BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Epigenetic landscape of drug responses revealed through large-scale ChIP-seq data analyses

Zhaonan Zou[1,2,3] , Michio Iwata[4] , Yoshihiro Yamanishi[4*] and Shinya Oki[1,5*]

*Correspondence:
yamani@bio.kyutech.ac.jp;
oki.shinya.3w@kyoto-u.ac.jp
[1] Department of Drug
Discovery Medicine, Kyoto
University Graduate School
of Medicine, 53 Shogoin
Kawahara-cho, Sakyo-ku,
Kyoto 606-8507, Japan[4]
Department of Bioscience
and Bioinformatics,
Faculty of Computer
Science and Systems
Engineering, Kyushu Institute
of Technology, 680-4 Kawazu,
Iizuka, Fukuoka 820-8502,
Japan
Full list of author information
is available at the end of the
article

Abstract

Background: Elucidating the modes of action (MoAs) of drugs and drug candidate compounds is critical for guiding translation from drug discovery to clinical application. Despite the development of several data-driven approaches for predicting chemical–disease associations, the molecular cues that organize the epigenetic landscape of drug responses remain poorly understood.

Results: With the use of a computational method, we attempted to elucidate the epigenetic landscape of drug responses, in terms of transcription factors (TFs), through large-scale ChIP-seq data analyses. In the algorithm, we systematically identified TFs that regulate the expression of chemically induced genes by integrating transcriptome data from chemical induction experiments and almost all publicly available ChIP-seq data (consisting of 13,558 experiments). By relating the resultant chemical–TF associations to a repository of associated proteins for a wide range of diseases, we made a comprehensive prediction of chemical–TF–disease associations, which could then be used to account for drug MoAs. Using this approach, we predicted that: (1) cisplatin promotes the anti-tumor activity of TP53 family members but suppresses the cancer-inducing function of MYCs; (2) inhibition of RELA and E2F1 is pivotal for leflunomide to exhibit antiproliferative activity; and (3) CHD8 mediates valproic acid-induced autism.

Conclusions: Our proposed approach has the potential to elucidate the MoAs for both approved drugs and candidate compounds from an epigenetic perspective, thereby revealing new therapeutic targets, and to guide the discovery of unexpected therapeutic effects, side effects, and novel targets and actions.

Keywords: Drug modes of action, Transcriptome, ChIP-seq, Transcription factor, Epigenetic landscape

Introduction

Elucidating the modes of action (MoAs) of drugs and candidate compounds is critical for guiding translation from drug discovery to clinical application. Understanding the complex responses of the human biological system to chemicals is of vital importance in medical and pharmaceutical research. For many chemicals, including some approved drugs, the MoAs remain elusive. The task of revealing MoAs can be moderately

Zou *et al. BMC Bioinformatics*   (2022) 23:51

Page 2 of 20

simplified to the identification of target proteins implicated in the pharmacological effects of chemicals on disease. Phenotype-based high-throughput screening (PHTS) is an efficient way to find candidate compounds with a desired phenotype [1–3]. Although PHTS can rapidly screen thousands of chemicals, the underlying molecular mechanisms of the hit compounds remain unknown. Identification of the target proteins associated with a phenotype requires considerable effort, e.g., analysis of drug–protein interactions using biochemical and chemoinformatic methods [4–11]. Recent developments in biotechnology have contributed to the increase in the amounts of omics data for chemicals and proteins in the genome, transcriptome, epigenome, and interactome, which can be useful sources for inferring the MoAs of drugs.

Based on the concept that gene expression changes are pivotal for pharmacophysiological effects, another major approach to drug discovery is transcriptome profiling following administration of compounds. A popular transcriptome-guided drug repositioning approach for finding novel drugs is to search for compounds whose gene expression patterns are inversely correlated with those of a disease of interest [12–16]. These methods involve genome-wide expression profiling of transcriptional responses to compound treatment. In recent years, chemically induced gene expression data based on large-scale transcriptome experiments have been made available from several public databases. The Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system (TG-GATEs) hosts the results of a toxicogenomics project implemented in Japan, in which 170 compounds were used to perturb cell homeostasis in vitro [17]. Subsequently, the Connectivity Map (CMap) database was constructed to provide the gene expression profiles of five cancer cell lines perturbed by 1,309 compounds [18]. Many more compounds have been examined by the National Institute of Health Library of Integrated Network-Based Cellular Signatures (LINCS) consortium, which analyzed the transcriptomic responses of 68 human cell lines to more than 20,000 compounds [19]. The LINCS consortium took advantage of a "reduced representation of the transcriptome", in which 978 landmark genes, termed L1000, are investigated as a representative gene set for biological significance. The Comparative Toxicogenomics Database (CTD) is another public resource that provides information about differentially expressed genes (DEGs) following administration of chemicals and medical drugs [20]. The distinguishing feature of CTD is that all records were generated by manual curation of more than 13,713 peer-reviewed publications, including expression changes of 23,081 genes upon administration of 4,121 compounds to human cells. Thus, relative to the aforementioned large-scale projects, CTD integrates a wider variety of genes in an unbiased manner.

Despite the construction of chemically induced gene expression signatures, the molecular cues that mediate gene expression changes in response to chemical administration remain to be clarified. One class of promising mediators are the transcription factors (TFs) that act upstream of sets of chemically induced DEGs to control their expression. TFs are master regulators that profoundly alter cell phenotype and behavior by modulating the epigenetic landscape, thereby organizing the expression of large sets of genes. Although it is true that TFs are not often directly targeted by drugs, recent studies have demonstrated the power of epigenetic drug discovery, e.g., by revealing the potential utility of inhibitors of bromodomain proteins and histone deacetylases (HDACs) against neoplasms [21–25]. These findings highlight the importance of modeling epigenetic

Zou *et al. BMC Bioinformatics*   (2022) 23:51

Page 3 of 20

landscapes in the context of pharmacological strategies. Although specifically targeting TF activity faces major hurdles, targeting effectors downstream of cell signaling (e.g., TFs) rather than upstream factors is likely to be a more specific approach [26]. In this study, instead of predicting directly druggable targets, we developed a computational method for identifying TFs pivotal to thousands of chemically induced DEGs, making full use of large-scale TF binding profiles obtained from tens of thousands of actual chromatin immunoprecipitation sequencing (ChIP-seq) datasets. The predicted chemical–TF associations provided clues about drug MoAs involved in drug efficacies and side effects. Our approach outperformed methods that directly evaluated the similarity of chemically induced and disease-specific DEGs without considering key TFs.

Results

Overview of TF-focused elucidation of drug MoAs

A considerable proportion of bioactive compounds and medical drugs exert their effects by modifying disease-elicited gene expression. To further understand the MoAs of chemicals and, in turn, to define chemical–disease associations, it is of crucial importance to focus on the master regulators that organize the expression of chemically perturbed DEGs. In the proposed approach, shown in Fig. 1, we identified TFs that integratively regulate chemically perturbed DEGs by analyzing large-scale comprehensive ChIP-seq data obtained from ChIP-Atlas [27]. In addition, we evaluated the matches of predicted target TFs with disease-associated proteins available from the DisGeNET database [28] (details shown below). The predicted chemical–TF and chemical–disease associations were validated with known chemical–protein associations from the Kyoto Encyclopedia of Genes and Genomes DRUG database (KEGG DRUG) [29] and the chemical–disease association dataset from CTD, respectively.

Evaluation of biological significance of genes and TFs in databases

To confirm the quality of chemically induced transcriptomes, we evaluated the biological significance of annotated chemically perturbed genes by comparing the statistics
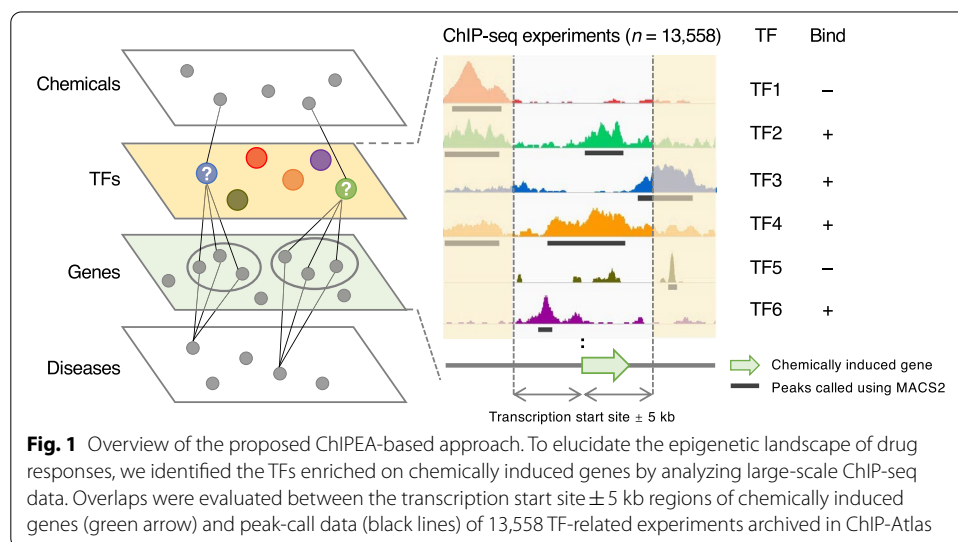

Fig. 1 Overview of the proposed ChIPEA-based approach. To elucidate the epigenetic landscape of drug responses, we identified the TFs enriched on chemically induced genes by analyzing large-scale ChIP-seq data. Overlaps were evaluated between the transcription start site ± 5 kb regions of chemically induced genes (green arrow) and peak-call data (black lines) of 13,558 TF-related experiments archived in ChIP-Atlas

of gene expression profile data in CTD and L1000 in order to estimate their overlap. Additional file 1: Figure S1a shows a Venn diagram of the genes that overlapped among the comparisons. Almost all (99.3%) of the L1000 landmark genes are also included in CTD. We sorted chemically induced genes in the order of the frequency with which they appeared in CTD, and assessed their matches with L1000 landmark genes (Additional file 1: Fig. S1b). The results confirmed that the representative L1000 genes were ranked generally higher among the genes annotated in CTD ($p = 2.5 \times 10^{-176}$ by Wilcoxon rank-sum test). In particular, 41 DEGs in the top 100 of CTD were designated as L1000 genes. These data suggest that CTD includes a wide variety of biologically significant genes.
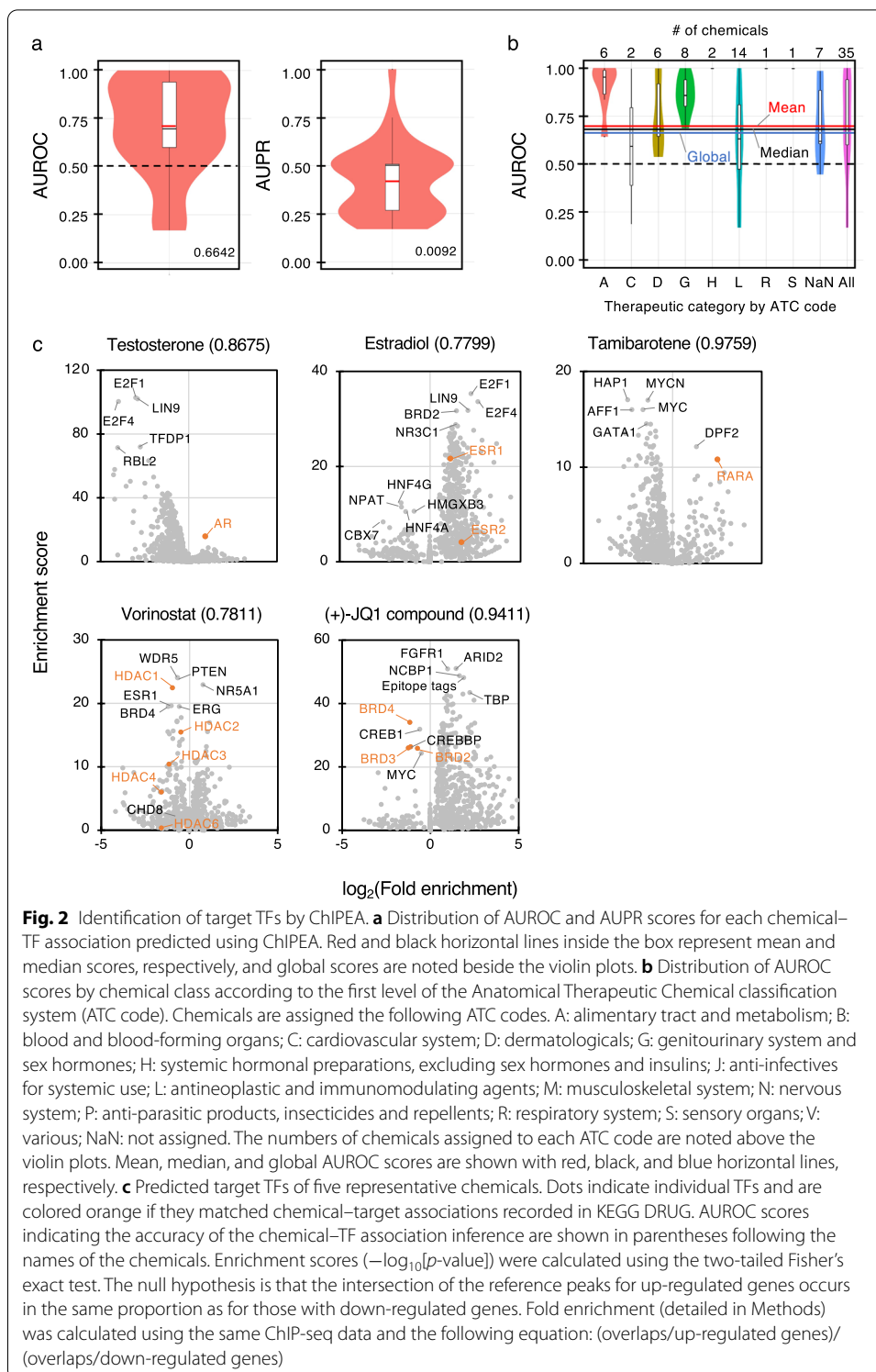
We then compared the annotated proteins in DisGeNET and ChIP-Atlas (Additional file 1: Fig. S1c). Most of the TFs profiled by ChIP-seq experiments (67.9% of ChIP-Atlas TFs) are also curated in DisGeNET. We sorted disease-associated proteins in the order of the frequency with which they appeared in DisGeNET, and assessed the match with TFs contained in ChIP-Atlas (Additional file 1: Fig. S1d). The results confirmed that the TFs were ranked generally higher among the proteins annotated in DisGeNET ($p = 3.2 \times 10^{-33}$ by Wilcoxon rank-sum test). These data suggest that the TFs analyzed by ChIP-seq associate with diseases more strongly than other proteins within DisGeNET. In summary, CTD, ChIP-Atlas, and DisGeNET include information about biologically significant genes and proteins, and are therefore suitable for elucidating the MoAs of chemicals and inferring chemical–disease associations.

Identification of master regulators that organize the expression of DEGs in response to drug treatment

Chemical perturbation of gene expression is organized by a series of TFs in an integrated manner. Therefore, identification of key TFs is critical for understanding drug MoAs. To this end, we combined chemically induced DEGs from CTD (Additional file 2: Table S1) and large-scale public ChIP-seq data ($n = 13,558$) from ChIP-Atlas (Fig. 1, Additional file 2: Table S2). We then performed ChIP-seq-based enrichment analysis (ChIPEA, detailed in Methods) to identify TFs that exhibited enriched binding around up- or down-regulated genes following drug administration (target range: transcription start site ± 5 kb).

We then asked whether the TFs with higher enrichment scores were involved in the MoAs of query compounds. As standard data, we used known chemical–protein inter-actions data obtained from KEGG DRUG and applied the receiver operating character-istic (ROC) curve, a plot of true-positive rates as a function of false-positive rates, as well as the precision-recall (PR) curve, which is a plot of precision (positive predictive value) as a function of recall (sensitivity). We then summarized the evaluation using the area under the ROC curve (AUROC) score, where 1 is perfect classification and 0.5 is random classification, and the area under the PR curve (AUPR) score, where 1 is perfect inference and the ratio of positive examples in the standard data is random inference.

The distribution of AUROC and AUPR scores for each chemical ($n = 35$) that directly targets TFs was visualized with a violin plot (Fig. 2a). Mean AUROC and mean AUPR across chemicals were 0.7063 and 0.4187, respectively. Given that the distribution pattern of AUROC varied depending on the highest enrichment score of each chemical (Additional file 1: Fig. S2), we were concerned that the item discrimination to distinguish

Fig. 2 Identification of target TFs by ChIPEA. **a** Distribution of AUROC and AUPR scores for each chemical–TF association predicted using ChIPEA. Red and black horizontal lines inside the box represent mean and median scores, respectively, and global scores are noted beside the violin plots. **b** Distribution of AUROC scores by chemical class according to the first level of the Anatomical Therapeutic Chemical classification system (ATC code). Chemicals are assigned the following ATC codes. A: alimentary tract and metabolism; B: blood and blood-forming organs; C: cardiovascular system; D: dermatologicals; G: genitourinary system and sex hormones; H: systemic hormonal preparations, excluding sex hormones and insulins; J: anti-infectives for systemic use; L: antineoplastic and immunomodulating agents; M: musculoskeletal system; N: nervous system; P: anti-parasitic products, insecticides and repellents; R: respiratory system; S: sensory organs; V: various; NaN: not assigned. The numbers of chemicals assigned to each ATC code are noted above the violin plots. Mean, median, and global AUROC scores are shown with red, black, and blue horizontal lines, respectively. **c** Predicted target TFs of five representative chemicals. Dots indicate individual TFs and are colored orange if they matched chemical–target associations recorded in KEGG DRUG. AUROC scores indicating the accuracy of the chemical–TF association inference are shown in parentheses following the names of the chemicals. Enrichment scores ($-\log_{10}[p\text{-value}]$) were calculated using the two-tailed Fisher's exact test. The null hypothesis is that the intersection of the reference peaks for up-regulated genes occurs in the same proportion as for those with down-regulated genes. Fold enrichment (detailed in Methods) was calculated using the same ChIP-seq data and the following equation: (overlaps/up-regulated genes)/(overlaps/down-regulated genes)

associated and not associated TFs within individual chemicals was in some sense limited. Therefore, we calculated "global" statistics using an inter-chemical merged enrichment score vector to emphasize the significance of the actual values of enrichment scores (detailed in the Methods section). The global AUROC and global AUPR scores

were 0.6642 and 0.0092, respectively. Figure 2b shows the distribution of AUROC scores for a number of chemical classes on the basis of the first level of the WHO Anatomical Therapeutic Chemical (ATC) Classification code; the detailed explanations on the ATC codes are shown in the figure caption. These results revealed that ChIPEA can generally predict chemical–TF associations with high efficiency, particularly for chemicals categorized as ATC code A (alimentary tract and metabolism, e.g., diabetes treatments and vitamins) and G (genitourinary system and sex hormones), which directly target nuclear receptors.

Figure 2c illustrates representative predictions of chemical–TF associations (Additional file 2: Table S3). For example, androgen receptor (AR), estrogen receptor (ESR) 1/2, and retinoic acid receptor alpha (RARA) were shown to significantly bind to the up-regulated genes induced by testosterone (primary male sex hormone), estradiol (major female sex hormone), and tamibarotene (synthetic retinoid) treatment, respectively [30–32]. Though they did not exhibit the highest enrichment, the binding of histone deacetylases (HDACs) was clearly detected among the TFs that bound in a biased manner to down-regulated genes after treatment with vorinostat, a pan-HDAC inhibitor. In addition, bromodomain-containing proteins (BRDs), which are crucial factors involved in the transcription elongation process, were enriched among the down-regulated genes after treatment with (+)-JQ1 compound (JQ1), a bromodomain and extra-terminal motif (BET) protein inhibitor and potential antineoplastic agent. These results suggest that testosterone, estradiol, and tamibarotene promote, whereas vorinostat and JQ1 suppress, the activities of the corresponding receptors or factors in a manner consistent with the evidence [30–34].

Highlighting MoAs along with presumed chemical–TF–disease associations of CTD chemicals

To examine the MoAs more deeply, and in turn construct the chemical–TF–disease associations, we linked the TFs enriched for chemically induced DEGs by ChIPEA with protein–disease associations according to the DisGeNET database. In the proposed method, the probabilities of chemical–TF–disease associations were simply represented by the ChIPEA enrichment scores. For comparison, we also used a conventional DEG-connected method for chemical–disease association analysis, which directly calculates the positive or negative correlation between chemically induced DEGs from CTD and disease-specific DEGs from Crowd Extracted Expression of Differential Signatures (CREEDS) [35]. The sets of genes included in CTD and CREEDS modestly overlapped with each other (69.3% and 81.1% of genes, respectively), and CREEDS was confirmed to include an extensive range of biologically significant genes suitable for inferring chemical–disease associations (Additional file 1: Fig. S1a, b).

To compare the accuracies of predicted chemical–disease associations between the proposed method and the DEG-connected method, we applied ROC and PR curves using known chemical–disease associations obtained from CTD as standard data. The global AUROC and global AUPR of the proposed approach were 0.6839 and 0.0574, respectively (mean AUROC $=0.7026$, mean AUPR $=0.3504$), higher than that of approaches that only compared DEGs versus non-DEGs (global AUROC $=0.6286$, mean AUROC $=0.6133$, global AUPR $=0.0461$, mean AUPR $=0.2790$; $p=2.7\times10^{-15}$ for
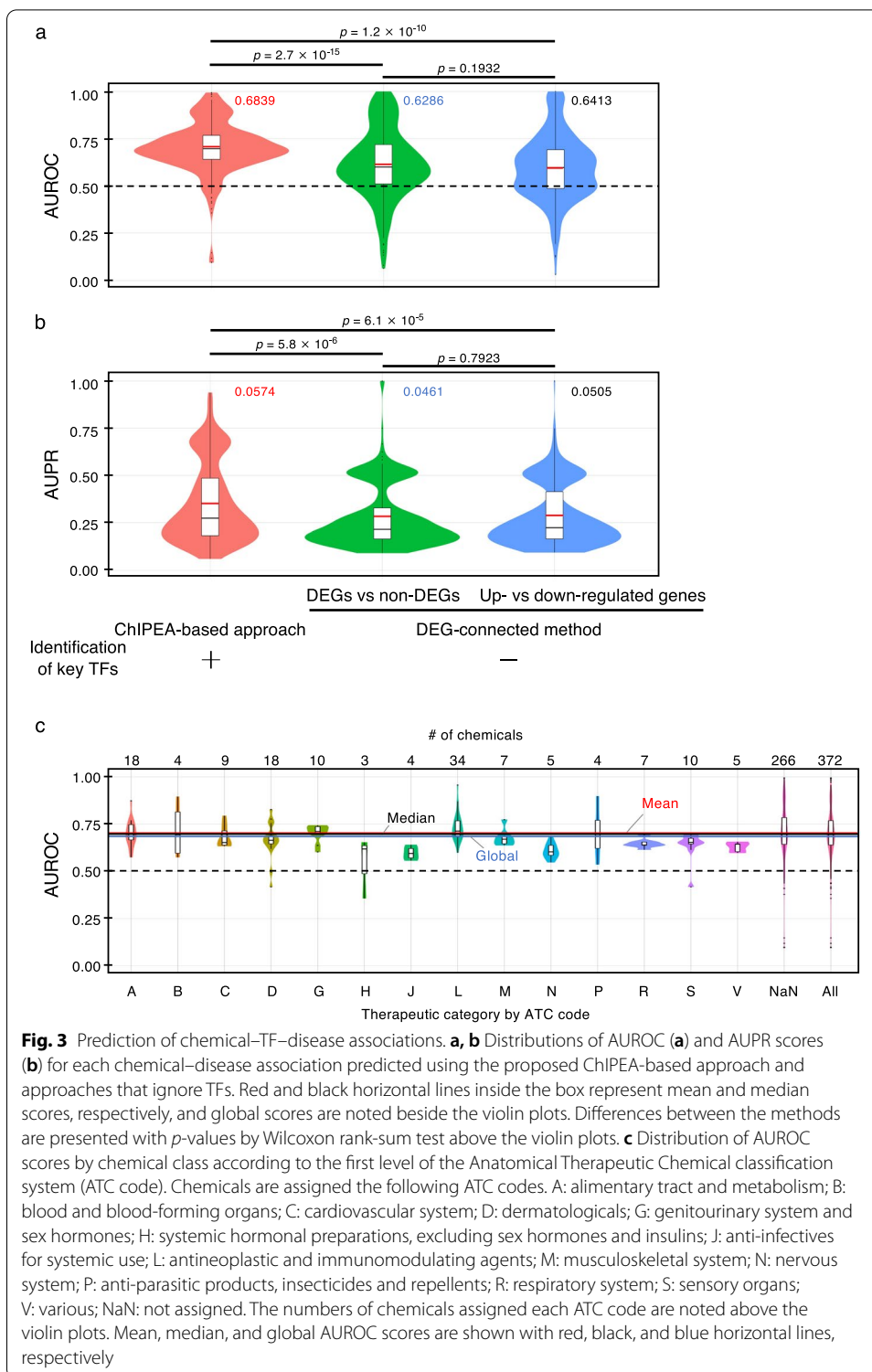
AUROC, $p = 5.8 \times 10^{-6}$ for AUPR by Wilcoxon rank-sum test) or up- versus down-regulated genes (global AUROC = 0.6413, mean AUROC = 0.5972, global AUPR = 0.0505, mean AUPR = 0.2846; $p = 1.2 \times 10^{-10}$ for AUROC, $p = 6.1 \times 10^{-5}$ for AUPR by Wilcoxon rank-sum test) (Fig. 3a). Figure 3b shows the distribution of AUROC scores of the purposed method, stratified by ATC code. These results suggest that the identification of target TFs using ChIPEA is a powerful approach that can be used to clarify pivotal factors for specific MoAs and is therefore useful for predicting the diseases associated with treatments, particularly for chemicals categorized as ATC code G (genitourinary system and sex hormones) or L (antineoplastic and immunomodulating agents). Use of this approach can ultimately lead to increased predictive power relative to directly calculating the commonalities between chemically induced and disease-specific DEGs.

Biological interpretations of chemical–TF–disease associations to estimate the pivotal TFs involved in efficacies and side effects of CTD chemicals

Figure 4 shows representative results of chemical–TF–disease associations constructed using the proposed approach for cisplatin, leflunomide, and valproic acid.

Cisplatin is an antineoplastic chemotherapy agent that acts by crosslinking DNA, resulting in DNA damage; in cancer cells, the drug activates ferroptosis and apoptosis [36–39]. Among the predicted target TFs, tumor suppressors TP53 and TP63 showed the most significant enrichment for genes up-regulated by cisplatin administration [40, 41]. By contrast, oncogenic factors MYC and MYCN were the most enriched TFs for the down-regulated genes [42]. These four TFs were assigned to various types of cancers in DisGeNET, leading to the prediction that cisplatin is associated with neoplasms; this prediction is firmly consistent with the evidence [38, 39]. Although the primary MoA of cisplatin is well known to involve crosslinking of genomic DNA, there is no evidence to support a direct interaction between TP53 or MYCs and cisplatin. Therefore, it is likely that cisplatin indirectly promotes the anti-tumor function of TP53 while suppressing the cancer-inducing function of MYCs. This suggests that our approach is useful for discovering indirect mediators in pharmacological processes that cannot be found by methods based on molecular structures, such as docking simulation and structure-based machine learning.

Leflunomide (LEF) is approved for treating adult rheumatoid arthritis. The drug acts mainly through direct inhibition of dihydroorotate dehydrogenase (DHODH), which is thought to impair proliferation of inflammatory T cells by blocking de novo pyrimidine biosynthesis [43–47]. Controversially, some reports have shown that impairment of cell proliferation by LEF is not rescued by uridine supplementation, implying that the drug also exerts DHODH-independent effects [48, 49]. Remarkably, our ChIPEA-based approach revealed that the most enriched TF for the LEF-induced up-regulated genes was TLE3, which suppresses cellular proliferation by inhibiting MAPK pathways [50, 51]. Another intriguing TF enriched for the up-regulated DEGs was AHR, an important regulator of inflammation in the immune system. This is consistent with the fact that LEF induces the AHR–ARNT interaction, thereby attenuating bone erosion in rheumatoid arthritis [52]. In addition, LEF was also predicted to inactivate the NF-κB pathway component RELA and the key cell cycle promoter E2F1, both of which activate cellular proliferation [53, 54]. These results suggest that LEF suppresses proliferation of

Zou *et al. BMC Bioinformatics*   (2022) 23:51

Page 8 of 20

Fig. 3 Prediction of chemical–TF–disease associations. **a, b** Distributions of AUROC (**a**) and AUPR scores (**b**) for each chemical–disease association predicted using the proposed ChIPEA-based approach and approaches that ignore TFs. Red and black horizontal lines inside the box represent mean and median scores, respectively, and global scores are noted beside the violin plots. Differences between the methods are presented with *p*-values by Wilcoxon rank-sum test above the violin plots. **c** Distribution of AUROC scores by chemical class according to the first level of the Anatomical Therapeutic Chemical classification system (ATC code). Chemicals are assigned the following ATC codes. A: alimentary tract and metabolism; B: blood and blood-forming organs; C: cardiovascular system; D: dermatologicals; G: genitourinary system and sex hormones; H: systemic hormonal preparations, excluding sex hormones and insulins; J: anti-infectives for systemic use; L: antineoplastic and immunomodulating agents; M: musculoskeletal system; N: nervous system; P: anti-parasitic products, insecticides and repellents; R: respiratory system; S: sensory organs; V: various; NaN: not assigned. The numbers of chemicals assigned each ATC code are noted above the violin plots. Mean, median, and global AUROC scores are shown with red, black, and blue horizontal lines, respectively

inflammatory cells in rheumatoid arthritis patients by influencing those TFs in a manner independent of DHODH. The top hit of RELA, along with the well-known interaction of NF-κB with the immunoglobulin light-chain enhancer in B cells, is consistent with previous findings that LEF prevents immunoglobulin production through inhibition of

Zou *et al. BMC Bioinformatics*   (2022) 23:51

Page 9 of 20

Fig. 4 Biological interpretations of the chemical–TF–disease associations. Representative diseases are associated with the predicted target TFs of three chemicals. Dots indicate individual TFs and are colored orange only if they match the chemical–disease associations recorded in CTD database. Chemical–disease associations are indicated in blue (marked as "therapeutic" in CTD) or red (marked as "marker/mechanism" in CTD). AUROC scores showing the accuracy of chemical–TF association inference are shown in parentheses following the names of chemicals

tyrosine kinase activity [55, 56]. Other than its antirheumatic effect, LEF has potential anticancer activity [57–59]. Although the underlying mechanism is elusive, our ChIPEA suggested that LEF suppresses tumor growth by activating TLE3 and inactivating RELA and E2F1, possibly via the same mechanisms involved in rheumatoid arthritis treatment. Thus, our proposed approach can be used to reveal MoAs that would not be expected from consensus interpretations.

Valproic acid (VPA), a structurally simple fatty acid, has anticonvulsant properties and has been widely applied in the treatment of epilepsy [60]. It is also a potent HDAC inhibitor and is under investigation as a treatment for various cancers [61–64]. In this analysis, several factors involved in chromatin remodeling, such as KDM1A, BRD2/4, and EP300, were significantly enriched to the down-regulated genes in response to VPA administration. Given that aberrant chromatin remodeling is a hallmark of oncogenesis [65], it is reasonable to speculate that chromatin landscapes would be moderately regulated by those factors, and that such effects could be responsible for the anticancer actions

of VPA. Furthermore, we detected significant enrichment of the chromatin remodeler CHD8, which genome-wide association studies have shown to be strongly associated with the risk of autism spectrum disorder in humans and in mouse models [66–68]. A well-known adverse effect of VPA is that exposure during pregnancy increases the risk of autism in children [69], although the molecular mechanism underlying this process remains poorly understood. Our results suggest that prenatal use of VPA affects the activity of CHD8, thereby perturbing the target genes involved in neural development. Interestingly, lower CHD8 binding enrichment was observed for the genes perturbed after treatments with the HDAC inhibitor vorinostat (Fig. 2c). This suggests that the potential key role of CHD8 in VPA-induced autism is HDAC-independent, an idea that is also supported by the lack of evidence showing a relationship between maternal use of vorinostat and neonatal autism. These findings demonstrate that our proposed approach has the potential to elucidate the MoAs involved in adverse effects of chemicals, and could therefore identify possible preventive strategies.

Discussion

In this paper, we present a novel computational approach for elucidating MoAs, focused on pivotal TFs, using large-scale data sets of chemically induced DEGs. This method enables estimation of the efficacies and side effects of given chemicals. In the proposed approach, we identified the TFs that organize the expression of chemically induced DEGs before addressing the associations with diseases based on gene/protein–disease databases. We also tested a method that did not consider key TFs involved in MoAs; in that approach, chemical–disease associations were defined based on the commonalities of chemically induced and disease-specific DEGs. In terms of accuracy, the performance of the proposed approach was superior. This is likely because gene expression changes are the final outcome of complex pharmacophysiological cascades; consequently, direct comparison of chemically induced DEGs with disease-specific DEGs will be influenced by secondary effects and other unknown factors. By contrast, TFs that integrate chemically induced DEGs are likely to be the direct targets of the corresponding compounds, or at least more proximal to them. Indeed, AR and ESR1 were identified as the targets of the sex hormones testosterone and estradiol, respectively, whereas HDACs and BRDs were suppressed upon treatment with the anti-tumor drugs vorinostat and JQ1. In addition, we found that TP53/TP63 and MYCs integratively organized the expression of the up- and down-regulated genes that were responsive to cisplatin administration, respectively, revealing these proteins as pivotal factors in the MoAs of cisplatin's anti-tumor activities. Thus, in the context of pharmaceutical research, it might be only part of a complex system: ChIPEA is useful for elucidating MoAs from the epigenetic standpoint by revealing pivotal factors within the black box between input (drug administration) and output (gene expression changes). Our proposed method is easily performed using the "Enrichment Analysis" tool from the ChIP-Atlas website (Additional file 1: Fig. S3), where one can identify TFs simply by submitting a list of the DEGs that were identified after drug administration.

Genome-wide identification of TF binding sites based on inference of binding motifs is widely used to understand transcriptional regulation [70]. Previously, a method was proposed to model gene regulatory relationships based on predicted TF binding motifs

[71]. On the other hand, we utilized sets of target genes for each TF, which were constructed based on actual ChIP-seq experimental data in a motif-independent manner [27]. Relative to motif-based methods, our approach, which is based on ChIP-seq data, has the advantage of taking into consideration the actual TF binding sites within a specific cellular state. It is important to note that TFs do not always bind to specific binding motifs, even those that are statistically well-defined. Similarly, it is not uncommon for a TF to bind to sites that totally differ from its modeled motifs. This observation may partially explain the undesirable results that are obtained when using a DNA sequence motif-based method to estimate chemical–TF associations for chemically induced genes (Additional file 1: Fig. S4) [72], though these motif-based methods are of value when ChIP-seq data for specific TFs cannot be obtained from publicly available resources.

To predict chemical–target interactions, chemogenomics methods have been developed based on the compound molecular structures and protein sequence motifs or structural features [4–8]. When aiming to find proteins that directly interact with given ligands, these methods are generally efficient, with high prediction accuracy as long as the structures of the proteins and compounds have been well characterized. As a complementary approach, our method is applicable to proteins and compounds with less defined structures, as we do not focus primarily on direct drug–target interactions. Our approach only requires transcriptome analysis data for given chemicals, and is therefore capable of providing novel insights into chemical–protein associations fully based on biological experiments and public ChIP-seq data. In recent years, supervised machine learning algorithms have been used to predict drug targets and novel indications [9, 10]. Predictions made using this kind of approach can be very accurate if the system is provided with sufficient unbiased knowledge about specific drugs, proteins, and diseases. Such machine learning approaches will become more powerful when combined with TF-based knowledge obtained from our proposed method, which will allow gene-regulatory networks based on actual ChIP-seq experiments to contribute more accurate and explainable predictions.

Although the ChIPEA procedure itself is not novel, we show in this paper that the proposed ChIPEA-based approach is capable of identifying key regulators that are the direct targets of, or are primarily involved in the MoAs of, given bioactive compounds, implying that it could be used to make important contributions to the pharmaceutical field. For instance, ChIPEA could be used to analyze data from high-throughput expression screening of thousands of chemicals. The conventional approaches for screening candidate compounds include Gene Ontology and pathway enrichment analyses, which can be used to identify common features among chemically induced genes. In addition, our proposed approach provides insight into the regulatory mechanism acting upstream of these genes, allowing identification of drug candidates targeting the desired TFs. Furthermore, our method could also be applied to the transcriptome data of unapproved drugs, including compounds under development and those that failed to be approved in clinical trials. Identification of TFs primarily involved in MoAs, together with the factors associated with potential side effects, could shed light on the potential utility of repositories of compounds hoarded in pharmaceutical industries. In addition, our method could be used to analyze approved drugs, including those with more or less well-defined MoAs. Other

possible applications include drugs composed of unidentified ingredients, such as traditional herbal medicines. Potential therapeutic effects, side effects, and novel targets and actions can be inferred by identifying TFs involved in unexpected physiological pathways.

Finally, we wish to discuss the limitations and the extensibility of the proposed ChIPEA-based method. In this study, we considered only TF binding sites adjacent to the chemically induced genes (target range: transcription start site ± 5 kb). In general, genes are regulated by a complex series of enhancers at a short or long distance, of which a fair proportion are considered to fall outside the target range of ChIPEA. Thus, it would be informative to use chromatin accessibility data (DNase-seq and ATAC-seq) to analyze TF binding at longer ranges from chemically perturbed genes. Furthermore, genome-wide chromatin conformation capture (Hi-C) data is most suited to identifying chemically perturbed genes and the long-range TF binding sites that are sometimes observed for cis-regulatory elements. If ATAC-seq data are available for a broad coverage of chemicals, a more direct type of input data for ChIPEA would be the open chromatin regions, rather than the genes, rearranged by chemical administration. In our proposed method, the gene–TF matrix needs to be binarily abstracted as "binding" or "non-binding", which does not take into account the broad range of binding affinities observed. Development of a weighted enrichment analysis method that includes an algorithm to factor in the binding affinity between each TF and specific gene loci, such as the statistical values obtained using the MACS2 peak calling procedure (q-values), and the distance between genes and TF binding sites [73] would address this issue. It should be kept in mind that ChIPEA focuses on the "binding" patterns of TFs to chemically induced genes; however, "binding" does not necessarily mean that there is a regulatory relationship. It would also be best to experimentally validate the predicted results for pivotal TFs in order to determine their relationships within regulatory networks. Finally, because we are using experimental ChIP-seq data, TFs lacking public ChIP-seq data cannot be analyzed using ChIPEA. However, because the number of ChIP-seq experiments is steadily increasing, the ChIPEA-based approach will become increasingly powerful in the future.

Conclusions

In this paper, we introduced a computational approach to elucidating the epigenetic landscape of drug responses, in which large-scale public ChIP-seq experiment data were analyzed to identify key TFs acting upstream of chemically induced genes. Chemical–TF associations were predicted by ChIPEA of chemically induced expression profiles and validated using a chemical–protein association database. Furthermore, chemical–TF–disease associations were constructed by linking the TFs with known disease-associated proteins. Together, our findings demonstrate that ChIPEA using public ChIP-seq data is an efficient way to identify master regulators involved in MoAs from an epigenetic perspective. Therefore, this approach is a powerful means of predicting chemical–TF and chemical–disease associations in a biologically interpretable manner, outperforming methods that do not consider information about the TFs involved in MoAs. Our proposed approach could be used to further understand the MoAs of candidate drugs, as well as to discover unexpected therapeutic effects and side effects of approved drugs.

Methods

Datasets

Chemically perturbed DEGs from CTD

Gene symbols of genes up- and down-regulated by environmental chemicals and medical drugs were obtained from CTD (chemical–gene interaction; download link, http://ctdbase.org/reports/CTD_chem_gene_ixns.csv.gz; downloaded on May 15th, 2020). CTD is a community-supported genomic resource that provides manually annotated associations among chemicals, genes/proteins, and diseases [20]. In the chemical–gene interaction database, each chemical–gene interaction is addressed in a declarative statement and qualified by a degree: increases, decreases, affects, or does not affect. We used only gene expression profiles of human cells in response to chemical administration with the interaction type "*C* (analog) results in increased/decreased expression of *G* mRNA/protein", where *C* and *G* represent a chemical and a gene, respectively. The number of DEGs in response to treatment with each chemical varied widely, from 1 to over 6,000. We extracted expression profiles with more than ten each of up- and down-regulated DEGs, yielding a total of 890 gene expression profiles related to 434 chemicals (Additional file 2: Table S1).

Genome-wide TF binding experimental data from ChIP-Atlas

We obtained information about genome-wide TF binding sites from ChIP-Atlas, an integrative database that covers almost all public ChIP-seq data submitted to the NCBI SRA [27]. The metadata of all experiments, such as names of antigens and cellular states, are manually curated according to commonly or officially adopted nomenclature. The sequence data are processed with a unified pipeline in which sequenced reads are aligned to a reference genome with Bowtie2 and subjected to peak calling with MACS2. We retrieved full sequencing data from 13,558 experiments (Additional file 2: Table S2) that identified 170,067,307 binding sites (peaks were called with MACS2; q-value $< 1 \times 10^{-10}$) of 997 TFs in the human genome (download link, http://dbarchive.biosciencedbc.jp/kyushu-u/hg19/allPeaks_light/allPeaks_light.hg19.05.bed.gz; genome version, hg19; downloaded on May 15th, 2020). It is worth noting that all of the binding data in ChIP-Atlas were determined experimentally and therefore are not binding-motif dependent.

Gene/protein–disease association

Gene/protein–disease association data were acquired from DisGeNET (version, v7.0; download link, https://www.disgenet.org/static/disgenet_ap1/files/downloads/curated_gene_disease_associations.tsv.gz; downloaded on June 4th, 2020), a large collection of genes and variants associated with human diseases [28]. DisGeNET integrates data from expert-curated repositories and the scientific literature. We retrieved only manually curated data labeled with referenced PubMed IDs, yielding 77,524 gene/protein–disease associations involving 9,334 genes/proteins and 7,687 diseases.

Disease-specific DEGs from CREEDS

Disease-specific gene expression profiles were constructed based on gene expression profiles in CREEDS (manual disease signatures v1.0; download link, http://amp.pharm.mssm.edu/CREEDS/download/disease_signatures-v1.0.json; download on June 16th,

2020), a crowdsourcing project aimed at annotating and reanalyzing a large number of gene expression profiles from Gene Expression Omnibus [35]. The gene expression profiles were associated with scores calculated using the characteristic direction method, in which gene expression levels in diseased tissues were compared with healthy controls. Genes with positive or negative expression scores were considered to be up- or down-regulated DEGs, respectively, yielding 554 gene expression profiles associated with 235 diseases.

Standard chemical–protein interactome
Compound–protein interaction data were acquired from KEGG DRUG (download link, ftp://ftp.biosciencedbc.jp/archive/kegg-medicus/LATEST/kegg_medicus_drug.csv.zip; downloaded on July 1st, 2019), a comprehensive collection of approved drugs and their target information, including 11,550 chemical–protein interactions involving 1,458 chemicals and 768 proteins [29].

Standard chemical–disease associations
Chemical–disease association data were acquired from CTD (chemical–disease interactions; download link, http://ctdbase.org/reports/CTD_chemicals_diseases.tsv.gz; downloaded on June 15th, 2020). In total, we obtained 219,317 chemical–disease associations involving 9,855 chemicals and 3,244 diseases, curated from 78,582 papers.

The proposed ChIPEA-based approach
Identification of key TFs that organize the expression of chemically induced genes using ChIPEA
Taking full advantage of this enormous quantity of data, we performed enrichment analysis termed ChIPEA to profile TFs whose binding sites were enriched around chemically induced genes of interest. In particular, starting with the gene symbols of up- and down-regulated genes induced by a query chemical, we counted the overlaps between the transcription start site ± 5 kb regions of chemically induced DEGs and peak-call data of all TF-related experiments archived in ChIP-Atlas, using the "intersect" command of BEDTools2 (version, v2.23.0) [74]. Enrichment scores ($-\log_{10}[p\text{-values}]$) were calculated using the two-tailed Fisher's exact probability test, with the null hypothesis that the two data sets (up- and down-regulated genes) overlap with the ChIP-seq peak-call data in the same proportion; fold enrichment values were returned at the same time. If a chemical–TF association was given by multiple ChIP-seq experiments, the highest enrichment score was adopted. This ChIPEA procedure was proposed previously [27].

Relating the chemical–TF matrix identified using ChIPEA to TF–disease associations
Data for gene/protein–disease associations derived from DisGeNET (formula 1; where P_m and D_n represent a protein and a disease, respectively) were correlated with the chemical–TF associations determined using ChIPEA (formula 2; where Ci, Tj, and E_{ij} represent a chemical, a TF, and an enrichment score, respectively) when T_j was also included in DisGeNET as P_m (formula 3). The enrichment scores calculated by ChIPEA were also used to evaluate the probability of each chemical–disease prediction (formula 4).

$$P_m - D_n \tag{1}$$

$$C_i - T_j - E_{ij} \tag{2}$$

$$T_j = P_m \tag{3}$$

$$C_i - T_j - D_n - E_{ij} \tag{4}$$

If a chemical–disease pair was predicted via multiple TFs, the highest enrichment score was adopted.

Calculation of global AUROC and AUPR

After evaluating their validity, all predicted chemical–TF or chemical–disease associations were arranged into a single $m \times n$ matrix consisting of enrichment scores with correctness information, where m was the total number of chemicals and n was the number of TFs or diseases, respectively. We then stored the maximum value of enrichment scores within each column (TF or disease) into a vector with n elements. We generated ROC and PR curves and summarized the results into global AUROC and AUPR scores as described in the Results section.

Baseline methods

Motif-based enrichment analysis using the regulatory genomics toolbox [72]

In the motif matching step, a set of TF motifs identified from several main repositories was compared to the genomic regions of chemically induced genes without taking into account the DNase-seq signals. Subsequently, motif enrichment analysis was performed for each chemical (download link, http://www.regulatory-genomics.org/wp-content/uploads/2017/03/RGT_MotifAnalysis_FullSiteTest.tar.gz; downloaded on July 12th, 2021). In particular, Fisher's exact test was used in order to determine whether the chemically induced genes were enriched for particular TFs. Enrichment scores ($-\log_{10}[p$-values]) were calculated using the two-tailed Fisher's exact probability test, with the null hypothesis that the two data sets (up- and down-regulated genes) overlap with the TF motif data in the same proportion; fold enrichment values were returned at the same time. Testosterone-, vorinostat-, and JQ1-induced genes were obtained from CTD database.

DEG-connected method for predicting chemical–disease associations

Chemically induced and disease-specific expression changes in 28,268 genes were classified as up-regulated, down-regulated, or non-DEG; RefSeq genes were obtained from the UCSC genome annotation database of the human genome (download link, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refFlat.txt.gz; genome, hg19; downloaded on July 7th, 2020). All chemical and disease profiles were further arranged into two-by-two cross tabulations for each of the chemical–disease pairs in two ways ("DEGs vs. non-DEGs" and "up- vs. down-regulated genes"), as illustrated in Table 1.

Zou *et al. BMC Bioinformatics*   (2022) 23:51

Page 16 of 20

Table 1 Cross tabulation showing the frequency distribution of chemically induced and disease-specific genes, with C and D representing a chemical and a disease, respectively

		Disease *D*				Disease *D*	
		DEGs	non-DEGs			Up	Down
Chemical *C*	DEGs	n1	n2	Chemical *C*	Up	n1	n2
	non-DEGs	n3	n4		Down	n3	n4
	"DEGs vs. non-DEGs" comparison				**"up- vs. down-regulated genes" comparison**		

The chemical–disease associations were evaluated based on *p*-values calculated using the two-tailed Fisher's exact probability test with the null hypothesis that the comparative gene expression patterns in response to a given *C* (chemical) and *D* (disease) (*n1*, *n2*, *n3*, and *n4*) were uniformly distributed. Chemical–disease pairs with smaller *p*-values were considered to be more firmly associated.

Abbreviations

AUPR: Area under the precision-recall curve; AR: Androgen receptor; AUROC: Area under the receiver operating characteristic curve; ChIPEA: Chromatin immunoprecipitation sequencing-based enrichment analysis; ChIP-seq: Chromatin immunoprecipitation sequencing; CMap: Connectivity Map; CREEDS: Crowd Extracted Expression of Differential Signatures; CTD: Comparative Toxicogenomics Database; DEGs: Differentially expressed genes; DHODH: Dihydroorotate dehydrogenase; ES: Enrichment score; ESR1: Estrogen receptor 1; ESR2: Estrogen receptor 2; KEGG DRUG: Kyoto Encyclopedia of Genes and Genomes DRUG database; LINCS: National Institute of Health Library of Integrated Network-Based Cellular Signatures; MoA: Mode of action; PHTS: Phenotype-based high-throughput screening; PR: Precision-recall; RARA: Retinoic acid receptor alpha; ROC: Receiver operating characteristic; TF: Transcription factor; TG-GATEs: Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system; VPA: Valproic acid.

Supplementary Information

Additional file 1: Fig. S1 Comparison of gene expression profiles and annotated proteins contained in different databases. a, c Venn diagram showing genes shared between the chemical–gene (CTD and L1000) and disease–gene (CREEDS) association databases (a), and proteins shared between the ChIP-seq experiment (ChIP-Atlas) and gene/protein–disease association (DisGeNET) databases (c). **b, d** Bar charts showing the frequencies with which the genes appeared in CTD and CREEDS, with L1000 genes colored in orange (CTD) and green (CREEDS) (b), and the frequencies with which proteins were defined as disease-associated proteins in DisGeNET, with ChIP-Atlas TFs colored in blue (d). **Fig. S2 Factors potentially affecting the distribution of AUROC scores.** AUROC scores, sorted according to the highest enrichment score for each (**a, c**) chemical and (**b, d**) the number of DEGs that were used to predict (**a, b**) chemical–TF and (**c, d**) chemical–disease associations using the proposed ChIPEA-based approach. **Fig. S3 Visual manual for GUI-based ChIPEA. a.** Submission form for ChIPEA on the website. GUI-based ChIPEA is provided on the ChIP-Atlas website (termed "Enrichment Analysis" tool; https://chip-atlas.org/enrichment_analysis). When used to identify pivotal TFs involved in drug MoAs, genome assembly should be set as "hg19/hg38" (hg19 was used in this paper). "TFs and others" needs to be selected in panel "1. Antigen Class". "2. Cell type Class" and "3. Threshold for Significance" may be changed by the user according to demand. The "4. Enter dataset A" dialog box is to be filled in with the list of up-regulated genes, and the box "5. Enter dataset B" is for down-regulated genes. After specifying the "Distance range from TSS" in the "6. Analysis description" panel, the user can press the "Submit" button to submit the parameters to the server, and ChIPEA will initialize immediately. **b.** Interpretation of the results. The overlaps between the genomic loci (originating from panels 4 and 5 of the submission form) and reference peak call data (specified on upper panels 1–3 of the submission form) are counted using the bedtools intersect command (BedTools2; ver 2.23.0). The results are returned in html and tsv format. *p*-values are calculated using a two-tailed Fisher's exact probability test. The null hypothesis is that the intersection between the reference peaks and the data submitted in panel 4 occurs at the same proportion as for the data in panel 5 of the submission form. *q*-values are calculated using the Benjamini & Hochberg method. Fold enrichment was calculated by dividing the result from column 6 by the result from column 7 for a given row of data. If the ratio > 1, the rightmost column is "TRUE", meaning that the protein from column 3 is more likely to bind to the variable from panel 4 than to that from panel 5. **Fig. S4 Identification of target TFs using a motif analysis-based method.** TF binding profiles of three chemicals predicted by motif analysis using DEGs from CTD (testosterone, vorinostat, and JQ1). Dots indicate individual TFs and are colored orange only if they match the chemical–target associations recorded in KEGG DRUG database. AUROC scores are shown in the upper right corner of the volcano plot of each chemical.

Additional file 2: Table S1. List of CTD chemicals. Column 1, CTD ID; column 2, chemical name. **Table S2. List of ChIP-seq experiments.** Column 1, Experiment ID; column 2, ChIP-seq antigen. **Table S3. Predicted chemical–TF associations using ChIPEA for Fig. 2c. a** Testosterone; **b** Estradiol; **c** Tamibarotene; **d** Vorinostat; **e** JQ1 compound.

Availability of data and materials
We used freely available datasets, and the identifiers of the datasets are provided in Table S1 and S2 in Additional file 2; specific version numbers or dates of access/download are indicated in the Methods section. The code related to the proposed approach is available free of charge via the Internet at https://github.com/zouzhaonan/moa_tfea/.

Declarations

Ethics approval and consent to participate
Not applicable.

Content for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan. [2]Kyoto University Graduate Program for Medical Innovation, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. [3]Kyoto University Graduate Division, Yoshida-Nihonmatsu-cho, Sakyo-ku, Kyoto 606-8501, Japan. [4]Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan. [5]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan.

References
1. Etzion Y, Muslin AJ. The application of phenotypic high-throughput screening techniques to cardiovascular research. Trends Cardiovasc Med. 2009;19(6):207–12. https://doi.org/10.1016/j.tcm.2009.12.006.
2. Gao J, Xu Y, Yang Y, et al. Identification of upregulators of human ATP-binding cassette transporter A1 via high-throughput screening of a synthetic and natural compound library. J Biomol Screen. 2008;13(7):648–56. https://doi.org/10.1177/1087057108320545.
3. Mayer TU, Kapoor TM, Haggarty SJ, King RW, Schreiber SL, Mitchison TJ. Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. Science. 1999;286(5441):971–4. https://doi.org/10.1126/science.286.5441.971.
4. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11(1):4. https://doi.org/10.1186/s13321-018-0325-4.
5. Stanton DT. QSAR and QSPR model interpretation using partial least squares (PLS) analysis. Curr Comput Aided Drug Des. 2012;8(2):107–27. https://doi.org/10.2174/157340912800492357.
6. Saxena AK, Prathipati P. Comparison of MLR, PLS and GA-MLR in QSAR analysis. SAR QSAR Environ Res. 2003;14(5–6):433–45. https://doi.org/10.1080/10629360310001624015.
7. Liu P, Long W. Current mathematical methods used in QSAR/QSPR studies. Int J Mol Sci. 2009;10(5):1978–98. https://doi.org/10.3390/ijms10051978.
8. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol. 2009;20(4):429–36. https://doi.org/10.1016/j.copbio.2009.08.003.

9.   Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics. 2008;24(19):2149–56. https://doi.org/10.1093/bioinformatics/btn409.

10. Keiser MJ, Setola V, Irwin JJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175–81. https://doi.org/10.1038/nature08506.

11. Fang J, Cai C, Wang Q, Lin P, Zhao Z, Cheng F. Systems pharmacology-based discovery of natural products for precision oncology through targeting cancer mutated genes. CPT Pharmacometr Syst Pharmacol. 2017;6(3):177–87. https://doi.org/10.1002/psp4.12172.

12. Cheng J, Yang L, Kumar V, Agarwal P. Systematic evaluation of connectivity map for disease indications. Genome Med. 2014;6(12):540. https://doi.org/10.1186/s13073-014-0095-1.

13. Cheng J, Xie Q, Kumar V, et al. Evaluation of analytical methods for connectivity map data. Pac Symp Biocomput. 2013. https://doi.org/10.1142/9789814447973_0002.

14. Dudley JT, Sirota M, Shenoy M, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med. 2011;3(96):96ra76. https://doi.org/10.1126/scitranslmed.3002648.

15. Kosaka T, Nagamatsu G, Saito S, Oya M, Suda T, Horimoto K. Identification of drug candidate against prostate cancer from the aspect of somatic cell reprogramming. Cancer Sci. 2013;104(8):1017–26. https://doi.org/10.1111/cas.12183.

16. van Noort V, Scholch S, Iskar M, et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. Cancer Res. 2014;74(20):5690–9. https://doi.org/10.1158/0008-5472.CAN-13-3540.

17. Igarashi Y, Nakatsu N, Yamashita T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. 2015;43(Database issue):D921-7. https://doi.org/10.1093/nar/gku955.

18. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35. https://doi.org/10.1126/science.1132939.

19. Duan Q, Flynn C, Niepel M, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42(Web Server issue):W449-60. https://doi.org/10.1093/nar/gku476.

20. Davis AP, Grondin CJ, Johnson RJ, et al. The comparative toxicogenomics database: update 2019. Nucleic Acids Res. 2019;47(D1):D948–54. https://doi.org/10.1093/nar/gky868.

21. Wilson S, Filipp FV. A network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer. NPJ Syst Biol Appl. 2018;4:24. https://doi.org/10.1038/s41540-018-0061-4.

22. Ameratunga M, Brana I, Bono P, et al. First-in-human phase 1 open label study of the BET inhibitor ODM-207 in patients with selected solid tumours. Br J Cancer. 2020;123(12):1730–6. https://doi.org/10.1038/s41416-020-01077-z.

23. Lewin J, Soria JC, Stathis A, et al. Phase Ib trial with birabresib, a small-molecule inhibitor of bromodomain and extraterminal proteins, in patients with selected advanced solid tumors. J Clin Oncol. 2018;36(30):3007–14. https://doi.org/10.1200/JCO.2018.78.2292.

24. van Tilburg CM, Milde T, Witt R, et al. Phase I/II intra-patient dose escalation study of vorinostat in children with relapsed solid tumor, lymphoma, or leukemia. Clin Epigenet. 2019;11(1):188. https://doi.org/10.1186/s13148-019-0775-1.

25. Ning ZQ, Li ZB, Newman MJ, et al. Chidamide (CS055/HBI-8000): a new histone deacetylase inhibitor of the benzamide class with antitumor activity and the ability to enhance immune cell-mediated tumor cell cytotoxicity. Cancer Chemother Pharmacol. 2012;69(4):901–9. https://doi.org/10.1007/s00280-011-1766-x.

26. Fontaine F, Overman J, François M. Pharmacological manipulation of transcription factor protein-protein interactions: opportunities and obstacles. Cell Regen. 2015;4(1):2. https://doi.org/10.1186/s13619-015-0015-x.

27. Oki S, Ohta T, Shioi G, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep. 2018;19(12):e46255. https://doi.org/10.15252/embr.201846255.

28. Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845–55. https://doi.org/10.1093/nar/gkz1021.

29. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(Database issue):D355-60. https://doi.org/10.1093/nar/gkp896.

30. Takayama K, Inoue S. Transcriptional network of androgen receptor in prostate cancer progression. Int J Urol. 2013;20(8):756–68. https://doi.org/10.1111/iju.12146.

31. Jensen EV, Jordan VC. The estrogen receptor: a model for molecular medicine. Clin Cancer Res. 2003;9(6):1980–9.

32. Wang X, Dasari S, Nowakowski GS, et al. Retinoic acid receptor alpha drives cell cycle progression and is associated with increased sensitivity to retinoids in T-cell lymphoma. Oncotarget. 2017;8(16):26245–55. https://doi.org/10.18632/oncotarget.15441.

33. Richon VM, Emiliani S, Verdin E, et al. A class of hybrid polar inducers of transformed cell differentiation inhibits histone deacetylases. Proc Natl Acad Sci U S A. 1998;95(6):3003–7. https://doi.org/10.1073/pnas.95.6.3003.

34. Filippakopoulos P, Qi J, Picaud S, et al. Selective inhibition of BET bromodomains. Nature. 2010;468(7327):1067–73. https://doi.org/10.1038/nature09504.

35. Wang Z, Monteiro CD, Jagodnik KM, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nat Commun. 2016;7:12846. https://doi.org/10.1038/ncomms12846.

36. Fraval HN, Rawlings CJ, Roberts JJ. Increased sensitivity of UV-repair-deficient human cells to DNA bound platinum products which unlike thymine dimers are not recognized by an endonuclease extracted from Micrococcus luteus. Mutat Res. 1978;51(1):121–32. https://doi.org/10.1016/0027-5107(78)90014-3.

37. Guo J, Xu B, Han Q, et al. Ferroptosis: a novel anti-tumor action for cisplatin. Cancer Res Treat. 2018;50(2):445–60. https://doi.org/10.4143/crt.2016.572.

38. Evans DL, Dive C. Effects of cisplatin on the induction of apoptosis in proliferating hepatoma cells and nonproliferating immature thymocytes. Cancer Res. 1993;53(9):2133–9.

39. Henkels KM, Turchi JJ. Cisplatin-induced apoptosis proceeds by caspase-3-dependent and -independent pathways in cisplatin-resistant and -sensitive human ovarian cancer cell lines. Cancer Res. 1999;59(13):3077–83.
40. Aubrey BJ, Strasser A, Kelly GL. Tumor-suppressor functions of the TP53 pathway. Cold Spring Harb Perspect Med. 2016;6(5): a026062. https://doi.org/10.1101/cshperspect.a026062.
41. Flores ER, Sengupta S, Miller JB, et al. Tumor predisposition in mice mutant for p63 and p73: evidence for broader tumor suppressor functions for the p53 family. Cancer Cell. 2005;7(4):363–73. https://doi.org/10.1016/j.ccr.2005.02.019.
42. Dang CV. MYC on the path to cancer. Cell. 2012;149(1):22–35. https://doi.org/10.1016/j.cell.2012.03.003.
43. Goldenberg MM. Leflunomide, a novel immunomodulator for the treatment of active rheumatoid arthritis. Clin Ther. 1999;21(11):1837–52. https://doi.org/10.1016/S0149-2918(00)86732-6.
44. Kraan MC, Reece RJ, Barg EC, et al. Modulation of inflammation and metalloproteinase expression in synovial tissue by leflunomide and methotrexate in patients with active rheumatoid arthritis. Findings in a prospective, randomized, double-blind, parallel-design clinical trial in thirty-nine patients at two centers. Arthritis Rheum. 2000;43(8):1820–30. https://doi.org/10.1002/1529-0131(200008)43:8%3c1820::AID-ANR18%3e3.0.CO;2-D.
45. Cherwinski HM, Byars N, Ballaron SJ, Nakano GM, Young JM, Ransom JT. Leflunomide interferes with pyrimidine nucleotide biosynthesis. Inflamm Res. 1995;44(8):317–22. https://doi.org/10.1007/BF01796261.
46. Cao WW, Kao PN, Chao AC, Gardner P, Ng J, Morris RE. Mechanism of the antiproliferative action of leflunomide. A77 1726, the active metabolite of leflunomide, does not block T-cell receptor-mediated signal transduction but its antiproliferative effects are antagonized by pyrimidine nucleosides. J Heart Lung Transplant. 1995;14(6 pt 1):1016–30.
47. Davis JP, Cain GA, Pitts WJ, Magolda RL, Copeland RA. The immunosuppressive metabolite of leflunomide is a potent inhibitor of human dihydroorotate dehydrogenase. Biochemistry. 1996;35(4):1270–3. https://doi.org/10.1021/bi952168g.
48. Buettner R, Morales C, Wu X, et al. Leflunomide synergizes with gemcitabine in growth inhibition of PC cells and impairs c-Myc signaling through PIM kinase targeting. Mol Ther Oncolytics. 2019;14:149–58. https://doi.org/10.1016/j.omto.2019.04.006.
49. Ringshausen I, Oelsner M, Bogner C, Peschel C, Decker T. The immunomodulatory drug Leflunomide inhibits cell cycle progression of B-CLL cells. Leukemia. 2008;22(3):635–8. https://doi.org/10.1038/sj.leu.2404922.
50. Peng LN, Deng XY, Gan XX, et al. Targeting of TLE3 by miR-3677 in human breast cancer promotes cell proliferation, migration and invasion. Oncol Lett. 2020;19(2):1409–17. https://doi.org/10.3892/ol.2019.11241.
51. Yang RW, Zeng YY, Wei WT, et al. TLE3 represses colorectal cancer proliferation by inhibiting MAPK and AKT signaling pathways. J Exp Clin Cancer Res. 2016;35(1):152. https://doi.org/10.1186/s13046-016-0426-8.
52. Liang C, Li J, Lu C, et al. HIF1α inhibition facilitates Leflunomide-AHR-CRP signaling to attenuate bone erosion in CRP-aberrant rheumatoid arthritis. Nat Commun. 2019;10(1):4579. https://doi.org/10.1038/s41467-019-12163-z.
53. Park MH, Hong JT. Roles of NF-kappaB in cancer and inflammatory diseases and their therapeutic approaches. Cells. 2016;5(2):15. https://doi.org/10.3390/cells5020015.
54. Chen HZ, Tsai SY, Leone G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. Nat Rev Cancer. 2009;9(11):785–97. https://doi.org/10.1038/nrc2696.
55. Bartlett RR, Schleyerbach R. Immunopharmacological profile of a novel isoxazol derivative, HWA 486, with potential antirheumatic activity—I. Disease modifying action on adjuvant arthritis of the rat. Int J Immunopharmacol. 1985;7(1):7–18. https://doi.org/10.1016/0192-0561(85)90003-7.
56. Sen R, Baltimore D. Multiple nuclear factors interact with the immunoglobulin enhancer sequences. Cell. 1986;46(5):705–16. https://doi.org/10.1016/0092-8674(86)90346-6.
57. Ren A, Fu G, Qiu Y, Cui H. Leflunomide inhibits proliferation and tumorigenesis of oral squamous cell carcinoma. Mol Med Rep. 2017;16(6):9125–30. https://doi.org/10.3892/mmr.2017.7755.
58. Bahr HI, Toraih EA, Mohammed EA, Mohammad HM, Ali EA, Zaitone SA. Chemopreventive effect of leflunomide against Ehrlich's solid tumor grown in mice: effect on EGF and EGFR expression and tumor proliferation. Life Sci. 2015;141:193–201. https://doi.org/10.1016/j.lfs.2015.10.003.
59. Zhu S, Yan X, Xiang Z, Ding HF, Cui H. Leflunomide reduces proliferation and induces apoptosis in neuroblastoma cells in vitro and in vivo. PLoS ONE. 2013;8(8): e71555. https://doi.org/10.1371/journal.pone.0071555.
60. Romoli M, Mazzocchetti P, D'Alonzo R, et al. Valproic acid and epilepsy: from molecular mechanisms to clinical evidences. Curr Neuropharmacol. 2019;17(10):926–46. https://doi.org/10.2174/1570159X17666181227165722.
61. Gottlicher M, Minucci S, Zhu P, et al. Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells. EMBO J. 2001;20(24):6969–78. https://doi.org/10.1093/emboj/20.24.6969.
62. Lin T, Ren Q, Zuo W, et al. Valproic acid exhibits anti-tumor activity selectively against EGFR/ErbB2/ErbB3-coexpressing pancreatic cancer via induction of ErbB family members-targeting microRNAs. J Exp Clin Cancer Res. 2019;38(1):150. https://doi.org/10.1186/s13046-019-1160-9.
63. Ma XJ, Wang YS, Gu WP, Zhao X. The role and possible molecular mechanism of valproic acid in the growth of MCF-7 breast cancer cells. Croat Med J. 2017;58(5):349–57. https://doi.org/10.3325/cmj.2017.58.349.
64. Chen Y, Tsai YH, Tseng SH. Combined valproic acid and celecoxib treatment induced synergistic cytotoxicity and apoptosis in neuroblastoma cells. Anticancer Res. 2011;31(6):2231–9.
65. Nair SS, Kumar R. Chromatin remodeling in cancer: a gateway to regulate gene transcription. Mol Oncol. 2012;6(6):611–9. https://doi.org/10.1016/j.molonc.2012.09.005.
66. Talkowski ME, Rosenfeld JA, Blumenthal I, et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell. 2012;149(3):525–37. https://doi.org/10.1016/j.cell.2012.03.028.
67. O'Roak BJ, Vives L, Girirajan S, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246–50. https://doi.org/10.1038/nature10989.
68. Katayama Y, Nishiyama M, Shoji H, et al. CHD8 haploinsufficiency results in autistic-like phenotypes in mice. Nature. 2016;537(7622):675–9. https://doi.org/10.1038/nature19357.

69. Christensen J, Gronborg TK, Sorensen MJ, et al. Prenatal valproate exposure and risk of autism spectrum disorders and childhood autism. JAMA. 2013;309(16):1696–703. https://doi.org/10.1001/jama.2013.2270.

70. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):D87-92. https://doi.org/10.1093/nar/gkz1001.

71. Balwierz PJ, Pachkov M, Arnold P, Gruber AJ, Zavolan M, van Nimwegen E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res. 2014;24(5):869–84. https://doi.org/10.1101/gr.169508.113.

72. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods. 2016;13(4):303–9. https://doi.org/10.1038/nmeth.3772.

73. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci U S A. 2009;106(51):21521–6. https://doi.org/10.1073/pnas.0904863106.

74. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.

Publisher's Note

Additional file 1: Figure S1

Additional file 1: Figure S2

Additional file 1: Figure S3

Additional file 1: Figure S4

主論文 1（Zou, *BMC Bioinform.*）の補足資料は下記 URL より入手可能である。

BMC Bioinformatics ホームページ：

https://doi.org/10.1186/s12859-022-04571-8

京都大学ファイル共有システム：

https://fsv.iimc.kyoto-u.ac.jp/public/Ie0sgAxSKYlAYIgBRwuDxif4kWsKqw5f4j0ixBKxy-40

主論文 2

ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data

Zhaonan Zou [1,2,3], Tazro Ohta [4], Fumihito Miura [5] and Shinya Oki [1,6,*]

[1]Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan, [2]Kyoto University Graduate Program for Medical Innovation, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan, [3]Kyoto University Graduate Division, Yoshida-Nihonmatsu-cho, Sakyo-ku, Kyoto 606-8501, Japan, [4]Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Yata 1111, Mishima, Shizuoka 411-8540, Japan, [5]Department of Biochemistry, Kyushu University Graduate School of Medical Sciences, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan and [6]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

ABSTRACT

ChIP-Atlas (https://chip-atlas.org) is a web service providing both GUI- and API-based data-mining tools to reveal the architecture of the transcription regulatory landscape. ChIP-Atlas is powered by comprehensively integrating all data sets from high-throughput ChIP-seq and DNase-seq, a method for profiling chromatin regions accessible to DNase. In this update, we further collected all the ATAC-seq and whole-genome bisulfite-seq data for six model organisms (human, mouse, rat, fruit fly, nematode, and budding yeast) with the latest genome assemblies. These together with ChIP-seq data can be visualized with the Peak Browser tool and a genome browser to explore the epigenomic landscape of a query genomic locus, such as its chromatin accessibility, DNA methylation status, and protein–genome interactions. This epigenomic landscape can also be characterized for multiple genes and genomic loci by querying with the Enrichment Analysis tool, which, for example, revealed that inflammatory bowel disease-associated SNPs are the most significantly hypo-methylated in neutrophils. Therefore, ChIP-Atlas provides a panoramic view of the whole epigenomic landscape. All datasets are free to download via either a simple button on the web page or an API.

GRAPHICAL ABSTRACT
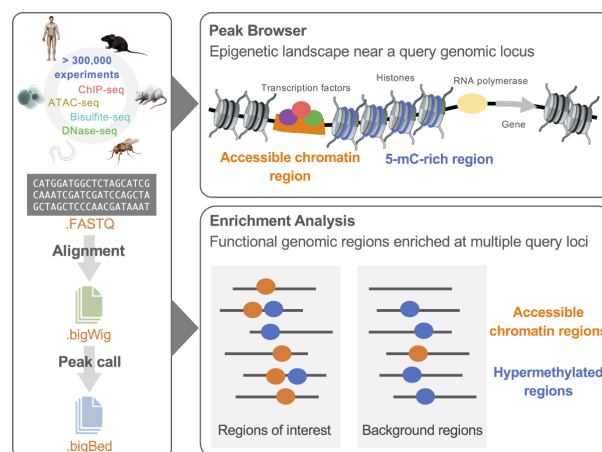
INTRODUCTION

In the past decade, despite the increasing number of high-throughput sequencing experiments, the secondary use of the obtained raw data has required complex and large-scale computational processing, and thus most data are still being hoarded. Since 2015, we have been comprehensively collecting, analyzing and integrating almost all chromatin immunoprecipitation sequencing (ChIP-seq) (1) and DNase-seq (2)—a method for profiling chromatin regions accessible to DNase—data derived from six representative model organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*) archived at the Sequence Read Archive (SRA). SRA is the largest publicly available data

---

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

repository that accepts submissions of high-throughput sequencing data, which is maintained by NCBI, EBI, and DDBJ. Data-mining tools powered by these data were provided through our web server, ChIP-Atlas (https://chip-atlas.org) (3), for visualization of assembled peak data and enrichment analysis for given genomic loci to identify transcription factor (TF) binding and histone modification status. ChIP-Atlas is powered by dedicated manual curation and annotation of experimental metadata, and a uniform data process pipeline to reveal the complex architecture of the transcription regulatory landscape.

Information on TF binding and histone marks is, however, still insufficient to fully understand the regulatory systems for gene expression control because an absence of nucleosomes and low methylation levels also characterize active promoters, enhancers, and other gene regulatory sequences (4–7). To detect accessible chromatin regions, several experimental methods including DNase-seq, formaldehyde-assisted isolation of regulatory elements (8), and assay for transposase-accessible chromatin with sequencing (ATAC-seq) (9) have been developed. Among these, ATAC-seq identifies accessible chromatin regions based on their increased accessibility to Tn5 transposase integration and is now predominantly used given its advantages of technical ease and sensitivity to a small number of cells. In contrast, bisulfite sequencing (Bisulfite-seq) is a well-established protocol to detect methylated cytosines in genomic DNA, which employs a chemical method that selectively deaminates unmodified cytosine to uracil while leaving 5-methylcytosine intact before DNA sequencing (10).

Here, we describe the ChIP-Atlas 2021 update, which collected all the ATAC-seq ($n = 66\ 104$) and whole-genome bisulfite sequencing (WGBS) data ($n = 51\ 074$) for the six representative model organisms. This update makes ChIP-Atlas, to our knowledge, the first and only web service that enables users to explore not only protein binding, but also chromatin accessibility and DNA methylation status, within a single given genomic region of interest (ROI) at the same time (Peak Browser tool). ChIP-Atlas can also be used to reveal the regulatory network involved in a batch of genomic ROIs (Enrichment Analysis tool) based on ATAC-seq and WGBS data in addition to ChIP-seq data in the previous version. ChIP-Atlas provides a panoramic view of the whole epigenomic landscape and should be of great interest to researchers in the fields of genetics and genomics, as well as those studying transcriptional regulation in general.

MATERIALS AND METHODS

Source data obtained from SRA

We obtained sample metadata from the NCBI BioSample database FTP server (ftp://ftp.ncbi.nlm.nih.gov/biosample). We also retrieved the metadata of SRA experiments (accession number with the prefix SRX, ERX or DRX; hereafter SRXs), such as library preparation or used sequencing instrument models, from the NCBI SRA FTP server (ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Metadata). ChIP-Atlas uses ATAC-seq and WGBS SRXs that meet the following criteria: LIBRARY_STRATEGY is 'ATAC-seq' or 'Bisulfite-seq'; LIBRARY_SOURCE is 'GENOMIC';

LIBRARY_SELECTION is 'PCR' or 'others' for ATAC-seq, and 'RANDOM' for WGBS; taxonomy_name is '*Homo sapiens*,' '*Mus musculus*', '*Rattus norvegicus*', '*Caenorhabditis elegans*', '*Drosophila melanogaster*' or '*Saccharomyces cerevisiae*'; and INSTRUMENT_MODEL includes 'Illumina', 'NextSeq' or 'HiSeq.' For ChIP-seq and DNase-seq data, refer to the previous paper on ChIP-Atlas (3).

Primary processing

Binarized sequence raw data (.sra) for each SRX were downloaded and decompressed into FASTQ format with the 'fasterq-dump' command of SRA Toolkit (ver. 2.9.4; https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.9.4/sratoolkit.2.9.4-ubuntu64.tar.gz) according to the default mode, with the exception of paired-end reads, which were decoded with the '-split-files' option. For ATAC-seq data, the subsequent alignment (with Bowtie2 [ver. 2.2.2] (11)) and peak call (with MACS2 [ver. 2.1.1; macs2 callpeak] (12)) process was the same as that for ChIP-seq and DNase-seq data, which was previously described in the first ChIP-Atlas paper. As for WGBS data, BMap (ver. 1.0; https://github.com/FumihitoMiura/Project-2/blob/master/Project-2.tar.gz) (13), a speedy aligner with small temporary file sizes, was used for alignment and MethPipe (ver. 4.1.1; https://github.com/smithlabcode/methpipe/releases/download/v4.1.1/methpipe-4.1.1.tar.gz) (14) was subsequently used as a region-caller for identifying hyper-, partially and hypo-methylated regions (MRs). After creation of the index files for all genome assemblies, we aligned the downloaded FASTQ files against the reference genomes with the '-fastq' and '-pfastq' options for single-end and paired-end reads, respectively. The alignment data (in bigWig format) containing methylation level and coverage information were generated for each SRX. Methylation level and coverage were then calculated for each CpG, and these data were next streamed into MethPipe. The 'hypermr,' 'pmd,' and 'hmr' sub-commands of MethPipe were used for identifying hyper-, partially, and hypo-MRs (in bigBed format), respectively, for each SRX according to the default mode.

Analysis of disease-associated SNPs

GWAS SNP data were downloaded from the website of UCSC genome browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/gwasCatalog.txt.gz; on 21 December 2021) (15). A BED format file was created by extending the genomic coordinates of all SNPs by 5 kb up- and downstream. Genomic regions containing SNPs were then categorized into 'inflammatory bowel disease (IBD)-associated' ($n = 393$) and 'Others' ($n = 189\ 036$) according to the disease/phenotype names listed in the 'trait' column. The IBD-associated SNPs (in BED format) were directly used as 'dataset A' for the Enrichment Analysis tool on the ChIP-Atlas web interface. Meanwhile, 3930 SNPs in the 'Others' category, which is ten times the number of IBD-associated SNPs, were randomly selected as the background data for 'dataset B'.

RESULTS

Overview of the ChIP-Atlas update

In this update, the major expansion of the experiment types results in significant growth in the number of experiments and the number of annotated functional genome regions by including all ATAC-seq and WGBS datasets archived at SRA (Figure 1A, Table 1). For unified management of the records, each experiment is assigned a unique ID (hereafter referred to as 'SRX') in ChIP-Atlas, which is exactly the same as the original SRA accession number. The number of SRXs collected in ChIP-Atlas is over 300 000 for the six organisms, which corresponds to 84.3% of the total number of ChIP-seq, ATAC-seq, DNase-seq and Bisulfite-seq SRXs in SRA for all of these organisms ($n = 362\ 121$ as of September 2021; Table 1). In this update, we adopted the latest version of genome assemblies in addition to previous ones (Table 2). Since the public release of ChIP-Atlas, the data have been updated monthly concurrent with the monthly update of the NCBI SRA metadata dump (Figure 1B), by which >100 000 ChIP-seq experimental datasets have been added since 2018 (Table 1).

We manually curated and annotated the cell types used in each experiment according to commonly or officially adopted nomenclature. The cell types were further categorized into superordinate 'cell type classes' (Figure 1C). We adapted a uniformed data process pipeline (Figure 1D; detailed in the Materials and Methods), in which the ChIP-seq, ATAC-seq, and DNase-seq data are aligned to corresponding reference genomes with Bowtie2 (11) and then subjected to peak calling with MACS2 (12); meanwhile, for WGBS data, we align the raw reads against the reference genomes using BMap before applying MethPipe for statistically calling hyper-, partially, and hypo-methylated regions (MRs). All alignment and peak-call data are freely downloadable from http://dbarchive. biosciencedbc.jp/kyushu-u/ (detailed on the documentation page of ChIP-Atlas [https://github.com/inutano/chip-atlas/wiki/#downloads_doc]), and can be browsed by users in the IGV genome browser (16) by entering the SRX ID or a given keyword (or keywords) in the corresponding Data Search page of ChIP-Atlas (Supplementary Figure S1 and Supplementary Material S1).

Example of use

Peak browser. All peak-call data recorded in ChIP-Atlas can be presented graphically with the Peak Browser tool. One can therefore easily understand not only protein–genome interactions (ChIP-seq), but also chromatin accessibility (ATAC-seq) and DNA methylation levels (WGBS), within any query genomic region of interest (ROI). To implement this tool, we integrated a large amount of peak-call data, indexed them for IGV, and constructed a web interface that externally controls IGV preinstalled on the user's machine (tested on Mac, Windows and Linux platforms). For instance, upon the specification of ChIP-seq, ATAC-seq or WGBS data of mouse spermatogonia on the web page (Figure 2A), the corresponding results are automatically streamed into IGV (Figure 2B). In this case, multiple ATAC-seq and WGBS in spermatogonia data characterize

an accessible chromatin region and hypo-MR, respectively, at the locus between the *Tcf19* and *Cchcr1* genes, where multiple factors such as Smarca4, Zbtb16, and Sall4 are colocalized, suggesting that the *Tcf19–Cchcr1* locus is robustly hypo-methylated and open to bind Sall4 and other TFs in spermatogonia. Representative ChIP-seq (SRX1284250), ATAC-seq (SRX5884282) and WGBS (SRX749893) alignment data are also exhibited in the top panel of Figure 2B. The IGV sessions can be saved as XML files at any moment, so that the results obtained by using the Peak Browser tool can be easily and precisely shared among collaborators (Supplementary Material S2). With the use of ChIP-Atlas, users can not only check individual experimental data, but also browse an integrative landscape of multiple epigenetic profiling results, potentially providing useful insight into the location of functional genomic regions (enhancers, promoters and insulators) and the corresponding regulators.

Enrichment analysis. Enrichment Analysis is a tool that allows a search for TFs, accessible chromatin regions, and hypo- or hyper-MRs enriched at a batch of genomic ROIs or gene loci. Upon the submission of two sets of genomic regions (ROIs and background regions), all SRXs are evaluated to count the overlaps between the peaks and submitted regions and to perform Fisher's exact test, before returning the enrichment analysis results in HTML and TSV formats. Although ChIP-Atlas can generate random background regions for comparison to ROIs, the users are strongly recommended to provide biologically appropriate background genomic intervals because most randomly selected regions are probabilistically devoid of any functional genomic annotation. The results are assigned unique URLs, which are permanently available to the public. As an example of usage, we selected inflammatory bowel disease (IBD)-associated SNPs identified by GWAS as the ROIs ($n = 393$) and other SNPs as the background ($n = 3930$), and we applied these selections to the Enrichment Analysis tool (Figure 2C; detailed in the Materials and Methods). The results in HTML format are shown in Figure 2D (upper), including SRX IDs (column 1), features (column 3), cell types (column 5), P-values (column 9) and fold enrichments (column 11), where hypo-MRs of neutrophils are significantly enriched ($P = 1 \times 10^{-17}$). To visually interpret the results, we downloaded the resultant TSV files (refer to Supplementary Table S1 for unique URLs for each analysis) and generated volcano plots (Figure 2D [bottom]; Supplementary Table S1). The IBD-associated SNPs were shown to be enriched in hypo-MRs and accessible chromatin regions of blood-related cells (red in Figure 2D), including the ATAC-seq peaks of macrophages ($P = 1 \times 10^{-20}$) and Th17 cells ($P = 1 \times 10^{-17}$), both involved in the inflammation of IBD. Furthermore, the enrichment analysis of TF ChIP-seq data revealed that the IBD-associated SNPs were preferentially bound by STAT1 in monocytes ($P = 1 \times 10^{-22}$) and SPI1 in macrophages ($P = 1 \times 10^{-21}$), essential factors for inflammation and macrophage differentiation, respectively, which is consistent with the nature of IBD as an autoimmune disease (17). All resultant URLs for generating Figure 2D are summarized in Supplementary Table S2. API of the Enrichment Analysis tool is also provided by ChIP-Atlas, the general
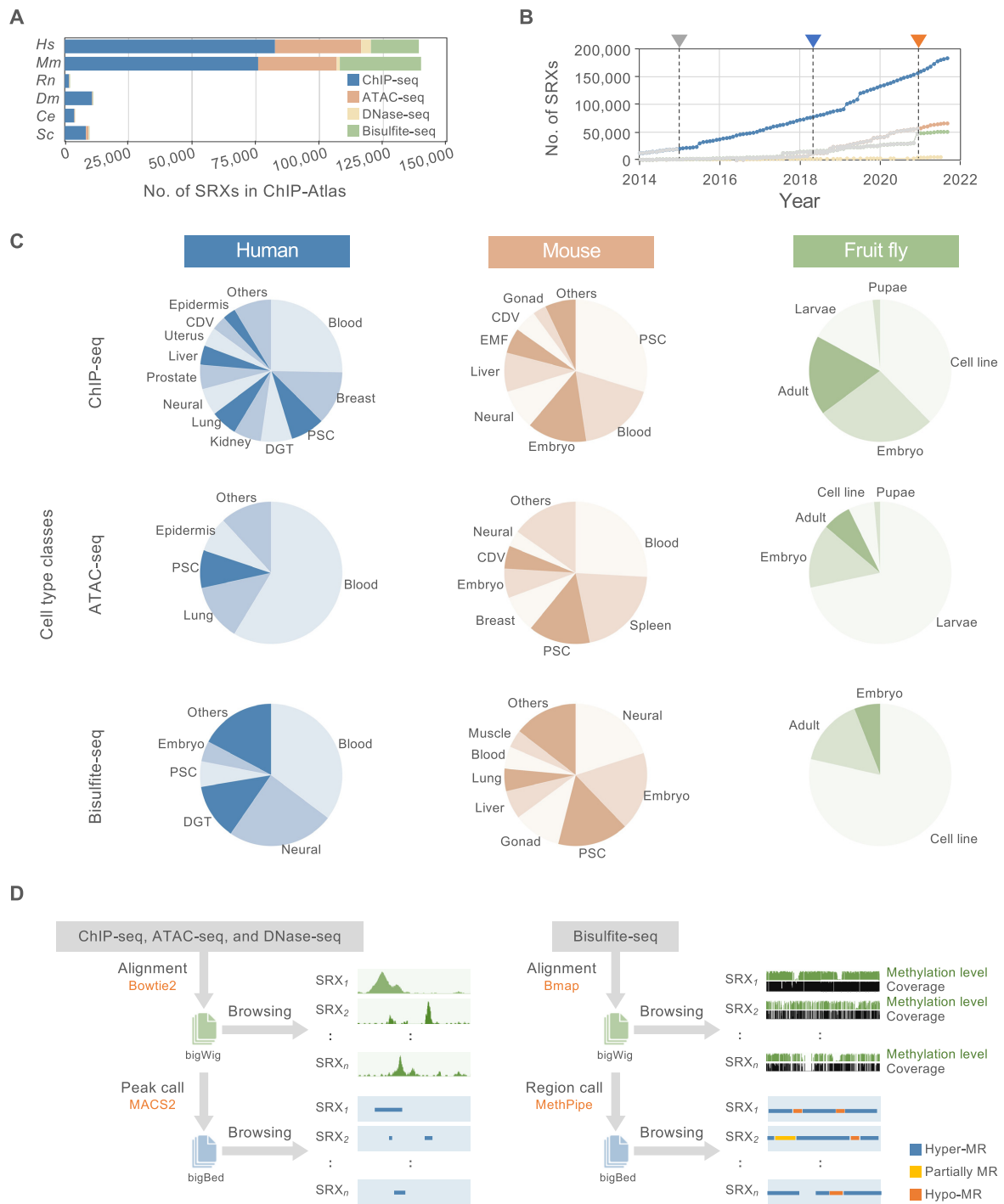
Figure 1. Overview of the ChIP-Atlas data set and computational processing. (**A**) Numbers of ChIP-seq, DNase-seq, ATAC-seq and WGBS experiments recorded in ChIP-Atlas (as of September 2021). Colors indicate different experiment types. Color legend is shown in the figure. *Hm, Homo sapiens; Mm, Mus musculus; Rn, Rattus norvegicus; Dm, Drosophila melanogaster; Ce, Caenorhabditis elegans; Sc, Saccharomyces cerevisiae*. (**B**) Cumulative number of SRX-based experiments recorded in ChIP-Atlas. Colors of the dots indicate different experiment types. Color legend is the same as that in (A). The gray, blue, and orange arrowheads indicate the timing of the release of ChIP-Atlas in 2015, the publication of the first ChIP-Atlas paper in 2018, and the inclusion of ATAC-seq and WGBS in 2021, respectively. ChIP-seq and DNase-seq data published before 2015, and ATAC-seq or WGBS data published before January 2021 are shown in gray. (**C**) Numbers of experiments according to cell type classes for human, mouse, and fruit fly data. PSC, pluripotent stem cell; CDV, cardiovascular; DGT, digestive tract; EMF, embryonic fibroblast. (**D**) Overview of data processing. Raw sequence data are downloaded from NCBI SRA, aligned to a reference genome, and subjected to peak calling, all of which can be monitored with the IGV genome browser. MR, methylated region.

Table 1. Data statistics of ChIP-Atlas

Experiment type	ChIP-seq		ATAC-seq	DNase-seq	Bisulfite-seq	All
Year	2021	2018	2021	2021	2021	2021
Number of experiments	182 891	74 076	66 104	5346	51 074	305 415
Number of ChIP antigens	2 477	1 979	NA	NA	NA	2 477
Number of cell types	2 486	1 615	772	406	553	2 918
Number of peaks	1 329 049 688	717 206 934	362 032 813	128 597 931	3 798 954 139	5 618 634 571

Table 2. Comparison of ChIP-Atlas with other similar web services

	ChIP-Atlas	Cistrom DB	ReMap	GTRD	MethBank
Data source	NCBI SRA	GEO, ENCODE, and Roadmap Epigenetics	GEO, ENCODE, and ENA	GEO, SRA, ENCODE, and modENCODE	NCBI SRA and Genome Sequence Archive
Experiments	ChIP-seq, **ATAC-seq**, DNase-seq, and **Bisulfite-seq**	ChIP-seq, ATAC-seq, and DNase-seq	ChIP-seq, ChIP-exo, and DAP-seq	ChIP-seq, ChIP-exo, ChIP-nexus, MNase-seq, DNase-seq, FAIRE-seq, ATAC-seq, and RNA-seq	Bisulfite-seq
Filtering of data for quality control	No	Yes	Yes	No	Yes
Number of experiments	76 217 → **305 415**	56 442	19 983	36 540	673
Organism	Hs, Mm, Rn, Dm, Ce, and Sc	Hs and Mm	Hs, Mm, Dm and At	Hs, Mm, At, Ce, Dr, Dm, Rn, Sc and Sp	Hs, Dr, Mm, Os, Gm, Me, Pv and Sl
Genome assembly	hg19, **hg38**, mm9, **mm10**, rn6, dm3, **dm6**, ce10, **ce11** and sacCer3	hg38 and mm10	hg38, mm10, dm6, and tair10 (can be lifted to hg19/mm9 with liftover)	hg38, mm10, tair10, ce11, danRer11, dm6, rn6, sacCer3 and spo2	hg38, danRer7, mm10, IRGSP-1.0, Gmax_275_v2.0, Mesculenta_305_v6, Pvulgaris_218_v1 and GCF_000188115.3_SL2.53
Alignment tool	Bowtie2/**BMap**	BWA	Bowtie2	Bowtie2	WSBA
Peak caller	MACS2/**MethPipe**	MACS2	MACS2	MACS2, MACS, GEM, PICS, SISSRs	NA
Display format for each experiment	Alignment and peaks	Alignment and peaks	Peaks	Peaks	Alignment
Browsing assembled peaks	Possible	None	Possible	Possible	NA
Genome browser	IGV and UCSC	WashU and UCSC	UCSC, ENSEMBL and IGV	Self-developed, UCSC and ENSEMBL	JBrowse
Integrative analysis tools	Search tool for target genes and colocalizing factors of given TF, and enrichment analysis tool for given genes and genomic coordinates	Search tool for target genes of single experiment, TFs binding to single given genomic locus or query gene	Enrichment analysis tool for given genomic coordinates relative to random background	Search tool for target genes of given TF	Predictor of DNA methylation, age of human blood and enrichment analysis tool for identification of differentially methylated promoters

Hm, *Homo sapiens*; *Mm*, *Mus musculus*; *Rn*, *Rattus norvegicus*; *Dm*, *Drosophila melanogaster*; *Ce*, *Caenorhabditis elegans*; *Sc*, *Saccharomyces cerevisiae*; *At*, *Arabidopsis thaliana*; *Dr*, *Danio rerio*; *Sp*, *Schizosaccharomyces pombe*; *Os*, *Oryza sativa*; *Gm*, *Glycine max*; *Me*, *Manihot esculenta*; *Pv*, *Phaseolus vulgaris*; *Sl*, *Solanum lycopersicum*.

instructions for which can be found on the documentation page (https://github.com/inutano/chip-atlas/wiki/Perform-Enrichment-Analysis-programmatically).

DISCUSSION

In this paper, we present a major update of ChIP-Atlas involving significant expansion in the number of experiments by including all public ATAC-seq and WGBS data. The updated web service enables users to explore not only protein binding, but also chromatin accessibility and DNA methylation levels within single (Peak Browser tool) or multiple (Enrichment Analysis tool) queries of genomic ROIs or gene loci. As an example of use, we performed enrichment analysis to reveal that IBD-associated SNPs are the most significantly hypo-methylated in the neutrophils, a granu-
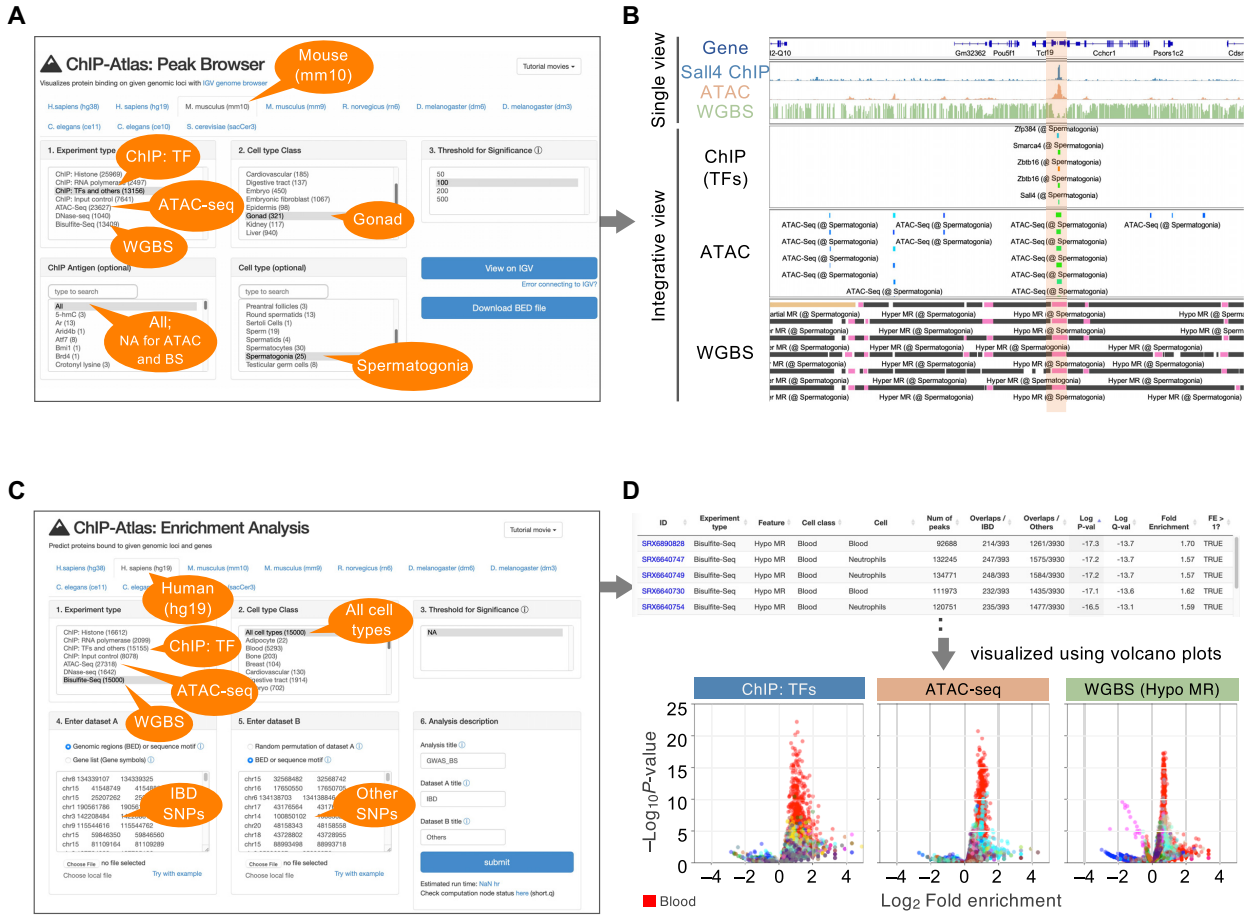
Figure 2. Examples of use of the integrative analysis tools in ChIP-Atlas. (**A**) Parameters when users perform a query to show chromatin accessibility, methylation status, and TF binding around mouse *Tcf19–Cchcr1* locus in spermatogonia using the Peak Browser tool. (**B**) Peak-call data for TF ChIP-seq, ATAC-seq and WGBS SRXs around the mouse *Tcf19–Cchcr1* locus are shown in the IGV genome browser. The highlighted region indicates an accessible chromatin and hypo-methylated region. Bars represent the peak regions, with the curated names of the ChIP antigens (only for ChIP-seq peaks) and cell types being shown below the bars. The color of the bars in the 'ChIP (TFs)' and 'ATAC' panel indicates the score calculated with the peak-caller MACS2 ($-\log_{10}[Q\text{-value}]$). Black, pink, and beige bars in the 'WGBS' panel indicate hyper-, hypo- and partially methylated regions, respectively. (**C**) Parameters when users perform a query to evaluate TF binding, chromatin accessibility, and methylation levels of multiple IBD-associated SNPs in all cell types. (**D**) The resultant HTML of enrichment analysis using WGBS data (upper). Volcano plots for visualizing the results (bottom). X-axis, $\log_2[\text{fold enrichment}]$; Y-axis, $-\log_{10}[P\text{-value}]$. Each dot represents an individual SRX. SRXs are categorized by cell type classes and indicated by different colors. Red, 'Blood' class. The whole color legend is summarized in Supplementary Table S2.

locyte subtype known to be involved in autoinflammatory IBD (17).

Before and after the public release of ChIP-Atlas, several similar web services have been released (Table 2). For ChIP-seq, ATAC-seq and DNase-seq data, Cistrome DB (http://cistrome.org/db/), ReMap (https://remap2022.univ-amu.fr/), and GTRD (https://gtrd.biouml.org/) are representatives providing thousands of preprocessed datasets (18–20). As for Bisulfite-seq data, MethBank (http://bigd.big.ac.cn/methbank) is widely used (21). ChIP-Atlas covers much more experimental data than all of these other services. ChIP-Atlas does not cover ChIP-exo (22) and MNase-seq (23) data, in contrast to ReMap and GTRD, and there are fewer available organisms than in GTRD and MethBank. Alignment data (in bigWig format) are available from ChIP-Atlas, Cistrome DB and MethBank. Integrative analysis

tools are provided in all services. Enrichment analysis is possible with ChIP-Atlas in both a GUI and an API, while ReMap provides such a tool in a CLI (R package named 'ReMapEnrich'). Since the publication of the first ChIP-Atlas paper, we have improved the service to make it compatible with the latest versions of reference genomes such as hg38 for human and mm10 for mouse, as is the case for other services. Meanwhile, alignment of raw reads against old references is still performed during the monthly update of ChIP-Atlas, and alignment and peak-call data in old versions are still provided for analyzing the data of users obtained years ago.

In addition to the examples of use mentioned in the RESULTS section, ChIP-Atlas has been cited by hundreds of peer-reviewed articles since its first release in 2015, including research for analyzing cis-regulatory elements of cer-

tain genes (24,25), and TF enrichment at genomic ROIs and query genes (26–30) (see http://chip-atlas.org/publications for the full list of publications citing ChIP-Atlas). Furthermore, because all alignment (bigWig) and peak-call (bigBed) data can be freely downloaded, ChIP-Atlas is now interconnecting with many other databases or web services such as UCSC Browser, DeepBlue (an epigenomic data server providing a central data access hub for large collections of epigenomic data), RegulatorTrail (a web service predicting target genes of TFs), jPOSTrepo (a data repository of sharing raw/processed mass spectrometry data), and the Signaling Pathways Project (a multi-omics knowledgemine based upon public transcriptomic and cistromic datasets) (31–35). Along with the inclusion of ATAC-seq and WGBS data, and the ongoing monthly updates with semiautomatic pipelines and systematic curation, the source data in ChIP-Atlas are continuously expanding. We are planning to include more experiment types such as CUT&Tag (36) and ChIL-seq (37) and more organisms including plants such as *Arabidopsis thaliana*. Integration of preprocessed 3D genome conformation data such as Hi-C datasets (38) into the Peak Browser and Enrichment Analysis tool is also on the agenda.

DATA AVAILABILITY

ChIP-Atlas (https://chip-atlas.org) is a publicly available web server with no sign up required. Documentation for data processing and downloadable data are available in the 'Documentation' section (https://github.com/inutano/chip-atlas/wiki).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
2. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
3. 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.
4. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
5. Cedar,H. and Bergman,Y. (2012) Programming of DNA methylation patterns. *Annu. Rev. Biochem.*, **81**, 97–117.
6. Owen-Hughes,T. and Gkikopoulos,T. (2012) Making sense of transcribing chromatin. *Curr. Opin. Cell Biol.*, **24**, 296–304.
7. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.
8. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.
9. 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.
10. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Nat. Acad. Sci. U.S.A.*, **89**, 1827–1831.
11. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
12. 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.
13. Miura,F., Enomoto,Y., Dairiki,R. and Ito,T. (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.*, **40**, e136.
14. Song,Q., Decato,B., Hong,E.E., Zhou,M., Fang,F., Qu,J., Garvin,T., Kessler,M., Zhou,J. and Smith,A.D. (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One*, **8**, e81148.
15. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC genome browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
16. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinf.*, **14**, 178–192.
17. Chang,J.T. (2020) Pathophysiology of inflammatory bowel diseases. *N. Engl. J. Med.*, **383**, 2652–2664.
18. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.H., Brown,M., Zhang,X., Meyer,C.A. *et al.* (2019) Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.
19. Hammal,F., de Langen,P., Bergon,A., Lopez,F. and Ballester,B. (2021) ReMap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.*, **50**, D316–D325.
20. Kolmykov,S., Yevshin,I., Kulyashov,M., Sharipov,R., Kondrakhin,Y., Makeev,V.J., Kulakovskiy,I.V., Kel,A. and Kolpakov,F. (2021) GTRD: an integrated view of transcription regulation. *Nucleic Acids Res.*, **49**, D104–D111.
21. Li,R., Liang,F., Li,M., Zou,D., Sun,S., Zhao,Y., Zhao,W., Bao,Y., Xiao,J. and Zhang,Z. (2018) MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic Acids Res.*, **46**, D288–D295.

22. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single nucleotide resolution. *Cell*, **147**, 1408–1419.

23. Chereji,R.V., Bryson,T.D. and Henikoff,S. (2019) Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol.*, **20**, 198.

24. Wang,R., Yamada,T., Kita,K., Taniguchi,H., Arai,S., Fukuda,K., Terashima,M., Ishimura,A., Nishiyama,A., Tanimoto,A. *et al.* (2020) Transient IGF-1R inhibition combined with osimertinib eradicates AXL-low expressing EGFR mutated lung cancer. *Nat. Commun.*, **11**, 4607.

25. Groff,A.F., Barutcu,A.R., Lewandowski,J.P. and Rinn,J.L. (2018) Enhancers in the peril lincRNA locus regulate distant but not local genes. *Genome Biol.*, **19**, 219.

26. Hänsel-Hertsch,R., Simeone,A., Shea,A., Hui,W.W.I., Zyner,K.G., Marsico,G., Rueda,O.M., Bruna,A., Martin,A., Zhang,X. *et al.* (2020) Landscape of G-quadruplex DNA structural regions in breast cancer. *Nat. Genet.*, **52**, 878–883.

27. Roels,J., Kuchmiy,A., De Decker,M., Strubbe,S., Lavaert,M., Liang,K.L., Leclercq,G., Vandekerckhove,B., Van Nieuwerburgh,F., Van Vlierberghe,P. *et al.* (2020) Distinct and temporary-restricted epigenetic mechanisms regulate human αβ and γδ t cell development. *Nat. Immunol.*, **21**, 1280–1292.

28. Zou,Z., Iwata,M., Yamanishi,Y. and Oki,S. (2022) Epigenetic landscape of drug responses revealed through large-scale chip-seq data analyses. *BMC Bioinf.*, **23**, 51.

29. Tanaka,H., Igata,T., Etoh,K., Koga,T., Takebayashi,S.i. and Nakao,M. (2020) The NSD2/WHSC1/MMSET methyltransferase prevents cellular senescence-associated epigenomic remodeling. *Aging Cell*, **19**, e13173.

30. Ferris,E., Abegglen,L.M., Schiffman,J.D. and Gregg,C. (2018) Accelerated evolution in distinctive species reveals candidate elements for clinically relevant traits, including mutation and cancer resistance. *Cell Rep.*, **22**, 2742–2755.

31. Lizio,M., Abugessaisa,I., Noguchi,S., Kondo,A., Hasegawa,A., Hon,C.C., De Hoon,M., Severin,J., Oki,S., Hayashizaki,Y. *et al.* (2019) Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.*, **47**, D752–D758.

32. Albrecht,F., List,M., Bock,C. and Lengauer,T. (2016) DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Res.*, **44**, W581–W586.

33. Kehl,T., Schneider,L., Schmidt,F., Stöckel,D., Gerstner,N., Backes,C., Meese,E., Keller,A., Schulz,M.H. and Lenhof,H.P. (2017) RegulatorTrail: a web service for the identification of key transcriptional regulators. *Nucleic Acids Res.*, **45**, W146–W153.

34. Okuda,S., Watanabe,Y., Moriya,Y., Kawano,S., Yamamoto,T., Matsumoto,M., Takami,T., Kobayashi,D., Araki,N., Yoshizawa,A.C. *et al.* (2017) jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.*, **45**, D1107–D1111.

35. Ochsner,S.A., Abraham,D., Martin,K., Ding,W., McOwiti,A., Kankanamge,W., Wang,Z., Andreano,K., Hamilton,R.A., Chen,Y. *et al.* (2019) The signaling pathways project, an integrated 'omics knowledgebase for mammalian cellular signaling pathways. *Scientific Data*, **6**, 252.

36. Kaya-Okur,H.S., Wu,S.J., Codomo,C.A., Pledger,E.S., Bryson,T.D., Henikoff,J.G., Ahmad,K. and Henikoff,S. (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.*, **10**, 1930.

37. Harada,A., Maehara,K., Handa,T., Arimura,Y., Nogami,J., Hayashi-Takanaka,Y., Shirahige,K., Kurumizaka,H., Kimura,H. and Ohkawa,Y. (2018) A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat. Cell Biol.*, **21**, 287–296.

38. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

主論文 2（Zou, *Nucleic Acid Res.*）の補足資料は下記 URL より入手可能である。

Nucleic Acid Research ホームページ：

https://doi.org/10.1093/nar/gkac199

京都大学ファイル共有システム：

https://fsv.iimc.kyoto-u.ac.jp/public/Ie0sgAxSKYlAYIgBRwuDxif4kWsKqw5f4j0ixBKxy-40