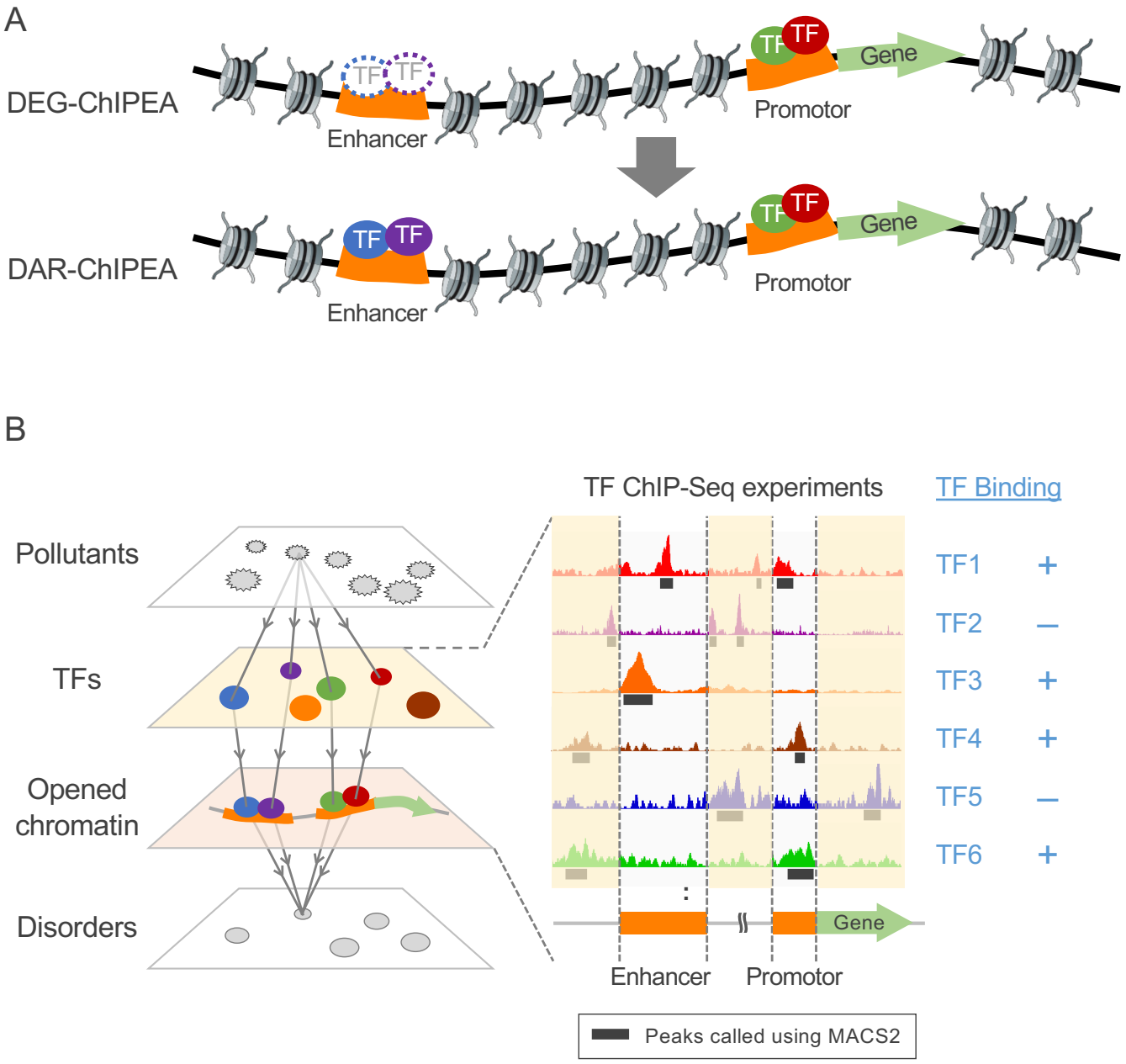
A
DEG-ChIPEA
TF
TF
Enhancer
TF
TF
Gene
Promotor
DAR-ChIPEA
TF
TF
Enhancer
TF
TF
Gene
Promotor
B
Pollutants
TFs
Opened chromatin
Disorders
TF ChIP-Seq experiments
TF Binding
TF1 +
TF2 –
TF3 +
TF4 +
TF5 –
TF6 +
Gene
Enhancer
Promotor
Peaks called using MACS2

Figure 1

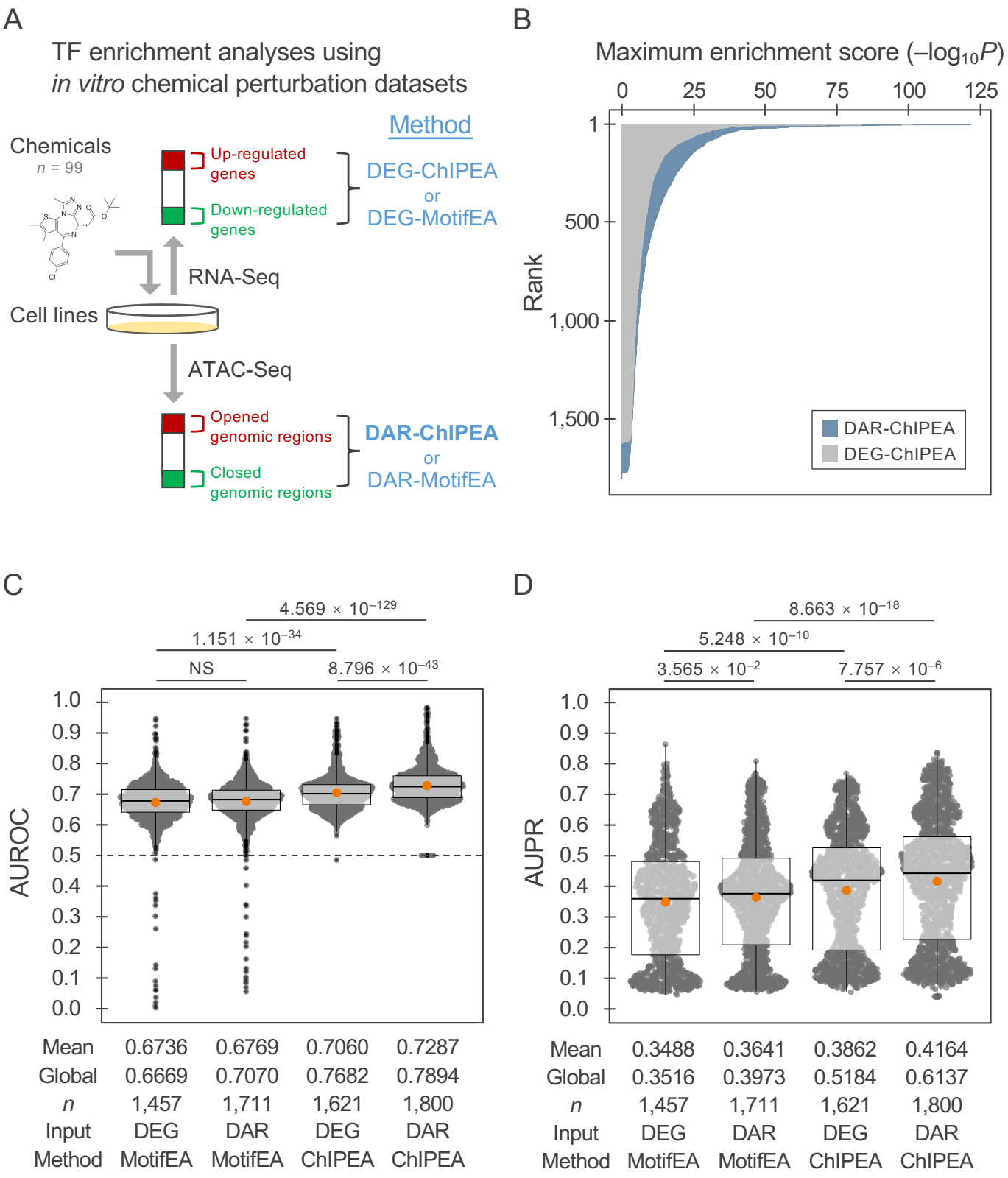
A
TF enrichment analyses using
in vitro chemical perturbation datasets
Method
Chemicals
n = 99
Up-regulated genes
Down-regulated genes
DEG-ChIPEA
or
DEG-MotifEA
RNA-Seq
Cell lines
ATAC-Seq
Opened genomic regions
Closed genomic regions
DAR-ChIPEA
or
DAR-MotifEA
B
Maximum enrichment score (–log10P)
0
25
50
75
100
125
Rank
1
500
1,000
1,500
DAR-ChIPEA
DEG-ChIPEA
C
4.569 × 10−129
1.151 × 10−34
NS
8.796 × 10−43
AUROC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Mean 0.6736 0.6769 0.7060 0.7287
Global 0.6669 0.7070 0.7682 0.7894
n 1,457 1,711 1,621 1,800
Input DEG DAR DEG DAR
Method MotifEA MotifEA ChIPEA ChIPEA
D
8.663 × 10−18
5.248 × 10−10
3.565 × 10−2
7.757 × 10−6
AUPR
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Mean 0.3488 0.3641 0.3862 0.4164
Global 0.3516 0.3973 0.5184 0.6137
n 1,457 1,711 1,621 1,800
Input DEG DAR DEG DAR
Method MotifEA MotifEA ChIPEA ChIPEA

Figure 2

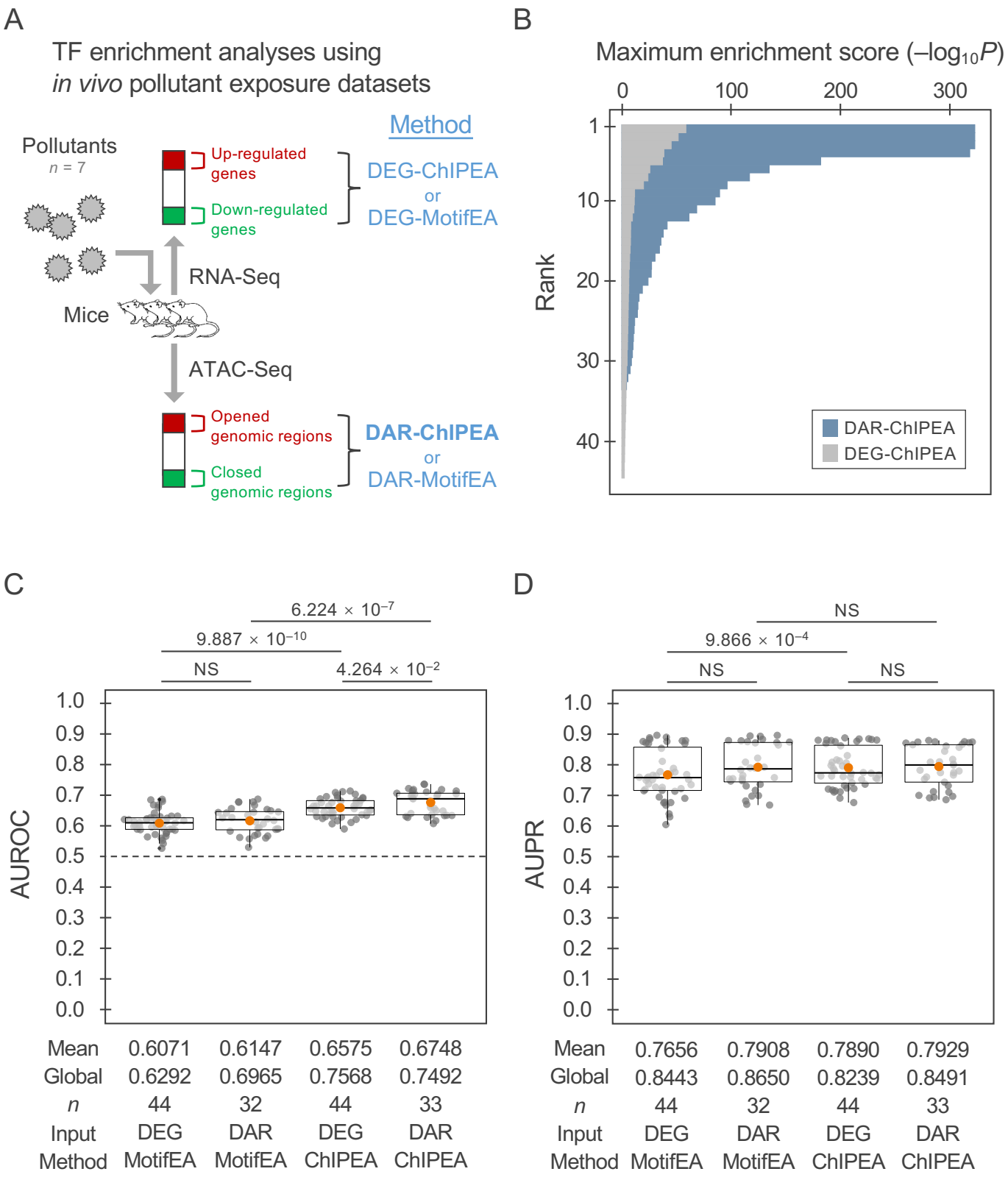
A
TF enrichment analyses using
in vivo pollutant exposure datasets
Pollutants
n = 7
Mice
RNA-Seq
ATAC-Seq
Up-regulated genes
Down-regulated genes
Opened genomic regions
Closed genomic regions
Method
DEG-ChIPEA
or
DEG-MotifEA
DAR-ChIPEA
or
DAR-MotifEA
B
Maximum enrichment score (–log10P)
Rank
DAR-ChIPEA
DEG-ChIPEA
C
AUROC
6.224 × 10−7
9.887 × 10−10
NS
4.264 × 10−2
Mean 0.6071 0.6147 0.6575 0.6748
Global 0.6292 0.6965 0.7568 0.7492
n 44 32 44 33
Input DEG DAR DEG DAR
Method MotifEA MotifEA ChIPEA ChIPEA
D
AUPR
NS
9.866 × 10−4
NS
NS
Mean 0.7656 0.7908 0.7890 0.7929
Global 0.8443 0.8650 0.8239 0.8491
n 44 32 44 33
Input DEG DAR DEG DAR
Method MotifEA MotifEA ChIPEA ChIPEA

Figure 3

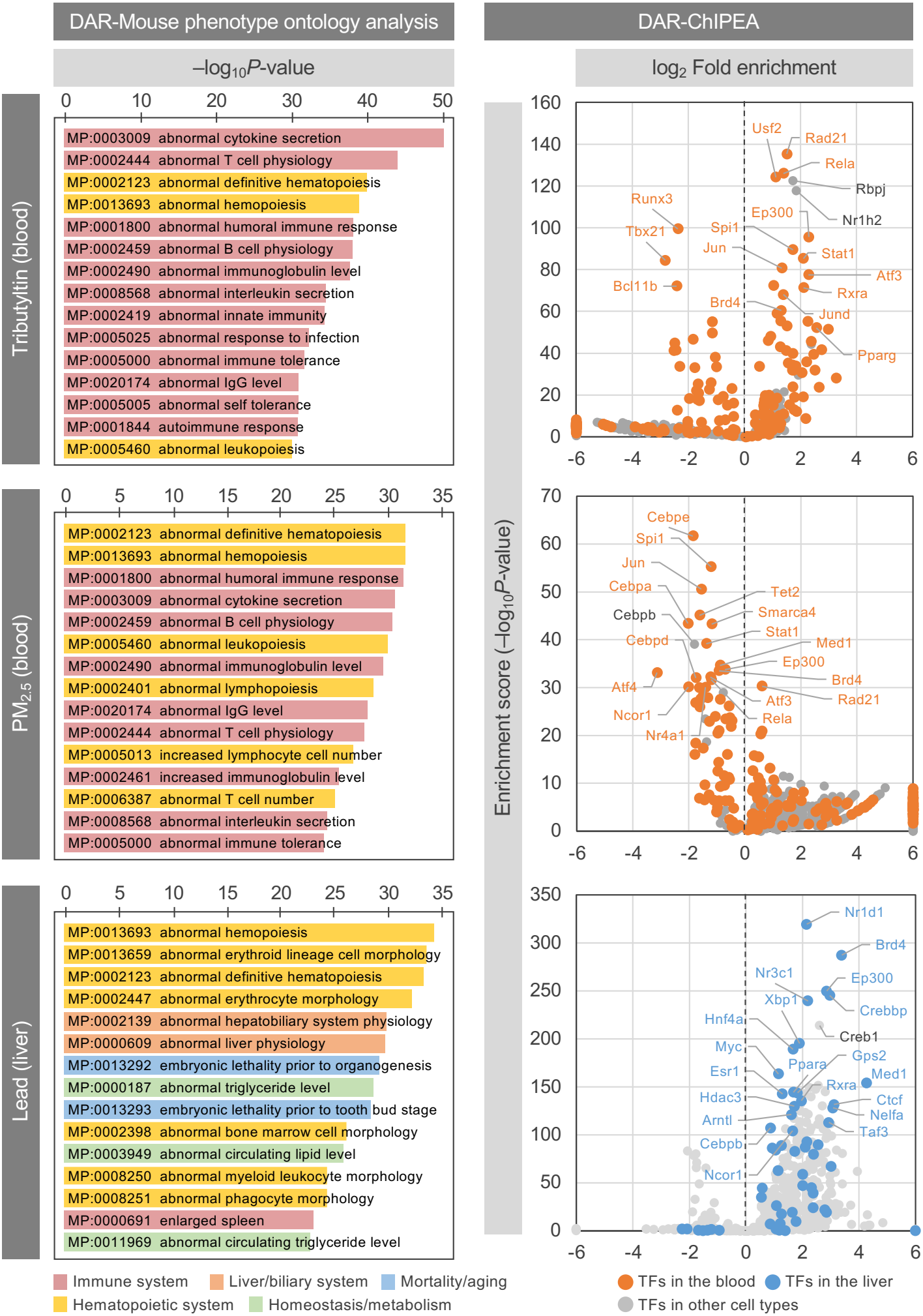
DAR-Mouse phenotype ontology analysis
–log10P-value
Tributyltin (blood)
MP:0003009 abnormal cytokine secretion
MP:0002444 abnormal T cell physiology
MP:0002123 abnormal definitive hematopoiesis
MP:0013693 abnormal hemopoiesis
MP:0001800 abnormal humoral immune response
MP:0002459 abnormal B cell physiology
MP:0002490 abnormal immunoglobulin level
MP:0008568 abnormal interleukin secretion
MP:0002419 abnormal innate immunity
MP:0005025 abnormal response to infection
MP:0005000 abnormal immune tolerance
MP:0020174 abnormal IgG level
MP:0005005 abnormal self tolerance
MP:0001844 autoimmune response
MP:0005460 abnormal leukopoiesis
PM2.5 (blood)
MP:0002123 abnormal definitive hematopoiesis
MP:0013693 abnormal hemopoiesis
MP:0001800 abnormal humoral immune response
MP:0003009 abnormal cytokine secretion
MP:0002459 abnormal B cell physiology
MP:0005460 abnormal leukopoiesis
MP:0002490 abnormal immunoglobulin level
MP:0002401 abnormal lymphopoiesis
MP:0020174 abnormal IgG level
MP:0002444 abnormal T cell physiology
MP:0005013 increased lymphocyte cell number
MP:0002461 increased immunoglobulin level
MP:0006387 abnormal T cell number
MP:0008568 abnormal interleukin secretion
MP:0005000 abnormal immune tolerance
Lead (liver)
MP:0013693 abnormal hemopoiesis
MP:0013659 abnormal erythroid lineage cell morphology
MP:0002123 abnormal definitive hematopoiesis
MP:0002447 abnormal erythrocyte morphology
MP:0002139 abnormal hepatobiliary system physiology
MP:0000609 abnormal liver physiology
MP:0013292 embryonic lethality prior to organogenesis
MP:0000187 abnormal triglyceride level
MP:0013293 embryonic lethality prior to tooth bud stage
MP:0002398 abnormal bone marrow cell morphology
MP:0003949 abnormal circulating lipid level
MP:0008250 abnormal myeloid leukocyte morphology
MP:0008251 abnormal phagocyte morphology
MP:0000691 enlarged spleen
MP:0011969 abnormal circulating triglyceride level
Immune system
Liver/biliary system
Mortality/aging
Hematopoietic system
Homeostasis/metabolism
DAR-ChIPEA
log2 Fold enrichment
Enrichment score (–log10P-value)
Usf2
Rad21
Rela
Rbpj
Nr1h2
Runx3
Ep300
Tbx21
Spi1
Jun
Stat1
Atf3
Bcl11b
Rxra
Brd4
Jund
Pparg
Cebpe
Spi1
Jun
Cebpa
Tet2
Cebpb
Smarca4
Stat1
Cebpd
Med1
Ep300
Brd4
Atf4
Atf3
Rad21
Ncor1
Rela
Nr4a1
Nr1d1
Brd4
Nr3c1
Ep300
Xbp1
Crebbp
Hnf4a
Creb1
Myc
Gps2
Ppara
Esr1
Med1
Rxra
Hdac3
Ctcf
Nelfa
Arntl
Taf3
Cebpb
Ncor1
TFs in the blood
TFs in the liver
TFs in other cell types

Figure 4