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

E-mail:

Zhaonan Zou	zou.zhaonan.56e@st.kyoto-u.ac.jp
Yuka Yoshimura	y-yoshimura1205@outlook.com
Yoshihiro Yamanishi	yamanishi@i.nagoya-u.ac.jp
Shinya Oki	oki.shinya.3w@kyoto-u.ac.jp

*Correspondence: oki.shinya.3w@kyoto-u.ac.jp; Tel: +81-75-366-7479

ABSTRACT

Background: Despite well-documented effects on human health, the action modes of environmental pollutants are incompletely understood. Although transcriptome-based approaches are widely used to predict associations between chemicals and disorders, the molecular cues regulating pollutant-derived gene expression changes remain unclear. Therefore, we developed a data-mining approach, termed "DAR-ChIPEA," to identify transcription factors (TFs) playing pivotal roles in the action modes of pollutants.

Methods: Large-scale public ChIP-Seq data (human, n = 15,155; mouse, n = 13,156) were used to predict TFs that are enriched in the pollutant-induced differentially accessible genomic regions (DARs) obtained from epigenome analyses (ATAC-Seq). The resultant pollutant–TF matrices were then cross-referenced to a repository of TF–disorder associations to account for pollutant modes of action. We subsequently evaluated the performance of the proposed method using a chemical perturbation dataset to compare the outputs of the DAR-ChIPEA and our previously developed differentially expressed gene (DEG)-ChIPEA methods using pollutant-induced DEGs as input. We then adopted the proposed method to predict disease-associated mechanisms triggered by pollutants.

Results: The proposed approach outperformed other methods using the area under the receiver operating characteristic curve score. The mean score of the proposed DAR-ChIPEA was significantly higher than that of our previously described DEG-ChIPEA (0.7287 vs. 0.7060; $Q = 5.278 \times 10^{-42}$; two-tailed Wilcoxon rank-sum test). The proposed approach further predicted TF-driven modes of action upon pollutant exposure, indicating that (1) TFs regulating Th1/2 cell homeostasis are integral in the pathophysiology of tributyltin-induced allergic disorders; (2) 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 (3) lead induces fatty liver by disrupting the normal regulation of lipid metabolism by altering hepatic circadian rhythms.

Conclusion: Highlighting genome-wide chromatin change upon pollutant exposure to elucidate the epigenetic landscape of pollutant responses outperformed our previously described method that focuses on gene-adjacent domains only. Our approach has the potential to reveal pivotal TFs that mediate deleterious effects of pollutants, thereby facilitating the development of strategies to mitigate damage from environmental pollution.

Keywords:

ChIP-Seq, ATAC-seq, Transcription factor, Differentially accessible genomic region, Epigenetic landscape, Action modes of environmental pollutants

53 **BACKGROUND**

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

73 Bioactive chemicals, including some pollutants, can target a wide variety of molecules,
74 such as membrane transport proteins, ion channels, and metabolic enzymes. In particular,
75 research on steroid receptors, which are activated by female (estrogens) and male
76 (androgens) hormones, is actively pursued owing to the reported effects of endocrine-

77 disrupting chemicals, also known as environmental hormones, on reproduction and
78 development [14, 15]. In addition, heavy metals, including mercury and cadmium, which have
79 been associated with pollution-related disorders such as Minamata disease and Itai-itai
80 disease, may induce epigenetic changes as a component of their disease mechanisms [16,
81 17]. Furthermore, both intranuclear steroid receptors and enzymes that regulate epigenomic
82 states, such as DNA methylation and histone modifications, can function as transcription
83 factors (TFs) to directly control gene expression. Accordingly, certain pharmaceuticals such as
84 estrogen receptor inhibitors and histone deacetylase inhibitors have been shown to exert
85 anticancer effects by directly regulating TF function [18, 19]. Together, these observations
86 suggest that TFs that collectively regulate the on/off switches of genes whose expression is
87 significantly altered by exposure to environmental chemicals may function as direct targets
88 or mediators of chemical MoAs.
89 To identify TFs whose binding is enriched to a given set of genes or genomic regions,
90 binding motif-based methods (MotifEA) such as HOMER and MEME have widely been used
91 [20, 21]. This approach predicts TF binding by extracting shared short nucleotide sequences,
92 commonly referred to as "motifs", from multiple input genomic regions. These motifs are
93 then compared to known TF–DNA binding motifs by comparative analysis. As an alternative
94 method for identifying key TFs in chemical MoAs, we recently developed an approach that is
95 independent of TF–DNA binding motifs, which we termed the DEG-ChIPEA method [22]. This
96 method is based on the ChIP-Atlas database, which extracts actual TF binding sites on the
97 reference genome by curating and integrating tens of thousands of previously published
98 ChIP-Seq experimental data examining TF binding to genomic DNA [23]. The DEG-ChIPEA
99 method involves inputting chemically induced DEGs and performing enrichment analysis
100 based on all ChIP-Seq experiments obtained from ChIP-Atlas (i.e., ChIP-Seq-based

101 enrichment analysis; ChIPEA) to identify TFs whose binding is significantly enriched in the
102 vicinity of the input DEGs. These identified TFs can be considered promising candidate
103 pivotal factors that mediate the expression of therapeutic or adverse chemical effects.
104 However, despite the generally high accuracy of DEG-ChIPEA, a potential limitation is that
105 only TFs in the immediate vicinity of the transcription start site (transcription start site ± 5 kb)
106 of the input chemically induced genes are analyzed. Thus, potential contributions to gene
107 expression by TFs that bind to distal transcriptional regulatory regions, such as enhancers
108 and topologically associating domain boundaries located far from the gene locus, are not
109 taken into account.

110 In this study, we aimed to gain further insights into the MoAs of chemicals, particularly
111 those of environmental pollutants, with regard to health effects. To achieve this, we propose
112 a novel approach, termed "DAR-ChIPEA", that considers TF binding events throughout the
113 genome, including both proximal and distal transcriptional regulatory regions of genes, in
114 contrast to the limited purview of the DEG-ChIPEA method. The DAR-ChIPEA method
115 involves extracting differentially accessible regions (DARs), which are genomic regions with
116 significantly different chromatin accessibility prior to and following chemical exposure, using
117 data from environmental chemical perturbation epigenome experiments, such as ATAC-Seq
118 datasets. These DARs are then used as input for ChIPEA to generate a chemical–TF
119 association matrix, which allows the identification of TFs whose binding is significantly
120 enriched in the input DARs. In addition, we constructed a chemical–TF–disorder triadic
121 association by incorporating known TF–disease associations into the resulting chemical–TF
122 matrix. Our approach outperformed methods based on TF–DNA binding motifs, as well as
123 the previously developed DEG-ChIPEA method [22], which focuses only on TFs that bind in
124 the vicinity of chemically induced DEGs.

125

126 **METHODS**

127 **Chemical- or pollutant-perturbed transcriptome and epigenome data**

128 To identify genomic regions with significant differences in chromatin accessibility (DARs) in
129 response to chemical or pollutant exposure as determined through epigenome perturbation
130 experiments, we used ATAC-Seq datasets retrieved from the DBKERO (*n* = 2,078;
131 DRA006903–DRA006930) [24, 25] and TaRGET (*n* = 383; GSE146508) databases [24, 25]. In
132 addition, RNA-Seq datasets corresponding to the ATAC-Seq datasets were made available in
133 these databases and used to identify DEGs induced by chemical or pollutant exposure as
134 input to the DEG-ChIPEA or DEG-MotifEA analyses (Additional file 1: Supplementary Fig. S1;
135 DBKERO, *n* = 2,012, DRA006875–DRA006902; TaRGET, *n* = 468, GSE146508). DBKERO
136 harbors perturbation data of 99 chemicals in human cell lines, whereas TaRGET curates *in-*
137 *vivo* exposure data for seven pollutants in blood and liver samples from mice. All data used
138 in this analysis were archived in the National Center for Biotechnology Information Sequence
139 Read Archive (NCBI SRA) [26] and are associated with accession IDs as listed in
140 Supplementary Tables S1–S4 (Additional file 2).

141

142 **Genome-wide TF binding and chromatin accessibility experimental data from ChIP-**
143 **Atlas**

144 Genome-wide TF binding site information was obtained from the ChIP-Atlas database [23],
145 which comprehensively integrates nearly all publicly available ChIP-Seq data. Peak call data
146 in BED format (MACS2 [27]; *Q*-value $<1 \times 10^{-10}$) were used in this study. Data from ChIP-Seq
147 experiments examining binding of 1,796 and 866 TFs on the human (GRCh38/hg38; hereafter
148 referred to as hg38) and mouse genomes (GRCm38/mm10; hereafter referred to as mm10),

149 respectively, were used in this study (hg38, *n* = 15,155; mm10, *n* = 13,156). In addition, the
150 information related to accessible chromatin used in this analysis to identify DARs induced by
151 chemical or pollutant exposure was also obtained from ChIP-Atlas, which also provides peak
152 call data for ATAC-Seq.
153
154 **TF–disease associations**
155 Association data between TFs and diseases were obtained from DisGeNET (v 7.0) [28], which
156 is a database of genetic and genomic information associated with human diseases. DisGeNET
157 integrates this information from other expert-curated repositories and PubMed literature.
158 Only manually curated data that were labeled with referenced PubMed IDs were retrieved,
159 resulting in a total of 12,857 associations between TFs and diseases, involving 1,004 TFs and
160 3,205 diseases.
161
162 **Known chemical–disease associations as standard data**
163 Manually curated chemical–disease association data were acquired from the Comparative
164 Toxicogenomics Database (CTD) [29]. In total, we incorporated 28,821 chemical–disease
165 associations involving 69 chemicals (DBKERO) and seven pollutants (TaRGET), and 2,150
166 diseases in this study.
167
168 **Identification of chemical- or pollutant-induced DEGs**
169 After obtaining the SRA accession IDs of the chemical or pollutant perturbation RNA-Seq
170 data from DBKERO and TaRGET, the raw sequencing data archived in NCBI SRA were
171 downloaded and decoded into FASTQ format using the "fastq-dump" command from the
172 SRA Toolkit (v3.0.0) [30]. Quality control was then performed using the "fastq_quality_filter"

173 command from the FASTX-Toolkit (v0.0.13) [31]. Alignment to the reference genomes
174 (DBKERO, hg38; TaRGET, mm10) was performed using HISAT2 (v2.1.0) [32], and expression
175 quantification was performed using featureCounts (v2.0.3) [33]. All commands were run with
176 default settings and parameters. DEGs were detected using the R Package "edgeR" (v3.40.2)
177 [34]. Information on the statistical significance used for identification of DEGs is detailed in
178 the Statistical Analyses section.
179

180 **Identification of chemical- or pollutant-induced DARs**

181 Referring to the SRA accession IDs for the chemical- or pollutant-perturbed ATAC-Seq data
182 obtained from DBKERO and TaRGET, the peak call data (in BED format) for the corresponding
183 ATAC-Seq experiments were acquired from ChIP-Atlas. Peak calling was performed using
184 MACS2 [27] in ChIP-Atlas; only peaks with a *Q*-value of $<1 \times 10^{-10}$ were considered as
185 statistically significant accessible chromatin regions. Next, common open chromatin regions
186 across all replicates for each chemical or pollutant perturbation condition were extracted
187 using BEDtools (v2.23.0) [35]. Differential analysis was then performed to identify DARs
188 induced by the chemicals or pollutants. Note that the term "differential analysis" in this
189 context refers to the identification of genomic regions with no overlap between groups with
190 and without chemical or pollutant exposure. In this study, DARs that exhibited relaxed or
191 condensed chromatin upon chemical or pollutant exposure are referred to as "opened" or
192 "closed" DARs, respectively.
193

194 **TF enrichment analysis fully incorporating public ChIP-Seq experiments (ChIPEA)**

195 Details of the ChIPEA procedure using the application programming interface are provided
196 online [22, 36]. To exploit the extensive ChIP-Seq data in ChIP-Atlas [23], we performed

197 ChIPEA on profile TFs whose binding sites were enriched around genomic regions or genes
198 of interest. In particular, using the genomic coordinates of opened and closed DARs, or the
199 gene symbols of up- and down-regulated genes by a query chemical or pollutant as input,
200 we first counted the number of intersects between the input data and peak-call data
201 (MACS2; Q-value $<1 \times 10^{-10}$) of all TF-related experiments archived in the ChIP-Atlas
202 database using the "intersectBed" command of BEDTools (v2.23.0) [35]. Note that 5 kb
203 upstream and downstream of the transcription start site were added for each gene when
204 chemically induced DEGs were queried. Enrichment scores ($-\log_{10}Q$) were then calculated
205 using the two-tailed Fisher's exact probability test to test whether the two datasets (opened
206 and closed DARs, or up- and down-regulated genes) overlapped with the ChIP-Seq peak-call
207 data with equivalent proportions before performing the Benjamini-Hochberg procedure for
208 multiple testing correction; fold enrichment values (opened DARs/closed DARs) were
209 concurrently obtained. If a chemical– or pollutant–TF association was indicated by multiple
210 ChIP-Seq experiments, only the highest enrichment score was adopted. Parameters used in
211 application programming interface-based ChIPEA were as follows: "genome=hg38 (DBKERO)
212 or mm10 (TaRGET); antigenClass=TFs and others; cellClass=All cell types; threshold=100."
213 Sample codes for DEG-ChIPEA and DAR-ChIPEA are provided as indicated in the Availability
214 of Data and Materials statement below.
215

216 **TF enrichment analysis focusing on TF–DNA binding motifs (MotifEA)**

217 In this method, TF binding to chemical- or pollutant-induced DARs and DEGs was estimated
218 using MotifEA with STREME and TOMTOM in MEME Suite (v5.4.1) [21], using default settings
219 termed DAR-MotifEA and DEG-MotifEA, respectively. STREME was first run to identify 1,000
220 sequence motifs from queried multiple genomic regions, i.e., chemical- or pollutant-induced

221 DARs or DEGs. Subsequently, TFs with enriched binding to the input genomic regions or
222 genes were estimated by matching the identified motifs with known TF binding motifs
223 obtained from HOCOMOCO v11 [37] using TOMTOM. Note that 5 kb upstream and
224 downstream of the transcription start site were added for each gene when chemically
225 induced DEGs were queried. Sample codes for DEG-MotifEA and DAR-MotifEA are provided
226 as indicated in the Availability of Data and Materials statement below.
227

228 **Establishment of chemical– or pollutant–TF–disorder triadic associations**

229 Using TFs enriched in chemical- or pollutant-induced DEGs and DARs as mediators, the TF–
230 disease associations retrieved from DisGeNET [28] were integrated with the chemical– or
231 pollutant–TF matrices obtained as outcomes of MotifEA or ChIPEA, thereby establishing
232 chemical– or pollutant–TF–disorder triadic associations. The enrichment scores obtained via
233 ChIPEA or MotifEA were continuously used as a quantitative measure to assess the degree of
234 newly constructed chemical– or pollutant–disorder associations.
235

236 **Calculation of global area under the receiver operating characteristic curve (AUROC)**
237 **and area under the precision recall curve (AUPR) scores**

238 All predicted chemical– or pollutant–disorder associations were first reorganized into a single
239 matrix containing enrichment scores with dimensions of $m \times n$, where m and n represent the
240 total number of chemical or pollutant perturbation conditions and the number of disorders,
241 respectively (Additional file 1: Supplementary Fig. S1). Among various TFs linking a specific
242 chemical–disease pair, we selectively adopted the most statistically significant TF, i.e., the one
243 with the highest enrichment score. We then utilized the maximum enrichment score as a
244 statistical measure to represent the association between this chemical and the disease. After

245 referencing known chemical–disease associations from CTD [29], chemical-wise ROC (which
246 plots the true positive rate against the false positive rate) and PR (which plots the precision
247 [positive predictive value] against the recall [sensitivity]) curves were generated (R Package
248 "ROCR" v1.0-11) [38]. The performance results were summarized into a global AUROC score,
249 in which 1 represents perfect classification and 0.5 represents random classification, and an
250 AUPR score, in which 1 represents perfect inference and the proportion of positive examples
251 in the standard data corresponds to random inference.
252
253 **Mouse phenotype ontology enrichment analysis**
254 GO enrichment analysis on opened DARs induced by pollutant exposure was performed
255 using the Genomic Regions Enrichment of Annotations Tool (GREAT) (v4.0.4) [39]. Default
256 parameters were used in this analysis; namely, "Species assembly: mm10; Association rule:
257 Basal+extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension,
258 curated regulatory domains included." Ontologies categorized as "Mouse Phenotype" were
259 then extracted. Parent category information for individual ontology terms was obtained with
260 reference to the Mouse Genome Informatics database [40].
261
262 **Statistical Analyses**
263 Statistical analysis was performed using R (v4.1.3). Unpaired two-tailed Wilcoxon rank-sum
264 tests were used to compare AUROC and AUPR scores between two groups, Bonferroni
265 correction was performed to presented adjusted *P*-values (*Q*-values). Fisher's exact test was
266 used for 2 × 2 contingency analysis, and the Benjamini-Hochberg procedure was used for
267 multiple testing correction. Sample sizes and *Q*-values are shown in the figures, and a *Q*-
268 value < 0.05 was considered statistically significant. NS denotes no significance. Chemical-

269 and pollutant-induced DEGs were identified using the R package edgeR (v3.40.2), and a *P*-
270 value <0.05 (DBKERO) or an adjusted *P*-value <0.1 (TaRGET) with an absolute value of $\log_2$
271 fold change >1 was recognized as statistically significant.
272
273
274 **RESULTS**
275 **Study Design: Elucidation of Chemical MoAs Focusing on Chemically Reorganized**
276 **Chromatin Accessibility**
277 To expand upon our prior approach (DEG-ChIPEA) [22], here, we present a novel analytical
278 approach to elucidate chemical MoAs focusing on chemically reorganized chromatin
279 assembly. Specifically, we utilize chemically induced DARs as input to cover the full spectrum
280 of TF binding to both proximal and distal transcription regulatory regions (e.g., promoters
281 and enhancers; Fig. 1a). Our approach identifies TFs that integratively regulate chemical- or
282 pollutant-induced DARs by comprehensively analyzing tens of thousands of publicly
283 available ChIP-Seq data obtained from the ChIP-Atlas database (Fig. 1b) [23]. In addition, we
284 integrated the predicted chemical–TF associations with disease-associated TF data obtained
285 from the DisGeNET database [28]. The resulting chemical–TF–disorder triadic associations
286 were then statistically validated using a known chemical–disease association dataset from
287 CTD as standard data (Additional file 1: Supplementary Fig. S1) [29].
288
289 **Validation of the Proposed DAR-ChIPEA Approach to Predict TF-Driven Actions for *In***
290 ***Vitro* Chemical Perturbation**
291 Because TaRGET [25] contains information on only a few types of pollutants, it is not
292 sufficient to statistically evaluate the reliability of DAR-ChIPEA. To validate the proposed

293 DAR-ChIPEA approach to predict TF-driven actions for *in vitro* chemical perturbation, we first
294 evaluated the prediction accuracy of our method using the DBKERO database [24], which
295 contains data on hundreds of chemical perturbations to cell cultures, providing correlated
296 RNA-Seq and ATAC-Seq datasets. After retrieving pre-processed peak call datasets of ATAC-
297 Seq from cell lines with and without chemical treatment in the DBKERO database from ChIP-
298 Atlas, we performed a comparative analysis to detect DARs opened or closed by chemical
299 perturbation (Additional file 1: Supplementary Fig. S2). Next, we fully exploited almost all
300 publicly available human TF ChIP-Seq data in ChIP-Atlas to identify TFs enriched for binding
301 to the extracted DARs using ChIPEA (DAR-ChIPEA). This allowed us to generate chemical–TF
302 associations for each drug treatment condition. Furthermore, to evaluate the pivotality of TFs
303 in drug MoAs, chemical–TF–disorder associations were constructed by assigning TF-related
304 disease information obtained from DisGeNET to the resulting chemical–TF association matrix.
305 For comparison, in the chemical–TF association generation step, we also used chemically
306 induced DEGs as input to ChIPEA (DEG-ChIPEA), as previously reported (Additional file 1:
307 Supplementary Fig. S2) [22]. For the same purpose, a more conventional and commonly used
308 TF enrichment analysis method focusing on TF–DNA binding motifs (MotifEA) was also
309 performed with DEGs or DARs as input (hereafter referred to as DEG-MotifEA and DAR-
310 MotifEA; Fig. 2a).

311 We observed that the statistically significant maximum enrichment scores ($-\log_{10}Q$-value)
312 obtained from the proposed DAR-ChIPEA method were globally higher than those from our
313 previously developed DEG-ChIPEA method (Fig. 2b; mean maximum enrichment score =
314 10.25 and 8.139, respectively; $P = 3.013 \times 10^{-4}$; two-tailed Wilcoxon rank-sum test),
315 indicating that DAR-ChIPEA has a wider dynamic range than DEG-ChIPEA. Thus, DAR-ChIPEA

316 offers a significant advantage in discriminating associated and non-associated TFs with
317 regard to individual chemicals.
318 We then examined whether the chemical–TF–disorder triadic associations derived from
319 DAR-ChIPEA that exhibited higher statistical significance were congruent with existing
320 knowledge. The distribution of AUROC and AUPR scores for each dosing condition,
321 generated using manually curated chemical–disease association data obtained from CTD as
322 reference data (1,457, 1,711, 1,621, and 1,800 conditions for DEG-MotifEA, DAR-MotifEA,
323 DEG-ChIPEA, and DAR-ChIPEA, respectively), was depicted using bee swarm plots (Fig. 2c, d;
324 Additional file 2: Supplementary Table S5). The mean AUROC and AUPR scores of the DAR-
325 ChIPEA approach across all dosage conditions were 0.7287 and 0.4164, respectively. These
326 scores were significantly higher than those of other methods performed for comparison
327 (DEG-MotifEA: mean AUROC 0.6736; mean AUPR 0.3488, DAR-MotifEA: mean AUROC
328 0.6769; mean AUPR 0.3641, and DEG-ChIPEA: mean AUROC 0.7060; mean AUPR 0.3862;
329 $P<0.05$ for all; two-tailed Wilcoxon rank-sum test).
330 In addition, because the maximum enrichment scores varied drastically depending on the
331 given dosing conditions, we were concerned that the discrimination of items was limited
332 when comparing TFs within a dosing condition with low maximum enrichment scores.
333 Therefore, we also calculated "global" statistics using an inter-chemical merged enrichment
334 score vector to emphasize the significance of the actual values of the enrichment scores
335 (detailed in the Methods section). The global AUROC and AUPR scores of the proposed
336 approach were 0.7894 and 0.6137, respectively, significantly outperforming those of other
337 methods ($P<0.05$; two-tailed Wilcoxon rank-sum test; Fig. 2c, d). These results suggest that
338 the ChIPEA method using chemically induced DARs as input is a powerful approach for
339 identifying pivotal TFs, elucidating key factors for specific chemical MoAs and predicting

340 disorders or phenotypes associated with chemical treatments. The use of this approach may
341 ultimately lead to increased predictive power compared to that of the existing DEG-ChIPEA
342 and MotifEA-based computational methods.
343
344 **Application of the DAR-ChIPEA Approach to Predict TF-Driven MoAs upon Pollutant**
345 **Administration *In Vivo***
346 We next employed the DAR-ChIPEA method to predict disorders caused by pollutants and
347 TF-driven MoAs upon pollutant administration *in vivo*. After retrieving ATAC-Seq peak call
348 datasets obtained with and without pollutant exposure from the TaRGET database available
349 in ChIP-Atlas, we identified opened and closed DARs induced by various pollutants, as
350 described in the Methods section (Additional file 1: Supplementary Fig. S2). We used publicly
351 available mouse TF ChIP-Seq data in ChIP-Atlas to generate pollutant–TF associations using
352 the proposed DAR-ChIPEA approach. We simultaneously performed analyses using the other
353 methods for comparison, namely, DEG-ChIPEA, DEG-MotifEA, and DAR-MotifEA (Fig. 3a). The
354 resulting pollutant–TF matrices were then used to generate pollutant–TF–disorder
355 associations, focusing on pivotal TFs, with reference to TF–disease associations from the
356 DisGeNET database. We evaluated the accuracy of the predicted pollutant–disorder
357 associations using ROC and PR curves using known chemical-disease association datasets
358 from CTD as a standard.
359 The maximum enrichment scores attained using the proposed DAR-ChIPEA approach
360 were globally higher than those achieved with our previously developed DEG-ChIPEA
361 method, similar to the results obtained using *in-vitro* input data, further supporting the wider
362 dynamic range for discrimination of associated TFs for individual pollutants (Fig. 3b; mean

363 maximum enrichment score = 76.77 and 11.57, respectively; $P = 8.421 \times 10^{-8}$; two-tailed
364 Wilcoxon rank-sum test).
365 In addition, we utilized a large set of publicly available Bisulfite-Seq experimental data
366 (mm10; n = 34,793) integrated into ChIP-Atlas to assess the DNA methylation status within
367 the DARs induced by representative pollutants. The results revealed that opened and closed
368 DARs were enriched by either hyper- or hypo-methylated regions, suggesting DNA
369 methylation change within pollutant-induced DARs (Additional file 1: Supplementary Fig. S3).
370 The distribution of AUROC and AUPR scores for each exposure condition (40, 28, 40, and
371 21 conditions for DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and DAR-ChIPEA, respectively)
372 was visualized using bee swarm plots (Fig. 3c, d; Additional file 2: Supplementary Table S6).
373 The mean AUROC of the proposed DAR-ChIPEA approach across exposure conditions was
374 0.6748, significantly higher than those of the other methods (DEG-MotifEA, 0.6071; DAR-
375 MotifEA, 0.6147; DEG-ChIPEA, 0.6575; $P<0.05$; two-tailed Wilcoxon rank-sum test). However,
376 the DAR-ChIPEA AUPR score did not significantly ($P>0.05$) exceed those of the other
377 methods (DEG-MotifEA, 0.7656; DAR-MotifEA, 0.7908; DEG-ChIPEA, 0.7890; proposed: DAR-
378 ChIPEA, 0.7929). The global AUROC and global AUPR scores of DAR-ChIPEA were 0.7492 and
379 0.8491, respectively. These results suggest that the DAR-ChIPEA method performed
380 satisfactorily in terms of AUROC; the absence of significant differences regarding AUPR may
381 be attributed to the limited number of datasets analyzed.
382
383 **Biological Interpretation of the Predicted Pollutant–TF–Disorder Associations**
384 In the proposed DAR-ChIPEA method, high AUROC scores were obtained for tributyltin (TBT)
385 (blood samples), $PM_{2.5}$ (blood samples), and lead (liver samples; AUROC = 0.7028, 0.7030,
386 and 0.7349, respectively; Additional file 2: Supplementary Table S4, S6; highlighted with bold

387 red font), which motivated us to evaluate the results of putative pivotal TFs involved in the
388 MoAs of these pollutants in further detail, particularly with regard to the biological
389 interpretation of the predicted pollutant–TF–disorder associations (Fig. 4; Additional file 2:
390 Supplementary Tables S7–S9).

391 The organotin compound TBT is widely utilized in various industrial applications such as
392 pesticides, paints, preservatives, and heat stabilizers for plastics. Nevertheless, TBT has been
393 associated with a potentially increased risk of a number of lifestyle-related diseases,
394 including obesity, fatty liver, cardiovascular disease, and abnormal glucose metabolism [41–
395 44]. We initially conducted ontology enrichment analysis to discern any common nature of
396 genes located proximal to opened DARs of blood samples upon TBT exposure using
397 mammalian phenotype (MP) ontology. Notably, we found that the enriched gene sets were
398 prominently manifested in phenotypic features of the immune system and hematopoietic
399 system (upper left panel in Fig. 4). To further understand the transcriptional regulatory
400 mechanisms involved in the response to TBT exposure, we then performed DAR-ChIPEA
401 (upper right panel in Fig. 4). The genes encoding the top-20 enriched TFs exhibited robust
402 expression in both control and TBT-exposed samples, and the same applied for $PM_{2.5}$ and
403 lead, as detailed below (Additional file 1: Supplementary Fig. S4). The results of DAR-ChIPEA
404 showed that the opening of specific DARs susceptible to TBT exposure could enhance the
405 enriched binding of Pparg, a master regulator of adipogenesis [45]. This is consistent with
406 the *in-vitro* findings that TBT stimulates lipid accumulation by activating Pparg [46].

407 In addition, TBT can induce apoptosis in diverse tissues and cell categories, such as
408 human peripheral blood lymphocytes and human amnion cells, and high-dose exposure to
409 TBT has been reported to induce T lymphocyte apoptosis and thymic atrophy [47–49].
410 Furthermore, TBT suppresses the differentiation of naive $CD4^+$ cells into Th1 cells, while

411 stimulating differentiation into Th2 cells, thereby causing allergic disorders through the
412 overproduction of Th2 cells, although the MoAs remain incompletely understood [50]. Our
413 results showed that TBT exposure led to the enrichment of TFs such as Rela, Stat1, and Rbpj
414 in opened DARs, consistent with the pro-inflammatory nature of TBT. Conversely, the binding
415 of Bcl11b and Tbx21 was markedly enriched in closed DARs. BCL11B has apoptosis-inhibitory
416 properties and is involved in maintaining T cell differentiation, as supported by the reduced
417 numbers of T cells observed in *Bcl11b*-deficient mice [51, 52]. Notably, TBX21 also plays a
418 key role in Th1 cell differentiation, with *Tbx21*-deficient naive Th cells exhibiting distorted
419 Th2 cell polarization, a phenotype somewhat similar to that observed upon TBT exposure
420 [53, 54]. These findings suggest that Bcl11b and Tbx21 may contribute to TBT-induced
421 disruption of the Th1/Th2 balance.

422 $PM_{2.5}$ is fine particulate matter suspended in the atmosphere with a diameter of 2.5 µm
423 or less, which has the ability to transport pollutants in the environment. Exposure to $PM_{2.5}$ is
424 associated with respiratory diseases such as asthma, bronchitis, and pneumonia, as well as
425 heart diseases such as arrhythmias and atherosclerosis, and significantly increases mortality
426 from lung cancer [55–60]. Moreover, $PM_{2.5}$ can worsen bacterial infections such as
427 pneumonia caused by *Streptococcus pneumoniae* by impairing the function of macrophages
428 in the lungs [61]. Oxidative and inflammatory damage and increased cellular autophagy and
429 apoptosis have been implicated in the induction of various diseases by $PM_{2.5}$; however, the
430 underlying MoAs remain to be elucidated. Notably, the results of the MP ontology
431 enrichment analysis for $PM_{2.5}$ exhibited strong similarity to those of TBT, highlighting the
432 prominent enrichment of phenotypes related to the immune and hematopoietic systems
433 following $PM_{2.5}$ exposure, consistent with previous research (middle left panel in Fig. 4) [62].
434 Furthermore, the results of DAR-ChIPEA revealed that the binding of blood cell

435 differentiation-related factors such as Cebpe, Cebpb, Cebpa, Rela, and Spi1 was enriched in
436 the DARs of blood samples closed by $PM_{2.5}$ exposure, i.e., showed binding in the unexposed
437 samples (middle right panel in Fig. 4). Among the enriched TFs, the C/EBP family members
438 Cebpe and Cebpb are essential factors for differentiation and functional maintenance of
439 granulocytes and the monocyte–macrophage axis, respectively, whereas Cebpa contributes
440 to the regulation of growth arrest and differentiation of various cell types, including
441 progenitor cells in bone marrow [63–65]. In addition, Rela is important to erythropoiesis, and
442 Spi1 is important to early differentiation of B cells and bone marrow cells [66, 67].
443 Lead (Pb) has historically been used in a variety of applications, including paints and
444 cosmetic dyes, and remains a common component of water pipes, solder, and other
445 materials. Pb poisoning is typically a chronic condition that may not present with acute
446 symptoms. Regardless of the absence of acute symptoms, Pb poisoning ultimately exhibits
447 irreversible effects such as cognitive impairment, peripheral neuropathy, and progressive
448 renal dysfunction [68–70]. Chronic Pb exposure may also induce fatty liver disease, which
449 may be associated with changes in the gut microbiota [71]. Consistent with these
450 observations, MP ontology enrichment analysis on liver samples exposed to Pb corroborated
451 the significant enrichment of genes associated with the regulation of liver physiology, as well
452 as lipid homeostasis and metabolism, in the proximity of DARs induced by Pb administration
453 (lower left panel in Fig. 4). DAR-ChIPEA further revealed that Nr1d1 (also known as Rev-
454 erbα), an important member of the feedback loop regulating the circadian clock composed
455 of Clock and Arntl (also known as Bmal1) [72], was the most enriched TF in Pb-induced
456 opened DARs. In addition, Ep300 and CREB-binding protein (Crebbp), transcriptional
457 coactivators and members of the Clock/Bmal1 activator complex, were also found to be
458 enriched in opened DARs (lower right panel in Fig. 4). These findings suggest that circadian

459 rhythms in the liver may be disrupted by Pb exposure. Moreover, clock genes have been
460 shown to act on the promoter and introns of adipogenic Ppara [73, 74]. In the liver, Ppara
461 enhances lipid metabolism and is generally considered to have an inhibitory effect on
462 hepatic steatosis, although some reports suggest that Ppara may promote the pathogenesis
463 of non-alcoholic steatohepatitis [75, 76]. These findings, together with those from MP
464 ontology analysis, suggest that Pb induces fatty liver not only through the gut microbiota
465 pathway, but also by disrupting the normal regulation of lipid metabolism by circadian
466 rhythms in the liver.
467

468 **DISCUSSION**

469 In this study, we used the ATAC-Seq dataset, derived from samples perturbed by
470 environmental chemicals, from public databases to identify pivotal TFs involved in chemical
471 MoAs. We extracted DARs and applied ChIPEA, taking full advantage of large-scale ChIP-Seq
472 experiments, to identify TFs that were enriched for binding to these genomic regions. In
473 addition, we integrated the TF-associated disease dataset with the putative chemical–TF
474 associations to estimate key TFs mediating chemical-induced disorder induction and in turn
475 understand chemical MoAs. The proposed DAR-ChIPEA approach outperformed other
476 existing computational methods in terms of accuracy.

477 In general, when cells or tissues are exposed to chemical substances, a series of events
478 occur, initialized by pioneer factors, which are TFs that modulate chromatin accessibility.
479 Subsequently, various cooperative TFs, often acting in complexes (e.g., Arntl/Ep300/Cebp in
480 Fig. 4), bind to the newly opened or closed genome regions, causing alterations in the
481 expression of genes regulated by these TFs. Moreover, the DNA methylation status
482 undergoes modifications, contributing to the fixation of gene expression alteration. This

483 entire cascade ultimately leads to phenotypic changes induced by chemical substances. The
484 proposed method can provide insights into identifying key DNA-binding proteins involved in
485 the MoAs of chemicals, including both pioneer factors and cooperative TFs. In addition, the
486 public Bisulfite-Seq dataset in our previously developed ChIP-Atlas database can also be
487 utilized to infer DNA methylation status change upon chemical exposure (Supplementary
488 figure S2).

489 When dealing with health hazards caused by air pollution or industrial wastewater, it can
490 be difficult to isolate a single pollution-causing molecule from such complex contexts.
491 Application of approaches that focus on the detailed molecular structure of chemicals to
492 analyze their MoAs, such as docking simulation and supervised learning using molecular
493 structure as features, can thus be challenging [8–10, 77]. In this respect, the enrichment
494 analysis-based approaches presented in this study are advantageous in that they use omics
495 data as input and thus are better suited to analyze complex mixtures. In parallel with the
496 processes used for causative molecule isolation, in the proposed DAR-ChIPEA method,
497 chemically induced DAR information is obtained from experiments involving exposure to
498 polluted air and water and passed through the proposed analysis pipeline, allowing for rapid
499 and rough estimation of the key TFs involved in the MoAs of pollutants. This approach is
500 expected to serve as an efficient complementary approach for investigating the causes of
501 clinically-relevant pollution damage.

502 We found that the ChIPEA approach, an enrichment analysis method based on actual
503 ChIP-Seq experiments, was more accurate than MotifEA, a numerical analysis method based
504 on DNA binding motifs of TFs, in estimating chemical–disorder associations via key TFs. The
505 moderate performance of MotifEA likely occurred, in part, because the presence of a
506 particular binding motif does not necessarily indicate the actual binding of a TF, and

507 conversely, it is not uncommon for a TF to bind to genomic regions that are distinct from
508 known motifs. Furthermore, binding motifs can vary greatly depending on the tissue or cell
509 type. In comparison, ChIPEA is highly reliable for studying the interaction between TFs and
510 genomic DNA because it extracts TF binding sites in the genome based on a large number of
511 ChIP-Seq experiments reported in the literature in a motif-independent manner [22, 36].
512 However, it should be acknowledged that the number of ChIP-Seq data currently available
513 varies considerably across different tissues and cell types, and thus, ChIPEA may be less
514 effective in situations in which available data are insufficient to generate statistically
515 significant results.

516 Notably, the pivotal TFs found using our proposed approach may also serve as potential
517 targets for drug discovery aimed at mitigating health hazards posed by environmental
518 pollutants. Therefore, our approach is not only useful for elucidating the MoAs of
519 environmental chemicals, but also for screening candidate compounds for potential
520 pharmaceuticals. Owing to its ability to achieve an adequate depth of analysis with as few as
521 500 cells, ATAC-Seq provides the possibility to perform large-scale analyses of various
522 compounds using small-well plates [78]. RNA-Seq may also be employed simultaneously to
523 evaluate the pharmacological effects of the input compounds. In addition to the DBKERO
524 database used in this study, other databases are also available that focus on predicting the
525 medicinal effects and toxicity of various chemicals by analyzing changes in gene signatures
526 upon chemical perturbation. For example, the LINCS L1000 project and Open TG-GATEs have
527 obtained extensive transcriptome data on tens of thousands of drugs applied to various cell
528 lines and laboratory animals [79, 80]. To further clarify the molecular cues that mediate gene
529 expression changes in response to chemical administration, ATAC-Seq data should be
530 incorporated into the compound screening pipeline in these projects, thereby facilitating the

531 effective application of the proposed DAR-ChIPEA method. Furthermore, because ChIPEA is
532 already publicly available through the ChIP-Atlas website in both a user-friendly graphical
533 user interface and a command-based application programming interface, it could be easily
534 applied in various situations to elucidate drug MoAs. Realization of this potential application
535 could help identify TFs that are critical for the expression of drug effects, allowing the
536 formulation of strategies to minimize unexpected risk, especially for compounds with
537 potential adverse effects. In addition, identification of novel targets will allow the
538 repositioning of drugs that have successfully passed safety testing as therapeutic agents for
539 other diseases. Thus, the proposed DAR-ChIPEA approach could also contribute to lower
540 drug discovery costs and a more efficient drug repositioning in the future.
541
542
543 **CONCLUSIONS**
544 In this paper, we introduced a computational approach for highlighting genome-wide
545 chromatin change upon pollutant exposure to elucidate the epigenetic landscape of
546 pollutant responses, which outperformed other methods focusing on gene-adjacent
547 domains only. The proposed method can be adopted to identify pivotal TFs involved in the
548 MoAs of pollutants from an epigenetic perspective, thereby facilitating the development of
549 strategies to mitigate environmental pollution damage. The proposed DAR-ChIPEA approach
550 can also be used to further understand the MoAs of candidate drug compounds as well as to
551 discover unexpected therapeutic and side effects of approved drugs to reduce costs during
552 drug discovery.
553
554

555 **DECLARATIONS**

556 **Ethics approval and consent to participate**

557 Not applicable.

558 **Consent for publication**

559 Not applicable.

560 **Availability of data and materials**

561 We used freely available datasets; specific version numbers or dates for accessed or
562 downloaded datasets are indicated as appropriate in the reference list. Sample codes and all
563 raw results of the TF enrichment analyses (i.e., DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and
564 DAR-ChIPEA) were deposited at https://github.com/zouzhaonan/dar_moa/.

565 **Competing interests**

566 The authors declare that they have no competing interests. The funders had no role in the
567 design of the study, the collection, analysis, and interpretation of data, or in writing the
568 manuscript.

569 **Funding**

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

577 **Authors' contributions**

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

582 **Acknowledgements**

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

586

587

588

589 **REFERENCES**

590 1. Lee SW, Yon DK, James CC, Lee S, Koh HY, Sheen YH et al. Short-term effects of multiple
591 outdoor environmental factors on risk of asthma exacerbations: Age-stratified time-series
592 analysis. J Allergy Clin Immunol 2019;144(6):1542–1550 e1541.
593 https://doi.org/10.1016/j.jaci.2019.08.037.
594 2. Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL et al. Association between air
595 pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-
596 Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. Lancet
597 2016;388(10045):696–704. https://doi.org/10.1016/S0140-6736(16)00378-0.
598 3. Yang BY, Qian Z, Howard SW, Vaughn MG, Fan SJ, Liu KK et al. Global association between
599 ambient air pollution and blood pressure: A systematic review and meta-analysis. Environ Pollut
600 2018;235:576–588. https://doi.org/10.1016/j.envpol.2018.01.001.
601 4. Zhu RX, Nie XH, Chen YH, Chen J, Wu SW, Zhao LH. Relationship Between Particulate Matter
602 (PM(2.5)) and Hospitalizations and Mortality of Chronic Obstructive Pulmonary Disease

603 Patients: A Meta-Analysis. Am J Med Sci 2020;359(6):354–364.
604 https://doi.org/10.1016/j.amjms.2020.03.016.
605 5. Wong CM, Tsang H, Lai HK, Thomas GN, Lam KB, Chan KP et al. Cancer Mortality Risks from
606 Long-term Exposure to Ambient Fine Particle. Cancer Epidemiol Biomarkers Prev
607 2016;25(5):839–845. https://doi.org/10.1158/1055-9965.EPI-15-0626.
608 6. Gianicolo EA, Bruni A, Rosati E, Sabina S, Guarino R, Padolecchia G et al. Congenital anomalies
609 among live births in a polluted area. A ten-year retrospective study. BMC Pregnancy Childbirth
610 2012;12:165. https://doi.org/10.1186/1471-2393-12-165.
611 7. Mahalingaiah S, Hart JE, Laden F, Farland LV, Hewlett MM, Chavarro J et al. Adult air pollution
612 exposure and risk of infertility in the Nurses' Health Study II. Hum Reprod 2016;31(3):638–647.
613 https://doi.org/10.1093/humrep/dev330.
614 8. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and
615 conformal prediction methods and their applications in drug discovery. J Cheminform
616 2019;11(1):4. https://doi.org/10.1186/s13321-018-0325-4.
617 9. Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach.
618 Bioinformatics 2008;24(19):2149–2156. https://doi.org/10.1093/bioinformatics/btn409.
619 10. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. Docking and chemoinformatic screens for new ligands
620 and targets. Curr Opin Biotechnol 2009;20(4):429–436.
621 https://doi.org/10.1016/j.copbio.2009.08.003.
622 11. Pham TH, Qiu Y, Liu J, Zimmer S, O'Neill E, Xie L et al. Chemical-induced gene expression ranking
623 and its application to pancreatic cancer drug repurposing. Patterns (N Y) 2022;3(4):100441.
624 https://doi.org/10.1016/j.patter.2022.100441.
625 12. Pilarczyk M, Fazel-Najafabadi M, Kouril M, Shamsaei B, Vasiliauskas J, Niu W et al. Connecting
626 omics signatures and revealing biological mechanisms with iLINCS. Nat Commun
627 2022;13(1):4678. https://doi.org/10.1038/s41467-022-32205-3.

628 13. Lee CW, Kim SM, Sa S, Hong M, Nam SM, Han HW. Relationship between drug targets and drug-
629 signature networks: a network-based genome-wide landscape. BMC Med Genomics
630 2023;16(1):17. https://doi.org/10.1186/s12920-023-01444-8.
631 14. DALYs GBD, Collaborators H. Global, regional, and national disability-adjusted life-years (DALYs)
632 for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and
633 territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.
634 Lancet 2018;392(10159):1859-1922. https://doi.org/10.1016/S0140-6736(18)32335-3.
635 15. Quenby S, Gallos ID, Dhillon-Smith RK, Podesek M, Stephenson MD, Fisher J et al. Miscarriage
636 matters: the epidemiological, physical, psychological, and economic costs of early pregnancy
637 loss. Lancet 2021;397(10285):1658-1667. https://doi.org/10.1016/S0140-6736(21)00682-6.
638 16. Go S, Kurita H, Matsumoto K, Hatano M, Inden M, Hozumi I. Methylmercury causes epigenetic
639 suppression of the tyrosine hydroxylase gene in an in vitro neuronal differentiation model.
640 Biochem Biophys Res Commun 2018;502(4):435-441.
641 https://doi.org/10.1016/j.bbrc.2018.05.162.
642 17. Huang D, Zhang Y, Qi Y, Chen C, Ji W. Global DNA hypomethylation, rather than reactive oxygen
643 species (ROS), a potential facilitator of cadmium-stimulated K562 cell proliferation. Toxicol Lett
644 2008;179(1):43-47. https://doi.org/10.1016/j.toxlet.2008.03.018.
645 18. van Tilburg CM, Milde T, Witt R, Ecker J, Hielscher T, Seitz A et al. Phase I/II intra-patient dose
646 escalation study of vorinostat in children with relapsed solid tumor, lymphoma, or leukemia.
647 Clin Epigenetics 2019;11(1):188. https://doi.org/10.1186/s13148-019-0775-1.
648 19. Bardia A, Kaklamani V, Wilks S, Weise A, Richards D, Harb W et al. Phase I Study of Elacestrant
649 (RAD1901), a Novel Selective Estrogen Receptor Degrader, in ER-Positive, HER2-Negative
650 Advanced Breast Cancer. J Clin Oncol 2021;39(12):1360-1370.
651 https://doi.org/10.1200/JCO.20.02272.

652 20. Duttke SH, Chang MW, Heinz S, Benner C. Identification and dynamic quantification of
653 regulatory elements using total RNA. Genome Res 2019;29(11):1836-1846.
654 https://doi.org/10.1101/gr.253492.119.
655 21. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res
656 2015;43(W1):W39-49. https://doi.org/10.1093/nar/gkv416.
657 22. Zou Z, Iwata M, Yamanishi Y, Oki S. Epigenetic landscape of drug responses revealed through
658 large-scale ChIP-seq data analyses. BMC Bioinformatics 2022;23(1):51.
659 https://doi.org/10.1186/s12859-022-04571-8.
660 23. Zou Z, Ohta T, Miura F, Oki S. ChIP-Atlas 2021 update: a data-mining suite for exploring
661 epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic
662 Acids Research 2022;50(W1):W175–W182. https://doi.org/10.1093/nar/gkac199.
663 24. Suzuki A, Kawano S, Mitsuyama T, Suyama M, Kanai Y, Shirahige K et al. DBTSS/DBKERO for
664 integrated analysis of transcriptional regulation. Nucleic Acids Res 2018;46(D1):D229-D238.
665 https://doi.org/10.1093/nar/gkx1001.
666 25. Wang T, Pehrsson EC, Purushotham D, Li D, Zhuo X, Zhang B et al. The NIEHS TaRGET II
667 Consortium and environmental epigenomics. Nat Biotechnol 2018;36(3):225-227.
668 https://doi.org/10.1038/nbt.4099.
669 26. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The
670 sequence read archive. Nucleic Acids Res 2011;39(Database issue):D19-21.
671 https://doi.org/10.1093/nar/gkq1019.
672 27. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE et al. Model-based analysis of
673 ChIP-Seq (MACS). Genome biology 2008;9(9):R137. https://doi.org/10.1186/GB-2008-9-9-R137.
674 28. Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, Ronzano F, Centeno E, Sanz F et al. The DisGeNET
675 knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48(D1):D845–
676 855. https://doi.org/10.1093/nar/gkz1021.

677 29. Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative
678 Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res 2023;51(D1):D1257-D1262.
679 https://doi.org/10.1093/nar/gkac833.
680 30. SRA ToolKit v3.0.0. https://github.com/ncbi/sra-tools. Accessed 31 March, 2022.
681 31. FASTX-Toolkit v0.0.13. http://hannonlab.cshl.edu/fastx_toolkit/. Accessed 31 March 2022.
682 32. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and
683 genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37(8):907-915.
684 https://doi.org/10.1038/s41587-019-0201-4.
685 33. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning
686 sequence reads to genomic features. Bioinformatics 2014;30(7):923-930.
687 https://doi.org/10.1093/bioinformatics/btt656.
688 34. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
689 expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139-140.
690 https://doi.org/10.1093/bioinformatics/btp616.
691 35. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
692 Bioinformatics 2010;26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.
693 36. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y et al. ChIP-Atlas: a data-mining suite
694 powered by full integration of public ChIP-seq data. EMBO Rep 2018;19(12):e46255.
695 https://doi.org/10.15252/embr.201846255.
696 37. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI et al.
697 HOCOMOCO: towards a complete collection of transcription factor binding models for human
698 and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res 2018;46(D1):D252-D259.
699 https://doi.org/10.1093/nar/gkx1106.
700 38. ROCR v1.0-11. .

701 39. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB et al. GREAT improves functional
702 interpretation of cis-regulatory regions. Nat Biotechnol 2010;28(5):495-501.
703 https://doi.org/10.1038/nbt.1630.
704 40. Blake JA, Baldarelli R, Kadin JA, Richardson JE, Smith CL, Bult CJ et al. Mouse Genome Database
705 (MGD): Knowledgebase for mouse-human comparative biology. Nucleic Acids Res
706 2021;49(D1):D981-D987. https://doi.org/10.1093/nar/gkaa1083.
707 41. Xu J, Ou K, Chen C, Li B, Guo J, Zuo Z et al. Tributyltin exposure disturbs hepatic glucose
708 metabolism in male mice. Toxicology 2019;425:152242.
709 https://doi.org/10.1016/j.tox.2019.152242.
710 42. Chamorro-Garcia R, Diaz-Castillo C, Shoucri BM, Kach H, Leavitt R, Shioda T et al. Ancestral
711 perinatal obesogen exposure results in a transgenerational thrifty phenotype in mice. Nat
712 Commun 2017;8(1):2012. https://doi.org/10.1038/s41467-017-01944-z.
713 43. Zuo Z, Chen S, Wu T, Zhang J, Su Y, Chen Y et al. Tributyltin causes obesity and hepatic steatosis
714 in male mice. Environ Toxicol 2011;26(1):79-85. https://doi.org/10.1002/tox.20531.
715 44. Mendes ABA, Motta NAV, Lima GF, Autran LJ, Brazao SC, Magliano DC et al. Evaluation of the
716 effects produced by subacute tributyltin administration on vascular reactivity of male wistar
717 rats. Toxicology 2022;465:153067. https://doi.org/10.1016/j.tox.2021.153067.
718 45. Lehrke M, Lazar MA. The many faces of PPARgamma. Cell 2005;123(6):993-999.
719 https://doi.org/10.1016/j.cell.2005.11.026.
720 46. Li X, Ycaza J, Blumberg B. The environmental obesogen tributyltin chloride acts via peroxisome
721 proliferator activated receptor gamma to induce adipogenesis in murine 3T3-L1 preadipocytes.
722 J Steroid Biochem Mol Biol 2011;127(1-2):9-15. https://doi.org/10.1016/j.jsbmb.2011.03.012.
723 47. Ueno S, Kashimoto T, Susa N, Asai T, Kawaguchi S, Takeda-Homma S et al. Reduction in
724 peripheral lymphocytes and thymus atrophy induced by organotin compounds in vivo. J Vet
725 Med Sci 2009;71(8):1041-1048. https://doi.org/10.1292/jvms.71.1041.

726 48. Zhang Y, Liang J, Sun L, Guo Z, Xu L. Inhibition of PP2A and the consequent activation of JNK/c-
727 Jun are involved in tributyltin-induced apoptosis in human amnionic cells. Environ Toxicol
728 2013;28(7):390-400. https://doi.org/10.1002/tox.20730.
729 49. Tada-Oikawa S, Murata M, Kato T. [Preferential induction of apoptosis in regulatory T cells by
730 tributyltin: possible involvement in the exacerbation of allergic diseases]. Nihon Eiseigaku
731 Zasshi 2010;65(4):530-535. https://doi.org/10.1265/jjh.65.530.
732 50. Kato T, Uchikawa R, Yamada M, Arizono N, Oikawa S, Kawanishi S et al. Environmental pollutant
733 tributyltin promotes Th2 polarization and exacerbates airway inflammation. Eur J Immunol
734 2004;34(5):1312-1321. https://doi.org/10.1002/eji.200324667.
735 51. Grabarczyk P, Przybylski GK, Depke M, Volker U, Bahr J, Assmus K et al. Inhibition of BCL11B
736 expression leads to apoptosis of malignant but not normal mature T cells. Oncogene
737 2007;26(26):3797-3810. https://doi.org/10.1038/sj.onc.1210152.
738 52. Ikawa T, Hirose S, Masuda K, Kakugawa K, Satoh R, Shibano-Satoh A et al. An essential
739 developmental checkpoint for production of the T cell lineage. Science 2010;329(5987):93-96.
740 https://doi.org/10.1126/science.1188995.
741 53. Szabo SJ, Kim ST, Costa GL, Zhang X, Fathman CG, Glimcher LH. A novel transcription factor, T-
742 bet, directs Th1 lineage commitment. Cell 2000;100(6):655-669. https://doi.org/10.1016/s0092-
743 8674(00)80702-3.
744 54. Zhu J, Jankovic D, Oler AJ, Wei G, Sharma S, Hu G et al. The transcription factor T-bet is induced
745 by multiple pathways and prevents an endogenous Th2 cell program during Th1 cell responses.
746 Immunity 2012;37(4):660-673. https://doi.org/10.1016/j.immuni.2012.09.007.
747 55. Kim Y, Park EH, Ng CFS, Chung Y, Hashimoto K, Tashiro K et al. Respiratory function declines in
748 children with asthma associated with chemical species of fine particulate matter (PM(2.5)) in
749 Nagasaki, Japan. Environ Health 2021;20(1):110. https://doi.org/10.1186/s12940-021-00796-x.

750 56. Chen PC, Mou CH, Chen CW, Hsieh DPH, Tsai SP, Wei CC et al. Roles of Ambient Temperature
751 and PM(2.5) on Childhood Acute Bronchitis and Bronchiolitis from Viral Infection. Viruses
752 2022;14(9). https://doi.org/10.3390/v14091932.
753 57. Shi W, Liu C, Annesi-Maesano I, Norback D, Deng Q, Huang C et al. Ambient PM(2.5) and its
754 chemical constituents on lifetime-ever pneumonia in Chinese children: A multi-center study.
755 Environ Int 2021;146:106176. https://doi.org/10.1016/j.envint.2020.106176.
756 58. He F, Yanosky JD, Fernandez-Mendoza J, Chinchilli VM, Al-Shaar L, Vgontzas AN et al. Acute
757 Impact of Fine Particulate Air Pollution on Cardiac Arrhythmias in a Population-Based Sample of
758 Adolescents: The Penn State Child Cohort. J Am Heart Assoc 2022;11(18):e026370.
759 https://doi.org/10.1161/JAHA.122.026370.
760 59. Chaulin AM, Sergeev AK. The Role of Fine Particles (PM 2.5) in the Genesis of Atherosclerosis
761 and Myocardial Damage: Emphasis on Clinical and Epidemiological Data, and Pathophysiological
762 Mechanisms. Cardiol Res 2022;13(5):268-282. https://doi.org/10.14740/cr1366.
763 60. Huang F, Pan B, Wu J, Chen E, Chen L. Relationship between exposure to PM2.5 and lung cancer
764 incidence and mortality: A meta-analysis. Oncotarget 2017;8(26):43322-43331.
765 https://doi.org/10.18632/oncotarget.17313.
766 61. Chen YW, Huang MZ, Chen CL, Kuo CY, Yang CY, Chiang-Ni C et al. PM(2.5) impairs macrophage
767 functions to exacerbate pneumococcus-induced pulmonary pathogenesis. Part Fibre Toxicol
768 2020;17(1):37. https://doi.org/10.1186/s12989-020-00362-2.
769 62. Ge J, Yang H, Lu X, Wang S, Zhao Y, Huang J et al. Combined exposure to formaldehyde and
770 PM(2.5): Hematopoietic toxicity and molecular mechanism in mice. Environ Int
771 2020;144:106050. https://doi.org/10.1016/j.envint.2020.106050.
772 63. Yamanaka R, Barlow C, Lekstrom-Himes J, Castilla LH, Liu PP, Eckhaus M et al. Impaired
773 granulopoiesis, myelodysplasia, and early lethality in CCAAT/enhancer binding protein epsilon-
774 deficient mice. Proc Natl Acad Sci U S A 1997;94(24):13187-13192.
775 https://doi.org/10.1073/pnas.94.24.13187.

776 64. Tanaka T, Akira S, Yoshida K, Umemoto M, Yoneda Y, Shirafuji N et al. Targeted disruption of the
777 NF-IL6 gene discloses its essential role in bacteria killing and tumor cytotoxicity by
778 macrophages. Cell 1995;80(2):353-361. https://doi.org/10.1016/0092-8674(95)90418-2.
779 65. Reckzeh K, Bereshchenko O, Mead A, Rehn M, Kharazi S, Jacobsen SE et al. Molecular and
780 cellular effects of oncogene cooperation in a genetically accurate AML mouse model. Leukemia
781 2012;26(7):1527-1536. https://doi.org/10.1038/leu.2012.37.
782 66. Grossmann M, Metcalf D, Merryfull J, Beg A, Baltimore D, Gerondakis S. The combined absence
783 of the transcription factors Rel and RelA leads to multiple hemopoietic cell defects. Proc Natl
784 Acad Sci U S A 1999;96(21):11848-11853. https://doi.org/10.1073/pnas.96.21.11848.
785 67. McKercher SR, Torbett BE, Anderson KL, Henkel GW, Vestal DJ, Baribault H et al. Targeted
786 disruption of the PU.1 gene results in multiple hematopoietic abnormalities. EMBO J
787 1996;15(20):5647-5658. https://doi.org/.
788 68. Lee H, Lee MW, Warren JR, Ferrie J. Childhood lead exposure is associated with lower cognitive
789 functioning at older ages. Sci Adv 2022;8(45):eabn5164.
790 https://doi.org/10.1126/sciadv.abn5164.
791 69. Rubens O, Logina I, Kravale I, Eglite M, Donaghy M. Peripheral neuropathy in chronic
792 occupational inorganic lead exposure: a clinical and electrophysiological study. J Neurol
793 Neurosurg Psychiatry 2001;71(2):200-204. https://doi.org/10.1136/jnnp.71.2.200.
794 70. Lin JL, Tan DT, Hsu KH, Yu CC. Environmental lead exposure and progressive renal insufficiency.
795 Arch Intern Med 2001;161(2):264-271. https://doi.org/10.1001/archinte.161.2.264.
796 71. Wan H, Wang Y, Zhang H, Zhang K, Chen Y, Chen C et al. Chronic lead exposure induces fatty
797 liver disease associated with the variations of gut microbiota. Ecotoxicol Environ Saf
798 2022;232:113257. https://doi.org/10.1016/j.ecoenv.2022.113257.
799 72. Allada R, Bass J. Circadian Mechanisms in Medicine. N Engl J Med 2021;384(6):550-561.
800 https://doi.org/10.1056/NEJMra1802337.

801 73. Oishi K, Shirai H, Ishida N. CLOCK is involved in the circadian transactivation of peroxisome-
802 proliferator-activated receptor alpha (PPARalpha) in mice. Biochem J 2005;386(Pt 3):575-581.
803 https://doi.org/10.1042/BJ20041150.
804 74. Inoue I, Shinoda Y, Ikeda M, Hayashi K, Kanazawa K, Nomura M et al. CLOCK/BMAL1 is involved
805 in lipid metabolism via transactivation of the peroxisome proliferator-activated receptor (PPAR)
806 response element. J Atheroscler Thromb 2005;12(3):169-174.
807 https://doi.org/10.5551/jat.12.169.
808 75. Kasano-Camones CI, Takizawa M, Ohshima N, Saito C, Iwasaki W, Nakagawa Y et al. PPARalpha
809 activation partially drives NAFLD development in liver-specific Hnf4a-null mice. J Biochem 2023.
810 https://doi.org/10.1093/jb/mvad005.
811 76. Klaunig JE, Babich MA, Baetcke KP, Cook JC, Corton JC, David RM et al. PPARalpha agonist-
812 induced rodent tumors: modes of action and human relevance. Crit Rev Toxicol
813 2003;33(6):655–780. https://doi.org/10.1080/713608372.
814 77. Stanton DT. QSAR and QSPR model interpretation using partial least squares (PLS) analysis. Curr
815 Comput Aided Drug Des 2012;8(2):107–127. https://doi.org/10.2174/157340912800492357.
816 78. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin
817 for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and
818 nucleosome position. Nature methods 2013;10(12):1213–1218.
819 https://doi.org/10.1038/NMETH.2688.
820 79. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T et al. Open TG-GATEs: a large-
821 scale toxicogenomics database. Nucleic Acids Res 2015;43(Database issue):D921–D927.
822 https://doi.org/10.1093/nar/gku955.
823 80. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X et al. A Next Generation
824 Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 2017;171(6):1437–1452
825 e1417. https://doi.org/10.1016/j.cell.2017.10.049.
826

827 **Supplementary Information**

828 The online version contains supplementary material available at:

829 **Additional file 1 (PDF): Supplementary Fig. S1**: Workflow of the proposed approach. First,
830 we detected the pollutant-induced DARs using ATAC-Seq datasets obtained from the
831 TaRGET database (*in vivo* pollutant exposure). We then performed ChIPEA with pollutant-
832 induced DARs to predict TFs that play a pivotal role in the pharmacological effects of
833 pollutants on health. The prediction phase was completed upon construction of pollutant–
834 TF–disorder triadic associations integrating TF-disease associations retrieved from DisGeNET
835 with the pollutant–TF matrices obtained as outcomes of ChIPEA. Subsequently, in the
836 validation phase, we checked the authenticity of the predicted pollutant–disorder
837 associations (shown as blue-colored continuous matrix) by referencing known chemical–
838 disorder associations from CTD (shown as green-colored boolean matrix) and summarized
839 the results as AUROC and AUPR scores. The *in vitro* chemical perturbation dataset obtained
840 from the DBKERO database was also used to evaluate the prediction accuracy of the
841 proposed method. The performance of other methods that utilize chemically induced DEGs
842 as input or employ motif-based enrichment analysis (MotifEA) rather than ChIPEA were
843 simultaneously assessed using the same workflow. **Supplementary Fig. S2**: Summary of the
844 number of DEGs or DARs used as input to the enrichment analyses. Ranking plots are
845 illustrated for individual chemical dosing or pollutant exposure conditions, and are sorted by
846 the summation of the number of up- and down-regulated DEGs or opened and closed DARs.
847 **Supplementary Fig. S3**: Results of the Bisulfite-Seq enrichment analysis for three
848 representative pollutants. Dots indicate individual Bisulfite-Seq peak sets from ChIP-Atlas.
849 Peak sets obtained from Bisulfite-Seq on blood (in the case of TBT and PM2.5 exposure) and
850 liver (in the case of lead exposure) samples are colored (hyper-methylated regions, red;

851 hypo-methylated, blue). Positive and negative values of log2 fold enrichment (opened
852 DARs/closed DARs; X-axis) indicate peak sets enriched in opened and closed chromatin
853 induced by pollutant exposure, respectively. The enrichment score ($-\log_{10}Q$-value; Y-axis)
854 indicates the statistical significance of Bisulfite-Seq enrichment analysis. **Supplementary Fig.**
855 **S4**: Expression levels of TFs mentioned in Figure 4 in mouse tissues before and after
856 exposure to TBT, PM2.5, and lead. Genes encoding the top 20 enriched TFs in Figure 4 (right
857 panels) are colored (rank 1–10, orange; rank 11–20, blue). CPM, count per million.
858
859 **Additional file 2 (XLSX): Supplementary Table S1:** *In vitro* chemical perturbation RNA-Seq
860 dataset (DBKERO). **Supplementary Table S2:** *In vitro* chemical perturbation ATAC-Seq
861 dataset (DBKERO). **Supplementary Table S3:** *In vivo* pollutant exposure RNA-Seq dataset
862 (TaRGET). **Supplementary Table S4:** *In vivo* pollutant exposure ATAC-Seq dataset (TaRGET).
863 Data shown in bold font are those that yielded high AUROC scores, as reported in the main
864 text; see also Supplementary Table S6. **Supplementary Table S5:** AUROC and AUPR scores
865 for *in vitro* chemical perturbation datasets (DBKERO). **Supplementary Table S6:** AUROC and
866 AUPR scores for *in vivo* pollutant exposure datasets (TaRGET). Data shown in bold font are
867 those that yielded high AUROC scores, as reported in the main text; see also Supplementary
868 Table S4. **Supplementary Table S7:** Result of DAR-ChIPEA for exposing blood samples to
869 TBT. **Supplementary Table S8:** Result of DAR-ChIPEA for exposing blood samples to $PM_{2.5}$.
870 **Supplementary Table S9:** Result of DAR-ChIPEA for exposing liver samples to lead.
871
872 **FIGURE LEGENDS**
873 **Fig. 1** Overview of the proposed DAR-ChIPEA approach. **a** Schematic diagram illustrating the
874 difference between our previously reported DEG-ChIPEA method [22] and the proposed

875 DAR-ChIPEA approach. The DEG-ChIPEA method is restricted to analyzing TF binding to the
876 vicinity of gene loci. In contrast, the DAR-ChIPEA approach comprehensively covers TF
877 binding to accessible chromatin regions (obtained from ATAC-seq) across the genome
878 including proximal (e.g., promoters) and distal (e.g., enhancers) regulatory elements. The
879 circles surrounded by dotted lines represent the TFs binding to distal regulatory elements
880 that are not detected in the DEG-ChIPEA method. **b** To demonstrate the feasibility of DAR-
881 ChIPEA, we identified the TFs enriched at pollutant-induced DARs by analyzing large-scale
882 ChIP-Seq data. Overlaps were evaluated between the DARs (green box) and peak-call data
883 (black lines) of TF-related ChIP-Seq experiments archived in ChIP-Atlas
884
885 **Fig. 2** Validation of the proposed DAR-ChIPEA approach to predict TF-driven MoAs of *in*
886 *vitro* chemical perturbation. **a** Schematic diagram for TF enrichment analyses using *in vitro*
887 chemical perturbation datasets. DEGs or DARs upon chemical perturbation were detected
888 using data from RNA-Seq or ATAC-Seq, respectively, followed by enrichment analyses to
889 profile TF binding. Methods used for TF enrichment analysis are shown in blue; the proposed
890 method (DAR-ChIPEA) is highlighted in bold. **b** Ranking plot of maximum enrichment scores
891 ($-\log_{10}Q$-value) for each dosing condition obtained using the DAR-ChIPEA and DEG-ChIPEA
892 methods. c and d Distribution of AUROC (**c**) and AUPR (**d**) scores for each chemical–TF–
893 disorder triadic association predicted using DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and
894 ChIPEA. Orange dots represent mean scores; global scores are noted below the plots.
895 Differences between the methods are presented as *Q*-values (two-tailed Wilcoxon rank-sum
896 test) above the bee swarm plots
897

898 **Fig. 3** Application of the DAR-ChIPEA approach to predict TF-driven modalities of pollutant
899 administration *in vivo*. **a** Schematic diagram for TF enrichment analyses using *in vivo*
900 pollutant exposure datasets. DEGs or DARs upon pollutant exposure were detected using
901 RNA-Seq or ATAC-Seq, respectively, followed by enrichment analyses to profile TF binding.
902 Methods used for TF enrichment analysis are shown in blue, with the proposed DAR-ChIPEA
903 method highlighted in bold. **b** Ranking plot of maximum enrichment scores ($-\log_{10}Q$-value)
904 for each exposure condition as obtained using the DAR-ChIPEA and DEG-ChIPEA methods. c
905 and d Distribution of AUROC (**c**) and AUPR (**d**) scores for each pollutant–TF–disorder triadic
906 association predicted using DEG-MotifEA, DAR-MotifEA, DEG-ChIPEA, and ChIPEA. Orange
907 dots represent mean scores; global scores are noted beneath the bee swarm plots.
908 Differences between the methods are presented as *Q*-values (two-tailed Wilcoxon rank-sum
909 test) above the bee swarm plots
910
911 **Fig. 4** Biological interpretations of the predicted pollutant–TF–disorder associations for three
912 representative pollutants. Mouse phenotype ontologies enriched for the genes proximal to
913 pollutant-induced opened DARs are shown in the left panels. Bars representing individual
914 ontology terms are colored based on their parent categories, which were obtained from the
915 Mouse Genome Informatics database. Volcano plots illustrating the results of the proposed
916 DAR-ChIPEA approach are shown in the right panels. Dots indicate individual TFs and are
917 colored based on cell type classes obtained from ChIP-Atlas. Positive and negative values of
918 $\log_2$ fold enrichment (opened DARs/closed DARs; X-axis) indicate TFs enriched in relaxed and
919 condensed chromatin induced by pollutant exposure, respectively. The enrichment score (–
920 $\log_{10}Q$-value; Y-axis) indicates the statistical significance of ChIPEA.
921

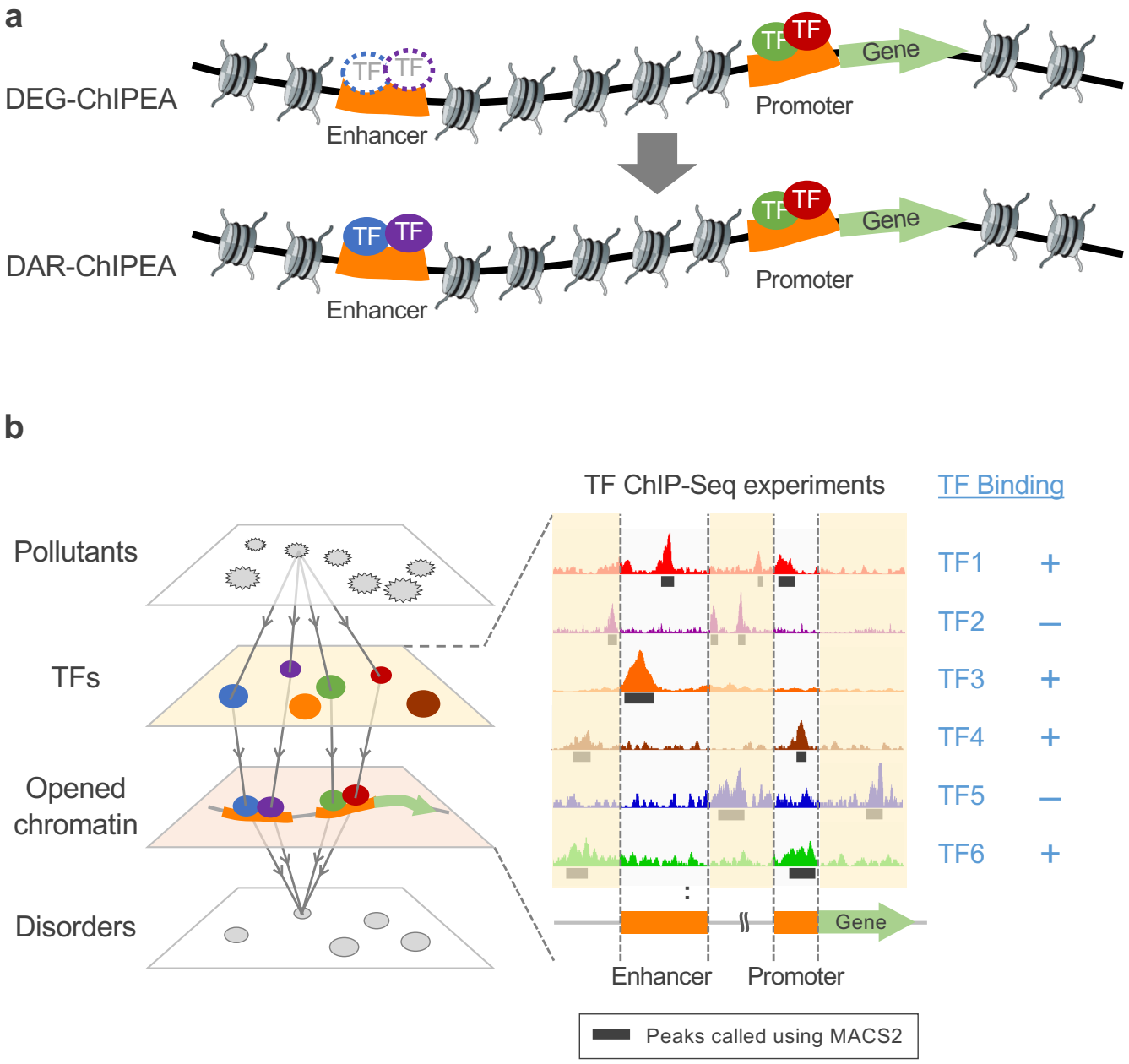
a
DEG-ChIPEA
TF
TF
Enhancer
TF
TF
Gene
Promoter
DAR-ChIPEA
TF
TF
Enhancer
TF
TF
Gene
Promoter
b
Pollutants
TFs
Opened chromatin
Disorders
TF ChIP-Seq experiments
TF Binding
TF1 +
TF2 –
TF3 +
TF4 +
TF5 –
TF6 +
Gene
Enhancer
Promoter
Peaks called using MACS2

Figure 1

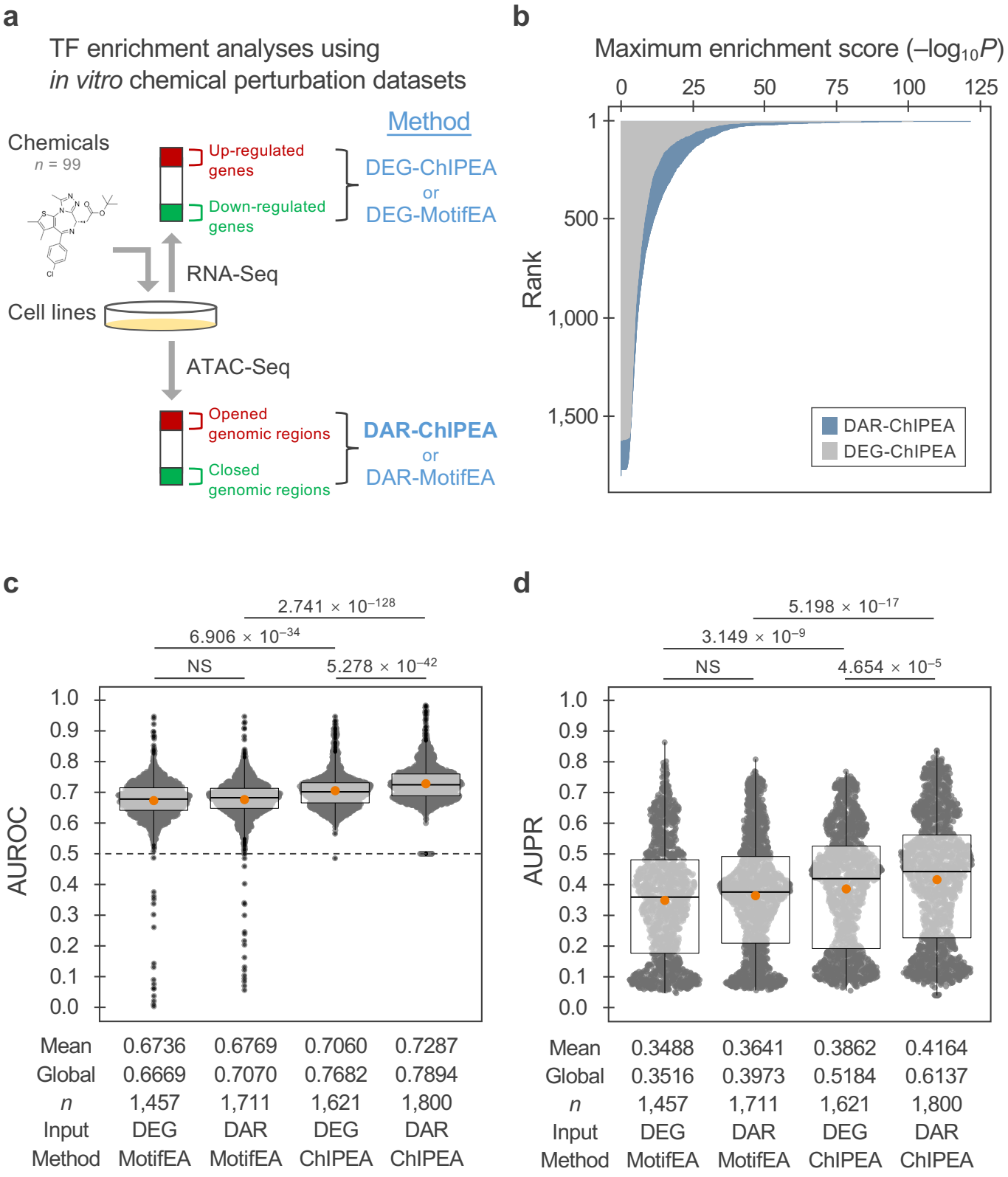
a
TF enrichment analyses using
in vitro chemical perturbation datasets
Method
Chemicals
n = 99
Up-regulated genes
Down-regulated genes
DEG-ChIPEA
or
DEG-MotifEA
RNA-Seq
Cell lines
ATAC-Seq
Opened genomic regions
Closed genomic regions
DAR-ChIPEA
or
DAR-MotifEA
b
Maximum enrichment score (–log10P)
0
25
50
75
100
125
Rank
1
500
1,000
1,500
DAR-ChIPEA
DEG-ChIPEA
c
2.741 × 10−128
6.906 × 10−34
NS
5.278 × 10−42
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
5.198 × 10−17
3.149 × 10−9
NS
4.654 × 10−5
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

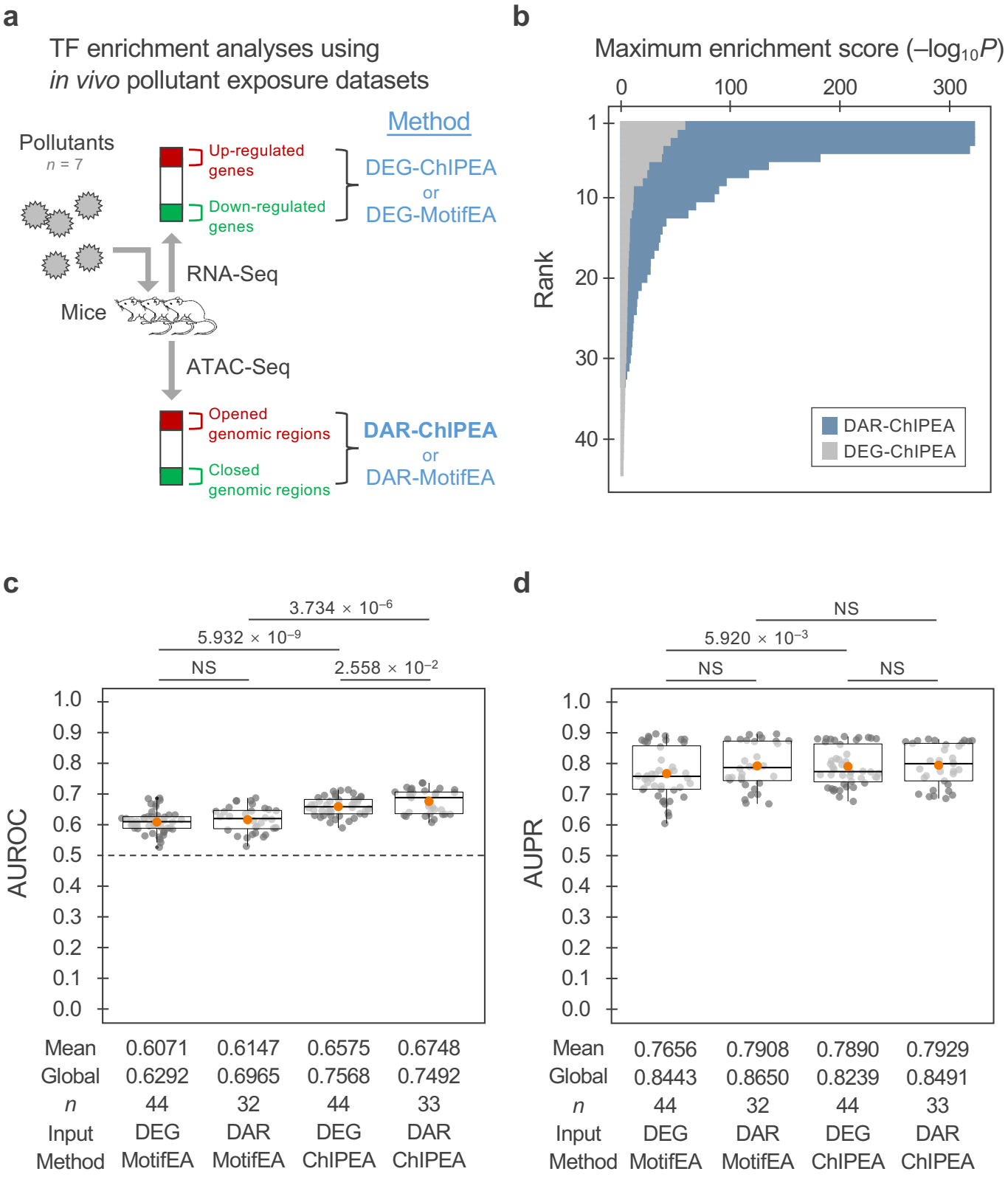
a
TF enrichment analyses using
in vivo pollutant exposure datasets
Method
Pollutants
n = 7
Up-regulated genes
Down-regulated genes
DEG-ChIPEA
or
DEG-MotifEA
RNA-Seq
Mice
ATAC-Seq
Opened genomic regions
Closed genomic regions
DAR-ChIPEA
or
DAR-MotifEA
b
Maximum enrichment score (–log10P)
Rank
DAR-ChIPEA
DEG-ChIPEA
c
3.734 × 10−6
5.932 × 10−9
NS
2.558 × 10−2
AUROC
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
NS
5.920 × 10−3
NS
NS
AUPR
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

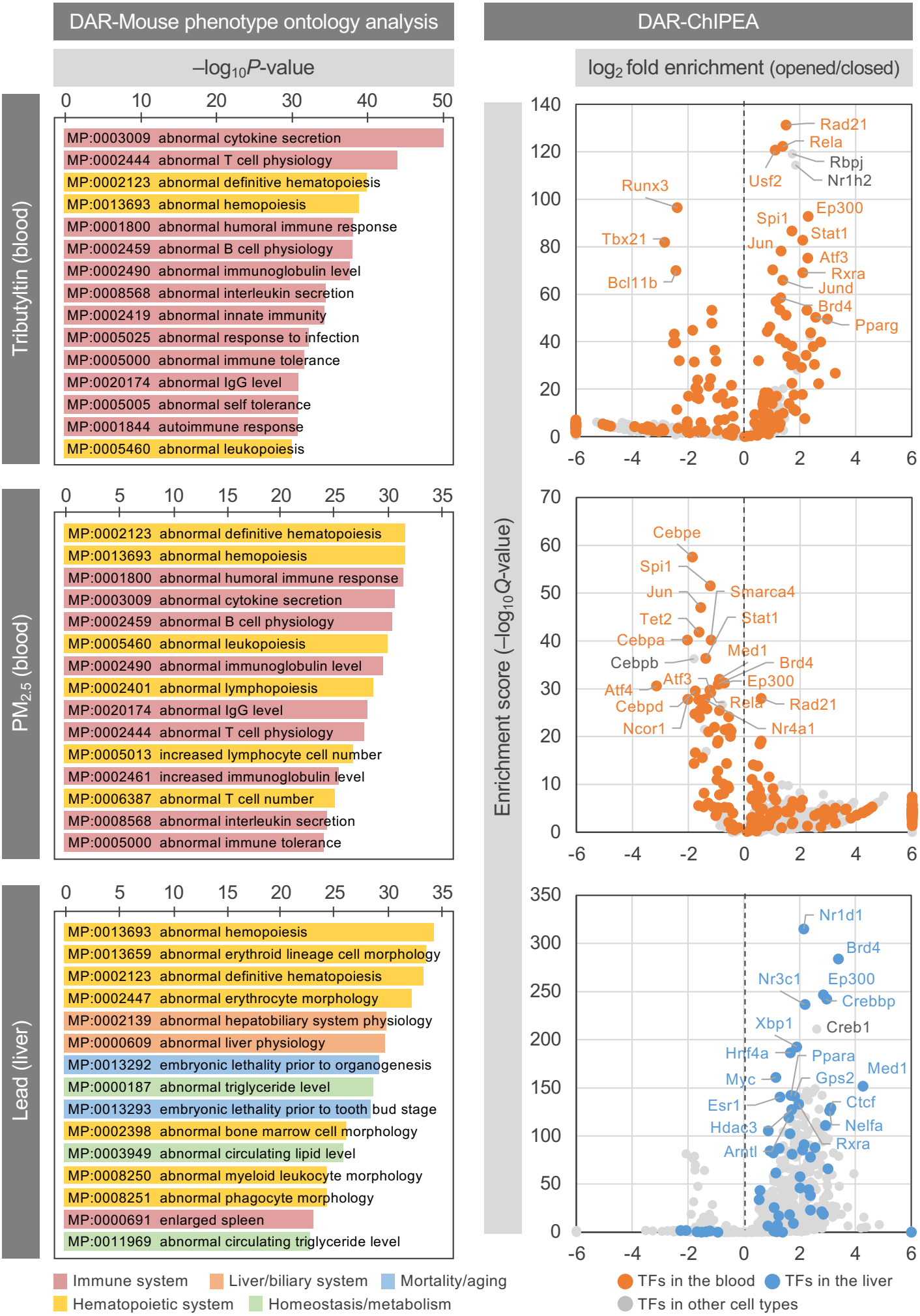
DAR-Mouse phenotype ontology analysis
–log10P-value
Tributyltin (blood)
MP:0003009 abnormal cytokine secretion
MP:0002444 abnormal T cell physiology
MP:0002123 abnormal definitive hematopoiesis
MP:0013693 abnormal hemopoiesis
MP:0001800 abnormal humoral immune response
MP:0002459 abnormal B cell physiology
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 (opened/closed)
Enrichment score (–log10Q-value)
Rad21
Rela
Rbpj
Nr1h2
Usf2
Runx3
Ep300
Spi1
Stat1
Tbx21
Jun
Atf3
Rxra
Bcl11b
Jund
Brd4
Pparg
Cebpe
Spi1
Jun
Smarca4
Tet2
Stat1
Cebpa
Med1
Cebpb
Brd4
Atf3
Ep300
Atf4
Cebpd
Rela
Rad21
Ncor1
Nr4a1
Nr1d1
Brd4
Nr3c1
Ep300
Crebbp
Xbp1
Creb1
Hnf4a
Ppara
Med1
Myc
Gps2
Esr1
Ctcf
Hdac3
Nelfa
Arntl
Rxra
TFs in the blood
TFs in the liver
TFs in other cell types

Figure 4