Published in final edited form as:
Comput Toxicol. 2019 November ; 12: . doi:10.1016/j.comtox.2019.100094.

Integration of curated and high-throughput screening data to elucidate environmental influences on disease pathways

Marissa B. Kosnik[a,b,c,*], **Antonio Planchart**[a,c,d], **Skylar W. Marvel**[b,c], **David M. Reif**[a,b,c,d], **Carolyn J. Mattingly**[a,c,d]

[a]Toxicology Program, North Carolina State University, North Carolina State University, Raleigh, NC 27695-7617, United States

[b]Bioinformatics Research Center, North Carolina State University, North Carolina State University, Raleigh, NC 27695-7617, United States

[c]Department of Biological Sciences, North Carolina State University, North Carolina State University, Raleigh, NC 27695-7617, United States

[d]Center for Human Health and the Environment, North Carolina State University, Raleigh, NC 27695-7617, United States

Abstract

Addressing the complex relationship between public health and environmental exposure requires multiple types and sources of data. An important source of chemical data derives from high-throughput screening (HTS) efforts, such as the Tox21/ToxCast program, which aim to identify chemical hazard using primarily *in vitro* assays to probe toxicity. While most of these assays target specific genes, assessing the disease-relevance of these assays remains challenging. Integration with additional data sets may help to resolve these questions by providing broader context for individual assay results. The Comparative Toxicogenomics Database (CTD), a publicly available database that builds networks of chemical, gene, and disease information from manually curated literature sources, offers a promising solution for contextual integration with HTS data. Here, we tested the value of integrating data across Tox21/ToxCast and CTD by linking elements common to both databases (i.e., assays, genes, and chemicals). Using polymarcine and Parkinson's disease as a case study, we found that their union significantly increased chemical-gene associations and disease-pathway coverage. Integration also enabled new disease associations to be made with HTS assays, expanding coverage of chemical-gene data associated with diseases. We demonstrate how integration enables development of predictive adverse outcome pathways using 4-nonylphenol, branched as an example. Thus, we demonstrate enhancements to each data source through database integration, including scenarios where HTS data can efficiently probe chemical space

[*]Corresponding author at: Box 7566, 1 Lampe Drive, North Carolina State University, Raleigh, NC 27695. mbkosnik@ncsu.edu (M.B. Kosnik).

Declaration of Competing Interest
The authors declare no conflicts of interest.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.comtox.2019.100094.

that may be understudied in the literature, as well as how CTD can add biological context to those results.

Keywords

Data integration; ToxCast; CTD; Disease; Disease pathways

1. Introduction

There are more than 80,000 chemicals registered for use in the United States with an estimated two thousand more introduced each year [1]. The majority of these have not been adequately tested for their human health effects despite the etiology of many chronic diseases involving interactions between environmental factors, including chemicals, and genes and pathways modulating physiological processes [2–4]. To address this challenge, high-throughput screening (HTS) efforts like the Toxicology in the 21st Century (Tox21) federal research collaboration [5] have been developed to automate *in vitro* biological assays and maximize efficiency of evaluating the activity of a large number of chemicals on a range of cellular processes. Members of the Tox21 collaboration seek to enhance the predictive capacity of toxicology studies and thereby improve efforts to protect human health and the environment. The goals of Tox21 include developing and improving models that predict biological responses to chemicals, identifying mechanisms of action that warrant further investigation, and prioritizing chemicals for further toxicological evaluation. Utilizing *in vitro* assays in a systematic, large-scale operation increases clarification of specific molecular endpoints compared to traditional animal toxicology studies. A major effort targeted at chemical prioritization is the Toxicity Forecaster (ToxCast) program [6–8], which is an ongoing, multiphase component of the U.S. Environmental Protection Agency’s (EPA) contributions to Tox21. ToxCast enables prioritization and profiling of chemicals of regulatory interest by their AC50/LEC (half-maximal activity concentration/lowest effective concentration) values or by mapping assay results onto canonical biochemical or physiological pathways by way of implicated genes. However, there is an urgent need to understand these data in a broader biological context, including their alignment with human disease or exposure implications.

One way to accomplish this is to develop evidence-based associations between HTS results and broader biological resources. With the exponential growth in environmental health data, new databases and tools have been developed to enable analysis of disconnected datasets [3]. The Comparative Toxicogenomics Database (CTD) [9–11] is one such publicly available database developed with the goal of advancing understanding about how environmental exposures affect human health. CTD accomplishes this goal by manually curating chemical-gene, chemical-phenotype, chemical-disease, and gene-disease relationships as well as exposure data from the biomedical literature. These data are integrated with functional and pathway data to inform hypotheses about the etiologies underlying environmentally influenced diseases. In addition, CTD also includes manually curated chemical-phenotype relationships for identifying pre-disease biomarkers associated with experimental and real-world environmental exposures [10–12].

Like many databases, the Tox21/ToxCast collaborative effort and CTD share objectives of better characterizing the role of chemicals on human health outcomes. Where Tox21/ToxCast assays generate evidence of a chemical affecting gene activity within an *in vitro* context, CTD curates evidence of chemical associations with genes, proteins or disease from diverse sources (e.g. model organisms, human populations, *in vitro*). Establishing methods for integrating Tox21/ToxCast results with curated data in CTD represents an initial step toward addressing well-known, long-term challenges facing the Tox21/ToxCast effort [4]. These challenges include how to extrapolate HTS data to human health by correlating perturbation of genes, proteins or cellular-based phenotypes to human disease. Integration of HTS results could also address information gaps in CTD owing to the comparatively narrow range of chemicals reported in the literature [12,13]. By integrating HTS and environmental health data, the respective enhancements to each resource can be analyzed in a human health context.

Here, we describe integration of HTS data with a broader environmental health resource using Tox21/ToxCast data and CTD. Through this data integration, we demonstrate the expansion of chemical-gene coverage. Using chemical-gene interactions associated with diseases and pathways in CTD as a case study, we describe the respective enhancements to each resource and demonstrate how this data integration can be used to identify new chemical-pathway and chemical-disease associations. We assess changes in coverage of chemical-gene information for diseases and pathways in CTD to quantify the value of this database integration and demonstrate how these integrative techniques can advance discovery through development of predictive chemical-pathway-disease frameworks.

2. Materials and methods

2.1. CTD data

CTD data from the October 05, 2018 release were downloaded as .CSV files from the CTD website (http://ctdbase.org/downloads). The disease vocabulary (MEDIC) contained 36 MEDIC-Slim categories encompassing 5,361 specific diseases [14]. The chemical-gene interactions file included 12,984 chemicals and 46,755 genes and proteins with data recorded from 87,428 curated references [15]. Also included were pathway-gene/protein associations, enriched pathway-chemical associations, gene-disease associations, and chemical-disease associations. Only chemicals and genes from the CTD chemical-gene interactions file were used as the basis for the CTD chemical and gene datasets.

2.2. HTS assay data from Tox21/ToxCast

The most-recent, full release of the invitrodb_v2 collection of ToxCast and Tox21 HTS assay information (October 2015) was obtained from the ToxCast website (These combined data are hereafter referred to as “HTS”, in reference to the data type) as a collection of .CSV files (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data) [16]. These data consisted of 9,019 chemicals, 359 assays, 1,112 assay endpoints, and 115,857 chemical-end-point pairs. Endpoints consist of molecular targets (i.e., gene(s) or protein(s) specified by NCBI GeneIDs) or biological processes (e.g., apoptosis) [17]. We focused on assays with molecular targets as specified in the “technological target” field of the assay summary file (Assay_Summary_151020.csv). Among these assays, 28 have multiple technological targets as designated by several GeneIDs reflecting cases in which a specific gene cannot be resolved (e.g., multiple retinoic acid receptors that can activate a response element in a reporter assay). The assay endpoint ID (aeid), which appears in the assay summary (Assay_Summary_151020.csv) and in the screening results (AllResults_flags_151020.csv) files, was used to map chemicals to molecular targets. The direction of gene activity (increase versus decrease) was determined using the assay component endpoint name and description for each chemical-gene association. For signaling assays, chemical-gene associations were characterized based on the direction of the assay the chemical was active in (up versus down). Chemicals were described as decreasing gene activity if the chemical was active in an inhibition assay. If a direction of activity could not be determined from the assay (e.g. binding assays) or a chemical was active in both the up and down signaling direction, then the interaction was classified as the chemical "affects" the gene rather than increases or decreases activity.

2.3. Integrating databases

Fig. 1 outlines the HTS/CTD data integration process. In brief, all chemicals and genes were integrated from the two databases (Fig. 1b and c), along with their respective chemical-gene associations (Fig. 1a and d). HTS chemicals are identified using names and Chemical Abstracts Service Registry Number (CASRNs). Chemicals in CTD are identified using names, synonyms and CASRNs derived from the MeSH Chemicals and Drugs category. CASRNs were used to map chemicals between HTS and CTD data sets where possible, or by comparing chemical names to names and synonyms in CTD (Fig. 1b). Among HTS chemicals, 242 did not have CASRNs. Using this approach, 3,237 chemicals (36% of all HTS chemicals and 25% of all CTD chemicals) were shared between these datasets (Fig. 1b, Grey). Genes in HTS and CTD are both designated using gene identifiers (GeneIDs) from the National Center for Biotechnology Information (NCBI) Entrez Gene database and were used to map genes across the datasets (Fig. 1c) [18]. Among the 366 genes implicated in HTS assays, 285 were identified in CTD (Fig. 1c, Grey). This fraction represents 78% of HTS genes and less than 1% of genes in the broader CTD database. The 285 genes and 3,237 chemicals common to both CTD and HTS form the intersecting dataset (Fig. 1a). HTS contained 5,782 chemicals and 81 genes without matches in CTD (Fig. 1b and c, Purple) and CTD contained 9,747 chemicals and 46,470 genes without matches in HTS (Fig. 1b and c, Blue). These remaining chemicals and genes that were unmatched between the two datasets formed the remainder of our union dataset, representing the entirety of the chemicals and genes in each resource integrated together (Fig. 1d). These linked chemical names and gene names were used to form connections between CTD chemical-disease, chemical-pathway, gene-disease, and gene-pathway data and HTS chemical-gene associations.

2.4. Context-defined overlap between datasets

Only chemicals designated as "actives" in HTS assays (i.e., ToxCast hit call = 1, indicating association between a given chemical and a given assay), were considered in the analysis of overlap between HTS and CTD. In addition, only actives with AC50 concentrations less than 1 μM for at least one assay were considered. This stringency is a tunable parameter that was set to maximize specificity of assay targets, versus allowing more general chemical cytotoxicity to cloud specific gene signals [19]. This filtered the number of chemicals in the intersecting dataset to 1,344 (42% of the 3,237 overlapping chemicals). An additional 1,808 chemicals from HTS fit these criteria but did not have chemical matches in CTD, thus forming part of our union database.

2.5. Disease and pathway analysis

Disease analyses were conducted using CTD MEDIC-Slim disease categories, disease-chemical, disease-gene, and disease-pathway association data. Pathway analyses were conducted using CTD pathway-chemical and pathway-gene association data. New, enriched diseases and pathways for chemicals were identified using genes associated with these chemicals in the union dataset and calculating the significance of enrichment using the hypergeometric test with a Bonferroni adjustment for multiple testing [20]. Only direct disease-gene associations in CTD were used to determine disease enrichment. The number of tests used for the Bonferroni adjustment was set as the number of diseases or pathways with at least one gene annotation for that gene or pathway in the union dataset. The adjusted significance threshold was p less than 0.01. This is the same methodology used by CTD’s online tools to determine disease and pathway enrichment (http://ctdbase.org/tools/analyzer.go). Network diagrams of pathway-gene connections were generated using the R/igraph package [21]. Pathway diagrams of gene interactions were developed using the genes from enriched chemicalpathway associations with the R/pathview package [22]. Heatmaps of disease chemical-gene associations were computed using the R/gplots package [23]. Filled curve plots for MEDIC-Slim chemical-disease and gene-disease coverage were generated using the R/ggplot2 package [24]. All analyses were done using R version 3.3.2 [25].

3. Results

3.1. Chemical and gene interactions forming the CTD/HTS integrated database

The union of the entire CTD and HTS datasets contains 14,756 chemicals and 46,836 genes and is referred to as the union dataset, with the majority of data coming from CTD (Fig. 1d). The full dataset of chemical-gene associations is available as Supplemental Table 1. The overlapping subset of the union dataset consists of 1,344 chemicals and 285 genes and is referred to as the intersecting dataset (Fig. 1a).

To assess the pathway and disease coverage of the union and intersecting datasets, we analyzed the associations between genes and chemicals in the entire dataset with pathways and diseases found in CTD. An overall network diagram showing gene-pathway annotations for all genes in the union dataset, overlaid with genes contained in the intersecting dataset was used to assess the relative density and connectivity of genes (Supplemental Fig. 1). Genes from the union dataset were connected collectively with 2,352 different pathways; genes from the intersecting dataset were connected with 1,181 pathways. Genes contained in the intersecting dataset were spread throughout the pathway space rather than concentrated in dense hubs, suggesting similar coverage in the two datasets (Supplemental Fig. 1). Next, we compared the coverage of disease categories between the union and intersecting dataset to determine if the disease associations were similar between the chemical space and the gene space. MEDIC-Slim, a set of terms from CTD that organizes diseases into 36 general categories, was used to provide a simplified perspective of the entire disease landscape [26]. Overall, there was strong agreement among categories of diseases for both the chemicals (rank order p = 0.60) and genes (rank order p = 0.72) in the union (Supplemental Fig. 2a) and intersecting datasets (Supplemental Fig. 2b). This indicates similar coverage of disease processes between the two datasets. For both the union and intersecting datasets, genes and chemicals were most commonly associated with the disease categories nervous system disease and cancer; MEDIC-Slim disease categories with few diseases were associated with low numbers of genes and chemicals in the union dataset. For example, the nutrition disorder disease category contains about 0.5% of the genes and chemicals in the union dataset and contains only 38 diseases compared to the nervous system disease category, which contains 1,556 diseases and close to 10% of the genes and chemicals in the union dataset.

3.2. Integration effects on chemical-gene associations from a disease perspective

To determine the effect of CTD/HTS data integration on understanding potential mechanisms and influences on individual diseases, we assessed chemical-gene associations in the intersecting dataset (Fig. 1a). Based on our analyses of chemical and gene associations with diseases among MEDIC-Slim categories (Supplemental Fig. 2), we determined that focusing on the intersecting dataset would be representative of disease associations from the entire union dataset. This enables better visualization of the small number of HTS genes without the larger number of CTD gene contributions overwhelming our results. As test cases, we examined Parkinson Disease (PD), a disease with many chemicals and genes associated in CTD, and malnutrition, a disorder with fewer chemical and gene associations in CTD.

Parkinson's Disease (PD): PD is a member of the most common disease category (nervous system diseases) associated with chemicals and diseases in the union and intersecting datasets (Supplemental Fig. 2). PD is a neurodegenerative disease affecting human populations of all ethnicities. Although familial cases are known, a significant proportion of PD patients have no known genetic risk factors; thus, environmental risk factors are suspected in the etiology of idiopathic PD [27–29]. Furthermore, with agricultural pesticides of emerging concern as suspected agents in the etiology of PD [28–30], we hypothesized that the pesticide-rich HTS data may augment chemical-gene representations in CTD that when integrated with CTD data would provide disease context for these chemicals [12]. Over 4,800 chemicals and 20,000 genes associated with PD were identified in chemical-disease and gene-disease association data for PD in the union dataset. Of these, 744 chemicals and 283 genes were in the intersecting dataset. We generated a heatmap to illustrate the differential coverage of genes within the source datasets (CTD and HTS), and to underscore the increased coverage that can be attained via integration (Fig. 2a). The total number of unique chemical-gene associations in the intersecting dataset was 35,617. Approximately ten percent (3,569) of the chemical-gene associations were the same in both databases prior to integration, whereas CTD accounted for 14,735 (41%) distinct chemical-gene associations and HTS accounted for 17,313 (49%) distinct chemical gene-associations. Integration increased the mean number of genes per chemical nearly 2-fold versus either dataset alone based on existing chemical-disease and gene-disease associations in CTD.

Malnutrition: Malnutrition (Fig. 2b) is from the nutrition disorder MEDIC-Slim category, one of the categories with the lowest percentage of genes and chemicals in the union and intersecting dataset (Supplemental Fig. 2). Unsurprisingly, there are fewer chemicals and genes associated with malnutrition than PD in the union dataset: 226 chemicals and 4,543 genes. Numbers were reduced to 79 chemicals and 134 genes in the intersecting dataset. As with PD, we generated a heatmap to analyze the differential contribution of chemical-gene associations from the original data sources. CTD incorporates significantly more information than HTS with 67% of the 4,771 chemical-gene associations coming from CTD compared to 15.8% coming exclusively from the HTS data.

3.3. Integration effects on chemical-disease and chemical-pathway associations from a specific chemical perspective

To characterize how differing sources of chemical-gene associations (CTD vs HTS) influence disease and pathway enrichment in the union and intersecting datasets and identify the scenarios where this integration is most useful, we analyzed data for a well-studied chemical (benzo[a]pyrene, or BaP) and a relatively understudied chemical (polymarcine, or metiram). BaP is a well-studied polycyclic aromatic hydrocarbon (PAH) [13] with a role in the etiology of many diseases [31,32]. PD has an inferred disease-association with BaP in CTD, and the majority of chemical-gene associations for BaP in the union dataset came from CTD. In contrast, polymarcine belongs to the ethylene-bisdithiocarbamate (EBDC) class of fungicides, and is understudied compared to other fungicides of the same class, such as maneb and mancozeb [33]. This class of fungicides is widely used and has been shown to lead to oxidative stress and neurotoxicity. Both maneb and mancozeb are suspected environmental agents in the etiology of idiopathic PD [30,33,34], and are associated with PD in CTD. In contrast, almost all chemical-gene associations in the union dataset for polymarcine came from the HTS data, and interestingly, polymarcine is not associated with PD in CTD. We hypothesized that by adding the HTS data to CTD, new diseases associated with polymarcine could be identified along with pathways involved in the etiology of these diseases. Since all the chemical-gene information for BaP is contained within CTD, we did not expect a similar enhancement in new disease and pathway associations to be seen following data integration with HTS.

To analyze disease enrichment for these two chemicals, existing chemical-disease associations in CTD and new, enriched disease associations based on the union chemical-gene set were identified (Table 1). In CTD, BaP has 11,894 chemical-gene associations and 4,520 chemical-disease associations, including PD. After integration with HTS data, there was no increase in the number of chemical-gene associations because the 13 genes identified in HTS were already associated with BaP in CTD. We used the 11,894 genes associated with BaP in the union dataset to identify enriched diseases using direct gene-disease associations from CTD with the hypergeometric test and identified 3,707 diseases. However, these diseases were already associated with BaP in CTD (Table 1).

In CTD, polymarcine is associated with 11 genes and has 135 chemical-disease associations. After CTD/HTS data integration, the number of genes associated with polymarcine increased to 147 with HTS adding 136 new genes. By using these 147 genes from the union dataset to identify enriched diseases as we did for BaP, 411 diseases were identified (Table 1). Of these, 307 were not previously associated with polymarcine in CTD, and one of these newly identified diseases was PD (corrected p-value less than 0.001).

To analyze pathway enrichment for these chemicals, existing gene-pathway annotations in CTD and new, enriched pathway associations based on the union gene set were identified. To assess the connectivity of the gene-pathway data, we developed network diagrams showing gene-pathway annotations for these chemicals before and after the addition of HTS chemical-gene associations to CTD. In CTD, BaP is associated with 1,357 pathways. Between these 1,357 pathways and the 11,894 genes associated with BaP in CTD, we generated a network diagram with 85,433 gene-pathway connections (Fig. 3a). After using the 11,894 genes associated with BaP in the union dataset to search for new, enriched pathways using gene-pathway associations from CTD with the hypergeometric test, the number of pathways associated with BaP increased to 1,545 with an additional 535 gene-pathway connections formed (Fig. 3b). In contrast, polymarcine is associated with only seven pathways in CTD. Between these seven pathways and the 10 genes associated with pathways in CTD, we generated a network with 16 gene-pathway connections (Fig. 3c). After incorporating the 147 genes associated with polymarcine in the union dataset and identifying enriched pathways the number of pathways increased to 453 with 4,401 total gene-pathway connections formed (Fig. 3d).

3.4. Data integration informs disease-pathway activity for chemical classes

Between the 135 diseases already associated in CTD and the 307 new, enriched diseases we identified, there were 442 diseases associated with polymarcine after data integration, where PD was among the new, enriched diseases. To demonstrate the utility of our data integration efforts in enhancing mechanistic understanding of a chemical’s role in disease progression, we studied the gene-pathway associations of polymarcine for pathways associated with PD. In CTD, PD is associated with 517 pathways, 242 of which are also associated with polymarcine. We narrowed down the overall network of gene-pathway connections for polymarcine to just those associated with both polymarcine and PD (Fig. 4a, left). To better characterize the gene-gene interactions involved in these polymarcine and PD pathways, we selected the tumor necrosis factor (TNF) signaling pathway for further analysis (shown in pink in network diagram with the associated genes shown in green, Fig. 4a left). TNF is an immune system regulator with an important role in inflammation and the potential to subsequently induce oxidative stress [35–38]. Consequently, the signaling pathway is believed to be involved in the progression of PD and other neurodegenerative diseases while also speculated to be a therapeutic target [35–37]. By determining the direction of chemical-gene activity from the ToxCast assay analysis, we assigned a direction of activity to each chemical-gene interaction for polymarcine and the genes associated with both the TNF signaling pathway and PD. Pathview [22] was used to develop a mechanistic diagram of these genes to characterize the role that polymarcine may play in the progression of PD via the TNF signaling pathway (Fig. 4a, right).

To demonstrate the utility of this process for groups of chemicals, we analyzed whether other EBDC fungicides covered similar pathways associated with PD to elucidate a potential toxicological response common to this disease progression. As with polymarcine, we used the union genes associated with maneb and mancozeb to identify enriched pathways. Of the 242 pathways associated with PD, 206 were common to all three EBDC fungicides. Of the 133 genes associated with polymarcine and PD pathways, 75 were common to all three EBDC fungicides. The gene-pathway network of PD pathways for polymarcine was adjusted to just those genes and pathways associated with all three EBDC fungicides (Fig. 4b, left). One of the pathways common to all three EBDC fungicides is the TNF signaling pathway. To better characterize the toxicological response of the EBDC fungicides that may be involved in PD, we generated a diagram of the TNF signaling pathway showing the direction of activity of EBDC fungicide-responsive genes (Fig. 4b, right). We found that there was general agreement in the pattern of gene activity among the three EBDC fungicides with multiple chemicals exhibiting activity in the same direction or exhibiting a general effect in some chemicals and directionality in others.

3.5. Data integration generates chemical-disease and chemical-pathway connections for chemicals without gene associations in CTD

In the union dataset, there were 1,808 chemicals that were either missing from or had no gene associations in CTD (Fig. 1b, Purple). The source of these chemicals' gene associations came solely from the HTS dataset with a median of 10 genes per chemical (range 1 to 117 genes). To assess how HTS results can be leveraged to mitigate information gaps in the literature and how CTD can provide biological context to HTS data, we used the hypergeometric test to identify enriched diseases for these 1,808 chemicals using the HTS chemical-gene associations and CTD's direct gene-disease associations. Based on the new chemical-disease enrichment we found for polymarcine after data integration, we hypothesized that there would be similar enrichment for these chemicals. In total, we identified 300,009 disease associations across the 1,808 chemicals, with a median of 132 diseases per chemical. These chemical-disease associations are presented in Supplemental Table 2. The total number of diseases enriched per chemical is shown in Fig. 5a. The most common diseases were liver neoplasms, female infertility, and adenocarcinomas, each enriched in over 1,700 chemicals. Bronopol and milbemectin had the most associations, each with over 430 enriched diseases. To determine which disease categories had the most enrichment from HTS chemicals, we calculated the percentage of enriched disease associations for these chemicals among MEDIC-Slim disease categories (Supplemental Fig. 3). Interestingly, the distribution of diseases differs from what we found for the union and intersecting datasets (Supplemental Fig. 2) with cancer still highly enriched, but nervous system disorders and other disease categories have a more even distribution.

To provide further context to these diseases and demonstrate the potential to study molecular mechanisms associated with these chemicals, we identified enriched pathway associations for the 1,808 chemicals using CTD gene-pathway associations and the same HTS chemical-gene associations. We identified 207,724 pathway associations across these chemicals with a median of 113 enriched pathways per chemical. These chemical-pathway associations are presented in Supplemental Table 3. The total number of pathways enriched per chemical is shown in Fig. 5b. The most common pathways enriched across chemicals were nuclear receptor transcription pathway and pathways in cancer, each enriched for over 1,700 chemicals. The chemicals associated with the most pathways were isopropyl triethanolamine titanate and PHA-00568487, each associated with over 360 pathways. Generally, the same chemicals had similar numbers of new disease and pathway associations (correlation coefficient = 0.49) and the number of new associations was correlated with the number of genes per chemical (correlation coefficient = 0.86 for diseases, 0.43 for pathways, Fig. 5a–d).

3.6. Data integration can inform adverse outcome pathways by connecting chemical, gene, disease, and pathway associations

An adverse outcome pathway (AOP) is a framework that organizes existing information to provide biologically supported explanations of how a toxicant can lead to adverse outcomes, such as diseases [39]. CTD has been demonstrated to aid in predictive AOP development by linking chemical-gene and phenotype data to disease and exposure data [10], however this process would not be possible for the 1,808 HTS chemicals without chemical-gene information in CTD. We hypothesized that by relating the chemical-gene activity from the HTS dataset, the pathway-disease data from CTD, and the chemical-pathway and chemical-disease associations identified through enrichment, we could begin to develop predictive AOPs for new chemical-disease linkages. For this study, we chose to continue focusing on PD and selected a toxicant that was also enriched for the TNF signaling pathway. Of the 1,808 chemicals with disease associations, 88 were enriched for PD (shown in blue and pink, Fig. 5a and c) of which 84 were also enriched for the TNF signaling pathway (shown in green and pink, Fig. 5b and d). One of these chemicals enriched for both PD and the TNF signaling pathway was 4-nonylphenol, branched (hereafter referred to as 4-nonylphenol, shown in pink, Fig. 5a–d). 4-nonylphenol is a recognized endocrine disruptor that is widespread throughout the environment and can produce reactive oxygen species leading to neurotoxicity [40–42], but its potential role in PD development is understudied. In developing an AOP for 4-nonylphenol, we characterized the HTS chemical-gene activity associated with 4-nonylphenol as the molecular initiating event (MIE) and the pathways associated with both 4-nonylphenol and PD as the downstream key events (Fig. 6). Of the 286 pathways enriched for 4-nonylphenol, 173 were associated with PD in CTD. For AOP development, we limited these pathways to the 40 that characterized phenotypes of PD (e.g. dopaminergic synapse) rather than comorbidities (e.g. asthma, breast cancer) and had a corresponding pathway ID in Pathview [22]. Each of these 40 pathways can be used to generate a diagram of HTS gene interactions to better characterize the connections between these key events (Fig. 6a–d). For example, as we did with polymarcine and the EBDC fungicides, we used Pathview to visualize the TNF signaling pathway for 4-nonylphenol (Fig. 6, panel c). In addition to the MIEs connected to and depicted in the TNF signaling pathway, the pathway also feeds into the PI3K-Akt pathway (Fig. 6, panel b) and the NF-κB signaling pathway (Fig. 6, panel d).

4. Discussion

Data integration can leverage information in public databases to generate new knowledge not contained in any single resource. We present a method for data integration using Tox21/ToxCast data and the Comparative Toxicogenomics Database as a case study to align high-throughput screening results with disease-relevant pathways. We found that integration augmented chemical-gene, gene-disease, and pathway-gene coverage. By assessing changes in coverage of, as well as connections between, chemical-gene information for diseases and pathways, we demonstrate that integration of HTS and CTD data can augment the biological context and depth of each resource. These data are currently available as a supplemental table (Supplemental Table 1) to provide valuable evidence for development of new hypotheses for under-studied chemicals. The analysis can be updated with future source database downloads by following the procedures described in our methods.

The connection of HTS data through CTD to a high-level, complex disease (PD) illustrates the value of integration through an increased number of chemical-gene associations. When compared to malnutrition, a disorder with fewer chemical and gene annotations in CTD, we did not see as much enhancement following integration. This is unsurprising given that biological and socioeconomic factors are considered the primary underlying determinants of malnutrition [43]. However, these results could suggest that the health effects of malnutrition may be compounded by exposure to environmental toxicants, which if identified, could lessen adverse consequences associated with diets of inferior nutritional quality.

Polymarcine, a chemical with the majority of chemical-gene associations coming from HTS, had a substantial increase in chemical-disease and chemical-pathway associations following integration. In contrast, we found less benefit from data integration for benzo[a]pyrene, a chemical with all chemical-gene associations already in CTD. BaP is well-studied and was identified as one of the top 20 compounds most frequently studied in a 2011 bibliometric analysis of research articles studying chemical compounds [13]. Owing to this, BaP has substantial information already contained in CTD, including the same chemical-gene associations identified from the HTS screening of BaP. This indicates that, for chemicals that are well-studied in the literature, HTS provides less benefit as compared to those chemicals like polymarcine that are less frequently studied.

We also identified a potential PD association with polymarcine, an understudied EBDC fungicide, demonstrating the utility of these data integration efforts to promote discovery. By relating the HTS chemical-gene activity to PD and the TNF signaling pathway, a pathway enriched for polymarcine and associated with PD in CTD, we were able to characterize the mechanistic role that polymarcine may play in activating a PD-related signaling pathway. We also demonstrated how identification of enriched pathways using HTS/CTD chemical-gene associations with CTD gene-pathway and pathway-disease associations could provide further mechanistic perspective into the etiology of diseases for classes of chemicals by analyzing pathways associated with PD and enriched for EBDC fungicides. Through this, we identified common pathways and genes that may be implicated in the toxicological response of these chemicals in the development of PD and characterized the chemical-gene responses for the TNF signaling pathway. These approaches demonstrate an application of our data integration efforts to clarify chemical-disease mechanisms of action that can be pursued for further analysis, and these efforts help to advance the goals of both Tox21 and CTD by creating associations between environmental chemicals and their mechanisms of action while connecting these mechanisms to human health.

For chemicals without gene associations in CTD, we demonstrate how HTS data can provide evidence for disease-gene and pathway-gene enrichment. By using the HTS chemical-gene associations with disease-gene and pathway-gene associations in CTD, we demonstrate the mutual enhancements to each dataset following integration. We show how HTS data can fill gaps in CTD owing to gaps in the literature while CTD can provide biological context to HTS results through disease and pathway associations. By forming these new chemical, disease, and pathway associations from existing datasets, we demonstrate the potential for data integration to spur hypothesis generation. Our analyses identified over 300,000 enriched chemical-disease associations. When we evaluated the disease categories, we found cancer to have more disease associations than nervous system disorders. This finding highlights the importance of testing these chemicals for a variety of disease associations, as they may have a role in the etiology of unexpected diseases. We also found over 200,000 chemical-pathway associations highlighting potential chemical-disease-pathway mechanisms for future study of these chemicals.

We provide evidence for the utility of this data integration strategy by demonstrating how linking these new gene, disease, and pathway associations for a chemical can aid in development of AOPs. We found 4-nonylphenol, a chemical with chemical-gene associations from the HTS screening dataset that are missing from CTD, was enriched for both PD and the TNF signaling pathway and demonstrated how linking these chemical-disease and chemical-pathway associations can provide a framework for toxicant-disease processes. By characterizing the connections between these pathways, this data integration technique can identify the connections between key events in AOPs and elucidate the connections between upstream signaling processes and adverse outcomes for understudied chemicals. This method could be replicated for any of the chemical-disease and chemical-pathway associations we identified (Supplemental Tables 2 and 3), providing potential to characterize thousands of new toxicant-disease mechanisms. These new chemical-disease and pathway associations demonstrate the utility of these integration efforts to provide disease and pathway context to understudied chemicals. These chemical-pathway-disease mechanisms can be prioritized for AOP development by identifying chemicals enriched for diseases and pathways of greatest concern. Associations could also be prioritized based on the most significant disease and pathway enrichment p-values. Additionally, by characterizing chemicals with the least redundant pathway-disease associations, unique toxicological responses can be identified. Alternatively, by characterizing AOPs with similar key events and outcomes, chemicals with similar toxicological responses may be identifiable. Future efforts can improve upon these AOPs by relating the key events we characterized to more specific biological responses, such as the phenotype data in CTD [10].

A challenge in our analysis is determining the relative value of each chemical-gene, chemical-disease, and chemical-pathway association. While we only incorporated HTS chemical-gene associations that were active, these data are likely to have inherent errors, such as false positives or chemical-gene associations resulting from activity downstream of the original chemical interaction. Similarly, biases inherent in the literature, including differences in data standards and experimental methods, affect data curation by CTD [44]. As described in our analysis of MEDIC-Slim categories, cancer and nervous system diseases have more chemical and gene associations due to the prevalence of studies of these disease categories in the literature. As such, it is probable that additional chemical-disease associations from less well-described disease categories (e.g. nutrition disorders) were not captured in our enrichment analysis because of the dearth of relevant gene-disease studies and their consequent underrepresentation in CTD. A similar study bias effect is likely at play in chemical-pathway associations owing to unresolved gene-pathway associations in the literature. Future efforts should focus on ways to identify the chemical-gene, chemical-disease, and chemical-pathway associations that are least likely to be influenced by underlying data errors and have the strongest associations. One approach for this would be to prioritize chemical-gene associations that were captured in both the HTS analysis and CTD or to refine the chemical-disease and chemical-pathway enrichment analysis.

Another limitation to our efforts is the small number of genes associated with chemicals in the Tox21 gene library. We propose that our integration efforts can be used to inform inclusion of additional assays to increase coverage of biomedical disease domains. For example, there are 19,216 genes associated with PD that are curated in CTD but were not assayed in HTS data. By using this set of genes to search for assays in resources such as the PubChemBioAssay Database, new assays could be identified targeting these PD-associated genes. Additional efforts could be focused on identifying assays for disease domains with less existing chemical-gene association data across either database (Supplemental Fig. 2). Further, while our study has focused on the utility of our strategy to identify associations between toxicants and adverse outcomes, our method can also be applied to identify new therapeutic routes for chemical-gene-pathway-disease connections. As we demonstrated, the chemical-gene activity for a toxicant can be identified from the HTS chemical-gene data and linked to the pathway-disease data from the union dataset. Therefore, a potential therapeutic may be characterized by identifying drugs that are associated with the same pathway-disease mechanisms but exhibiting opposite activity.

Integrating large datasets, whether from structured databases or unstructured collections of data, requires identification of common data elements. These elements serve as a foundation upon which more complex interrelationships between datasets can be discerned using a variety of data mining techniques. Unique terms or keys such as gene and chemical identifiers that are shared between databases like Tox21/ToxCast and CTD, offer a straightforward means of identifying intersections between the datasets. These shared data can then be used to leverage associated data from both resources and construct broader biological contexts such as pathway and disease relationships [3,45,46]. Our work presents a method for data integration between two distinct datasets and demonstrates the resulting benefits to promote discovery. Additional datasets could be integrated by using our data-matching methods. Many chemical datasets use CASRNs as a chemical identifier, such as the Chemical/Product Categories Database [47] and the Hazardous Substances Database [48]. Other chemical links could be used in place of CASRNs such as Standard IUPAC Chemical Identifier (InChI) representations as found in ChEMBL [49]. The same linkages could be done for genes, diseases, assays, or any other data so long as the datasets of interest contain a common element to link across. We also demonstrate the utility of linking broader biological data to HTS results and provide a method for new HTS data to be annotated.

We highlight the importance of database integration by demonstrating improvements in chemical-gene, chemical-disease, and chemical-pathway coverage following the integration of HTS data with curated data from the literature. Where HTS provided new chemical-gene information to CTD and enhanced CTD's chemical-disease and chemical-pathway connections, CTD provided human health context to HTS results by aligning HTS chemical-gene associations in a disease framework. These efforts can inform new chemical-disease and chemical-pathway hypotheses as well as aid in AOP development by linking these chemical-gene-pathway-disease processes into a structured framework. Future efforts will be aimed at incorporating additional datasets with the CTD/HTS integrated dataset and further understanding the connections between chemical-gene information and human health outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

5. Funding information

This work was supported by the National Institutes of Health [R01 ES023788, R01 ES014065, P30 ES025128, T32 ES007046] and the US Environmental Protection Agency [STAR R83502].

References

[1]. Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, Dellarco V, Henry T, Holderman T, Sayre P, Tan S, Carpenter T, Smith E, The toxicity data landscape for environmental chemicals, Environ. Health Perspect 117 (2009) 685–695, 10.1289/ehp.0800168. [PubMed: 19479008]

[2]. NRC, Toxicity Testing in the 21st Century, The National Academies Press, Washington, DC, 2007. doi:10.17226/11970.

[3]. Mattingly CJ, Boyles R, Lawler CP, Haugen AC, Dearry A, Haendel M, Laying a community-based foundation for data-driven semantic standards in environmental health sciences, Environ. Health Perspect 124 (2016) 1136–1140, 10.1289/ehp.1510438. [PubMed: 26871594]

[4]. Tice RR, Austin CP, Kavlock RJ, Bucher JR, Improving the human hazard characterization of chemicals: a Tox21 update, Environ. Health Perspect 121 (2013) 756–765, 10.1289/ehp.1205784. [PubMed: 23603828]

[5]. Collins FS, Gray GM, Bucher JR, Transforming Environmental Health Protection, Science (80-.). 319 (2008) 906–7. doi:10.1126/science.1154619.Transforming.

[6]. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ, The toxcast program for prioritizing toxicity testing of environmental chemicals, Toxicol. Sci 95 (2007) 5–12, 10.1093/toxsci/kfl103. [PubMed: 16963515]

[7]. Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, Reif DM, Rotroff DM, Shah I, Richard AM, Dix DJ, In vitro screening of environmental chemicals for targeted testing prioritization: The toxcast project, Environ. Health Perspect 118 (2010) 485–492, 10.1289/ehp.0901392. [PubMed: 20368123]

[8]. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D, Richard A, Rotroff D, Sipes N, Dix D, Update on EPA's ToxCast program:

providing high throughput decision support tools for chemical risk management, Chem. Res. Toxicol 25 (2012) 1287–1302, 10.1021/tx3000939. [PubMed: 22519603]

[9]. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ, The Comparative Toxicogenomics Database: Update 2017, Nucleic Acids Res 45 (2017) D972–D978, 10.1093/nar/gkw838. [PubMed: 27651457]

[10]. Davis AP, Wiegers TC, Wiegers J, Johnson RJ, Sciaky D, Grondin CJ, Mattingly CJ, Chemical-Induced Phenotypes at CTD Help Inform the Predisease State and Construct Adverse Outcome Pathways, Toxicol. Sci (2018) 1–12, 10.1093/toxsci/kfy131.

[11]. Grondin CJ, Davis AP, Wiegers TC, Wiegers JA, Mattingly CJ, Accessing an expanded exposure science module at the comparative toxicogenomics database, Environ. Health Perspect 126 (2018) 1–5, 10.1289/EHP2873.

[12]. Grondin CJ, Davis AP, Wiegers TC, King BL, Wiegers JA, Reif DM, Hoppin JA, Mattingly CJ, Advancing Exposure Science through Chemical Data Curation and Integration in the Comparative Toxicogenomics Database, Environ. Health Perspect (2016), 10.1289/EHP174.

[13]. Grandjean P, Eriksen ML, Ellegaard O, Wallin JA, The Matthew effect in environmental science publication: a bibliometric analysis of chemical substances in journal articles, Environ. Health 10 (1) (2011) 96, 10.1186/1476-069X-10-96. [PubMed: 22074398]

[14]. Davis AP, Wiegers TC, Rosenstein MC, Mattingly CJ, MEDIC: A practical disease vocabulary used at the comparative toxicogenomics database, Database 2012 (2012) 1–9, 10.1093/database/bar065.

[15]. CTD, Curated data were retrieved from CTD, MDI Biol. Lab. Salisbury Cove, Maine, NC State Univ. Raleigh, North Carolina, (n.d.). http://ctdbase.org/.

[16]. U.S. EPA, ToxCast and Tox21 Summary Files from invitrodb_v2, (2015). https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data.

[17]. Phuong J, Truong L, Sipes N, Connors K, Houck K, Judson R, Martin M, ToxCast Assay Annotation Version 1.0 Data User Guide, (2014).

[18]. Maglott D, Ostell J, Pruitt KD, Tatusova T, Entrez gene: Gene-centered information at NCBI, Nucleic Acids Res 39 (2011) 52–57, 10.1093/nar/gkq1237.

[19]. Judson R, Houck K, Martin M, Richard AM, Knudsen TB, Shah I, Little S, Wambaugh J, Woodrow Setzer R, Kothya P, Phuong J, Filer D, Smith D, Reif D, Rotroff D, Kleinstreuer N, Sipes N, Xia M, Huang R, Crofton K, Thomas RS, Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space, Toxicol. Sci kfw092 (2016), 10.1093/toxsci/kfw092.

[20]. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G, GO::TermFinder - Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, Bioinformatics 20 (2004) 3710–3715, 10.1093/bioinformatics/bth456. [PubMed: 15297299]

[21]. Gabor C, Nepusz T, The igraph software package for complex network research, Inter Journal Complex Sy (2006) 1695. http://igraph.org.

[22]. Luo W, Brouwer C, Pathview: an R/Bioconductor package for pathway-based data integration and visualization, Bioinformatics 29 (2013) 1830–1831, 10.1093/bioinformatics/btt285. [PubMed: 23740750]

[23]. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B, gplots: Various R Programming Tools for Plotting Data, R Packag. Version 3.0.1 (2016) https://cran.r-project.org/package=gplots.

[24]. Wickham H, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016 http://ggplot2.org.

[25]. R Core Team R: A language and environment for statistical computing, Vienna, Austria, 2016. https://www.r-project.org/.

[26]. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ, The Comparative Toxicogenomics Database: update 2013, Nucleic Acids Res 41 (2013) D1104–D1114, 10.1093/nar/gks994. [PubMed: 23093600]

[27]. Pringsheim T, Jette N, Frolkis A, Steeves TDL, The prevalence of Parkinson's disease: a systematic review and meta-analysis, Mov. Disord 29 (2014) 1583–1590, 10.1002/mds.25945. [PubMed: 24976103]

[28]. Schapira AH, Jenner P, Etiology and pathogenesis of Parkinson's disease, Mov. Disord 26 (2011) 1049–1055, 10.1002/mds.23732. [PubMed: 21626550]

[29]. Wirdefeldt K, Adami H-O, Cole P, Trichopoulos D, Mandel J, Epidemiology and etiology of Parkinson's disease: a review of the evidence, Eur. J. Epidemiol 26 (2011) 1–58, 10.1007/sl0654-011-9581-6. [PubMed: 20845063]

[30]. Uversky VN, Neurotoxicant-induced animal models of Parkinson's disease: Understanding the role of rotenone, maneb and paraquat in neurodegeneration, Cell Tissue Res 318 (2004) 225–241, 10.1007/s00441-004-0937-z. [PubMed: 15258850]

[31]. Gao D, Wu M, Wang C, Wang Y, Zuo Z, Chronic exposure to low benzo[a]pyrene level causes neurodegenerative disease-like syndromes in zebrafish (Danio rerio), Aquat. Toxicol 167 (2015) 200–208, 10.1016/j.aquatox.2015.08.013. [PubMed: 26349946]

[32]. IARC, Chemical agents and related occupations, IARC Monogr 100 F (2010) 111–133. https://www.ncbi.nlm.nih.gov/books/NBK304416/pdf/Bookshelf_NBK304416.pdf.

[33]. Bjørling-Poulsen M, Andersen HR, Grandjean P, Potential developmental neurotoxicity of pesticides used in Europe, Environ. Heal 7 (2008) 50, 10.1186/1476-069X-7-50.

[34]. Grosicka-Maciag E, Kurpios-Piec D, Szumiło M, Grzela T, Rahden-Staroń I, Dithiocarbamate fungicide zineb induces oxidative stress and apoptosis in Chinese hamster lung fibroblasts, Pestic. Biochem. Physiol 102 (2012) 95–101, 10.1016/j.pestbp.2011.11.003.

[35]. Caullet C, Le Nôtre J, Vegetable Oil Biorefineries, in: Ind. Biorefineries White Biotechnol, Elsevier, 2015: pp. 247–270. doi:10.1016/B978-0-444-63453-5.00007-0.

[36]. Tansey MG, Goldberg MS, Neuroinflammation in Parkinson's disease: Its role in neuronal death and implications for therapeutic intervention, Neurobiol. Dis 37 (2010) 510–518, 10.1016/j.nbd.2009.11.004. [PubMed: 19913097]

[37]. Wang Q, Liu Y, Zhou J, Neuroinflammation in Parkinson's disease and its potential as therapeutic target, Transl. Neurodegener 4 (2015) 1–9, 10.1186/s40035-015-0042-0. [PubMed: 25671103]

[38]. Tatton WG, Chalmers-Redman R, Brown D, Tatton N, Schapira, Hunot, Olanow, Isacson, Stocchi, Apoptosis in Parkinson's disease: Signals for neuronal degradation, Ann. Neurol 53 (2003) 61–72, 10.1002/ana.10489.

[39]. Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, Ottinger MA, Vergauwen L, Whelan M, Adverse outcome pathway (AOP) development I: Strategies and principles, Toxicol. Sci 142 (2014) 312–320, 10.1093/toxsci/kfu199. [PubMed: 25466378]

[40]. Okai Y, Sato EF, Higashi-Okai K, Inoue M, Enhancing effect of the endocrine disruptor para-nonylphenol on the generation of reactive oxygen species in human blood neutrophils, Environ. Health Perspect 112 (2004) 553–556, 10.1289/ehp.6584. [PubMed: 15064160]

[41]. Jie X, Li J, Zheng F, Lei G, Biao Z, Jie Y, Neurotoxic effects of nonylphenol: A review, Wien. Klin. Wochenschr 125 (2013) 61–70, 10.1007/s00508-012-0221-2. [PubMed: 23334477]

[42]. Litwa E, Rzemieniec J, Wnuk A, Lason W, Krzeptowski W, Kajta M, Apoptotic and neurotoxic actions of 4-para-nonylphenol are accompanied by activation of retinoid X receptor and impairment of classical estrogen receptor signaling, J. Steroid Biochem. Mol. Biol 144 (2014) 334–347, 10.1016/j.jsbmb.2014.07.014. [PubMed: 25092517]

[43]. Silva P, Environmental factors and children's malnutrition in Ethiopia: World Bank policy research working paper 3489, 2005. http://library1.nida.ac.th/worldbankf/fulltext/wps03489.pdf.

[44]. Davis AP, Murphy CG, Rosenstein MC, Wiegers TC, Mattingly CJ, The Comparative Toxicogenomics Database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study, BMC Med. Genomics. 1 (2008) 1–12, 10.1186/1755-8794-1-48. [PubMed: 18237448]

[45]. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC,

Richardson JE, Ringwald M, Rubin GM, Sherlock G, G.O. Consortium, Gene Ontology: Tool for The Unification of Biology, Nat. Genet 25 (2000) 25–29, 10.1038/75556. [PubMed: 10802651]

[46]. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M, KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res 40 (2012) 109–114, 10.1093/nar/gkr988.

[47]. Dionisio KL, Frame AM, Goldsmith MR, Wambaugh JF, Liddell A, Cathey T, Smith D, Vail J, Ernstoff AS, Fantke P, Jolliet O, Judson RS, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, Toxicol. Rep 2 (2015) 228–237, 10.1016/j.toxrep.2014.12.009. [PubMed: 28962356]

[48]. US DHHS, Hazardous substances data bank (HSDB, online database), Natl. Toxicol. Inf. Program, Natl. Libr. Med. Bethesda, MD. (1993).

[49]. Gaulton A, Beilis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res 40 (2012) 1100–1107, 10.1093/nar/gkr777.

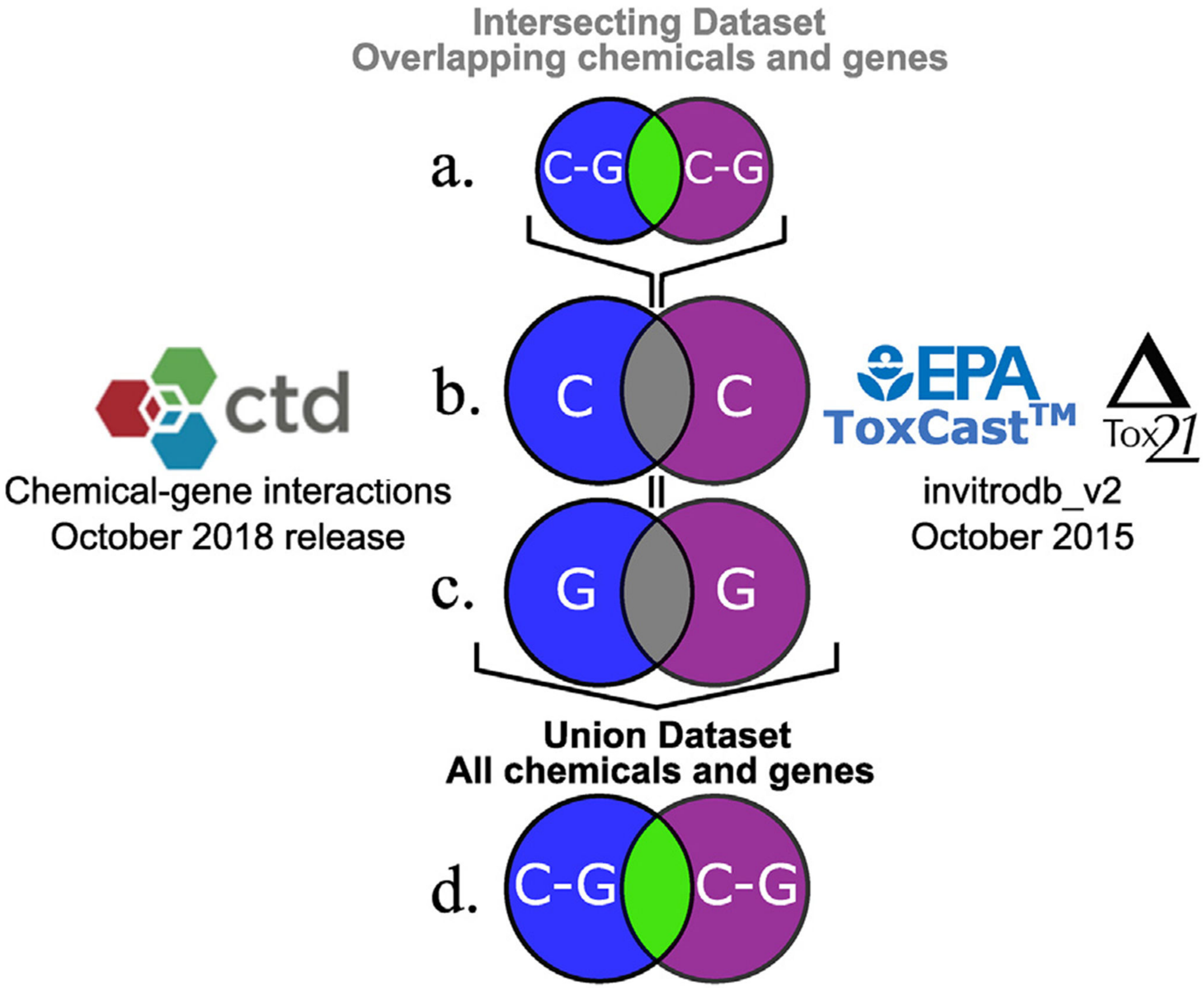

Fig. 1.

Overview of the data integration process. C = Chemical, G = Gene. C-G = Chemical-gene association. Blue = Genes/chemicals/chemical-gene associations found only in CTD, Purple = Genes/chemicals/chemical-gene associations found only in HTS, Grey = Genes/chemicals found both in CTD and HTS, Green = Chemical-gene associations found in both CTD and HTS. (A) Intersecting dataset chemical-gene associations. Formed from overlapping chemicals and genes in B and C. (B) Chemical integration. (C) Gene integration. (D) Union dataset chemical gene associations. Formed from all chemicals and genes in B and C.

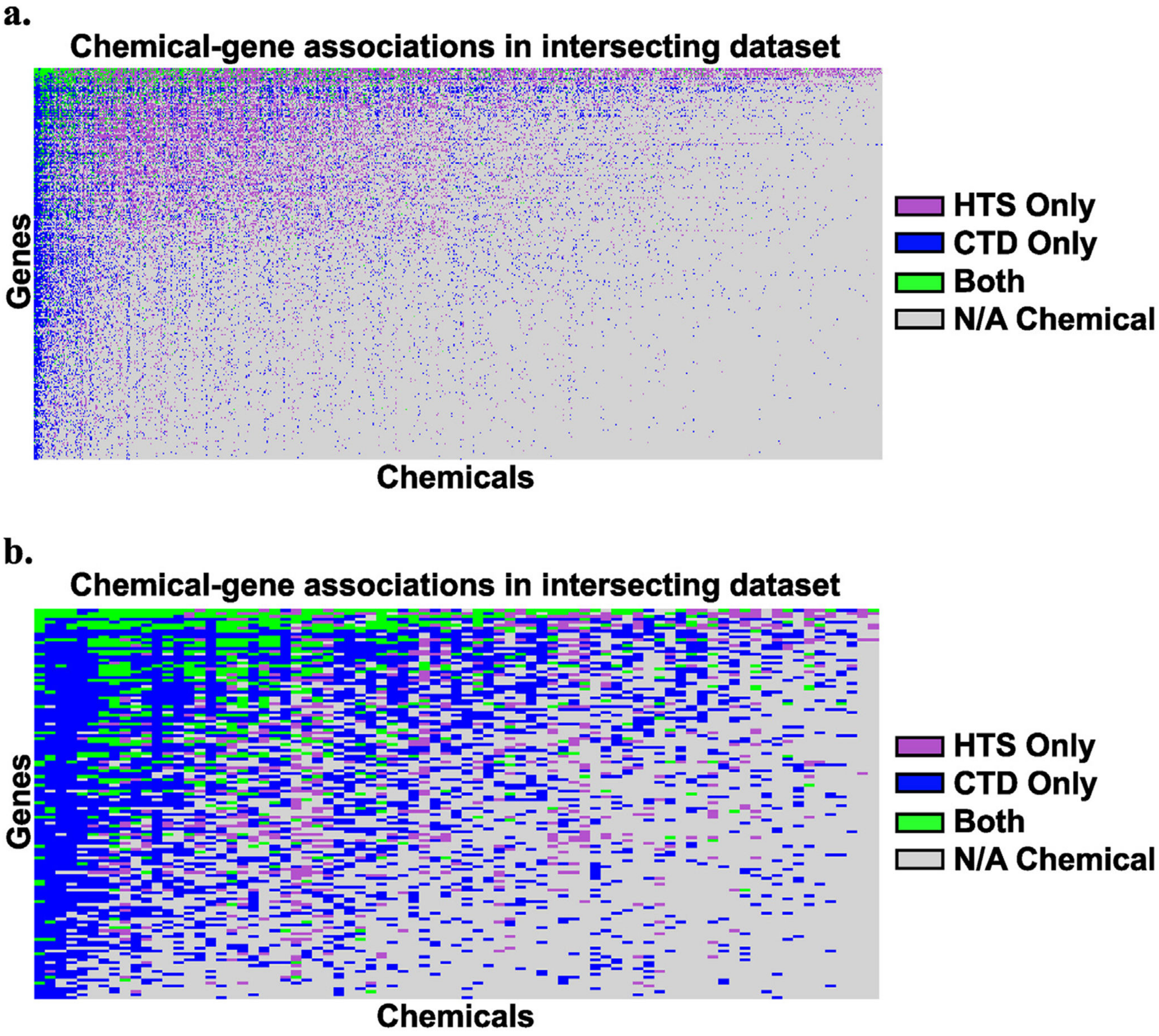

Fig. 2.

Chemical-gene associations are varied between sources for different diseases. Chemical-gene associations for chemicals and genes common to both HTS and CTD (the intersecting dataset) associated with Parkinson's disease or malnutrition in CTD. Rows = Genes. Columns = Chemicals. For each chemical in the intersecting dataset, the gene was originally mapped to that chemical in: Purple = HTS, Blue = CTD, Green = Both databases, Grey = No gene association. (A) Parkinson's Disease. (B) Malnutrition.

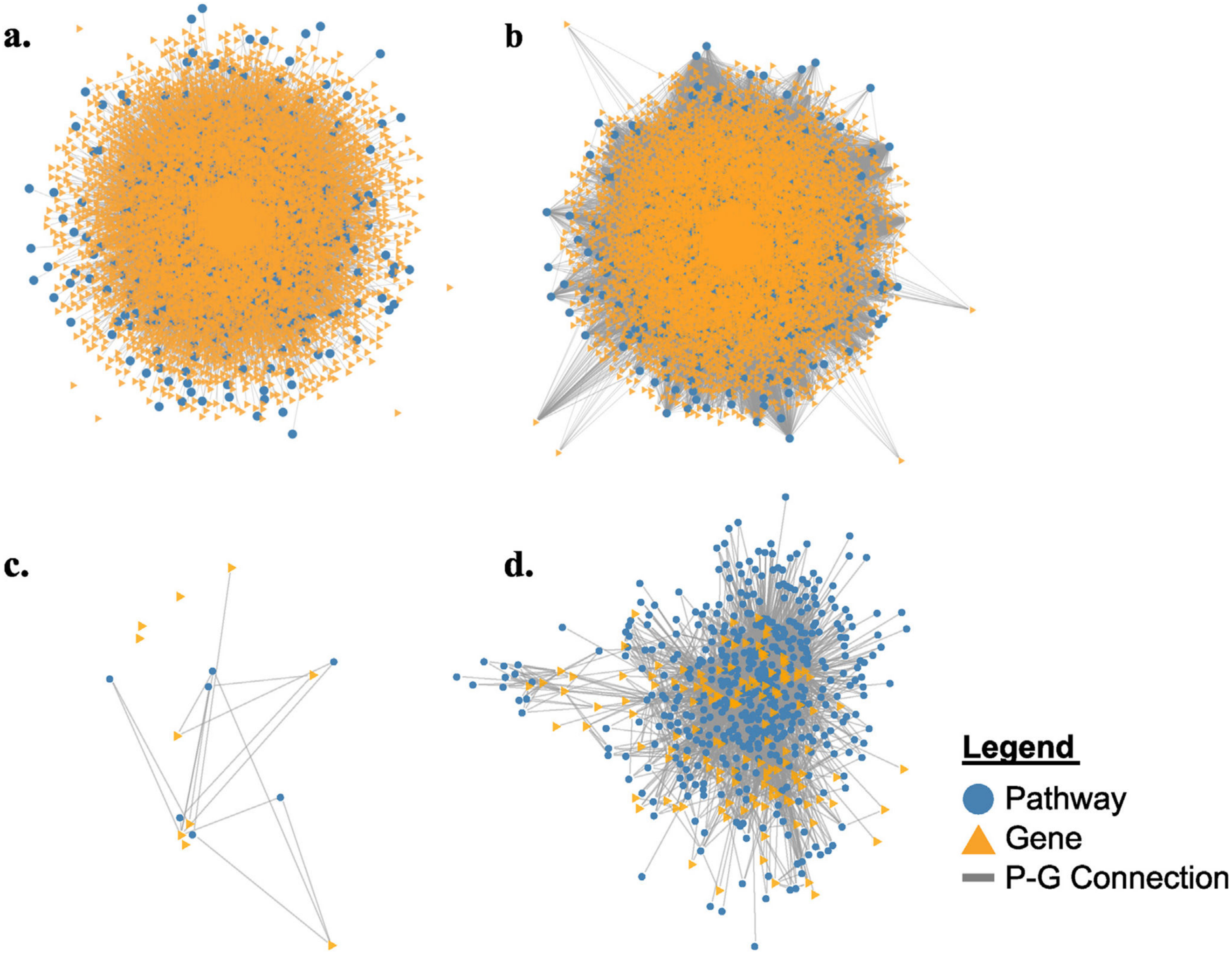

Fig. 3.

Integration of HTS and CTD augments pathway coverage for chemicals. Network diagrams of gene-pathway connections for benzo(a)pyrene and polymarcine. (A) BaP, CTD-only genes and pathways. (B) BaP, union dataset genes and new, enriched pathways. (C) Polymarcine, CTD-only genes and pathways. (D) Polymarcine, union dataset genes and new, enriched pathways.

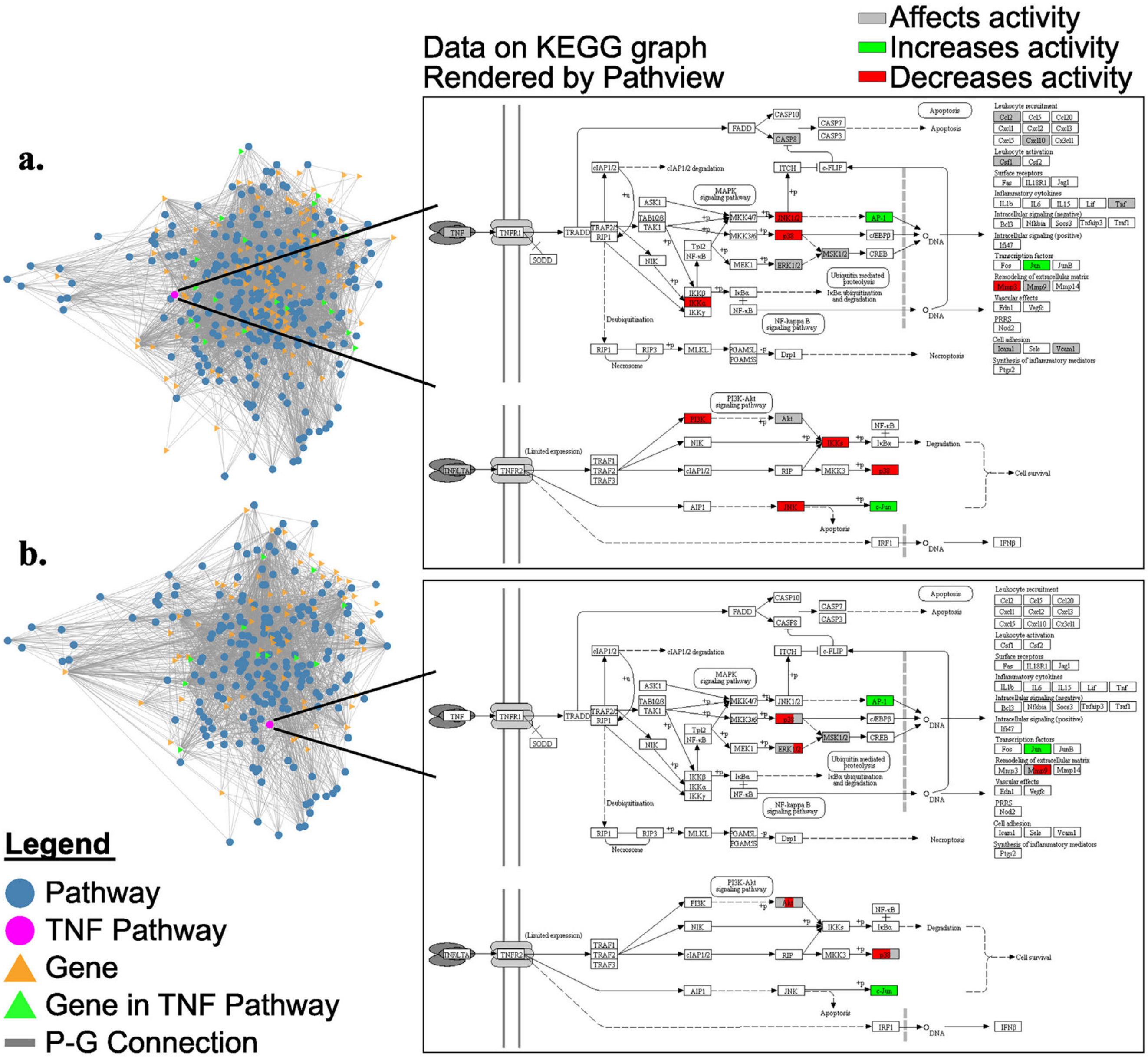

Fig. 4.

Integration of HTS and CTD informs disease-pathway connections. Left: Network diagrams of gene-pathway connections for pathways associated with Parkinson's disease in CTD. Networks generated for (A) Polymarcine, (B) EBDC fungicide class, from genes in the union dataset (HTS + CTD). Right: Mechanistic view of TNF signaling pathway generated using Pathview. Genes color coded by direction of activity determined from chemical-gene association in HTS assay. (A) Polymarcine, gene-pathway connections from Fig. 3d narrowed down to just PD connections. (B) EBDC fungicides (polymarcine, maneb, mancozeb), gene-pathway connections from Fig. 3.6a narrowed down to just polymarcine and other EBDC fungicides. Activity in the TNF signaling pathway is shown for all three chemicals colored by: Left = Polymarcine, Middle = Maneb, Right = Mancozeb.

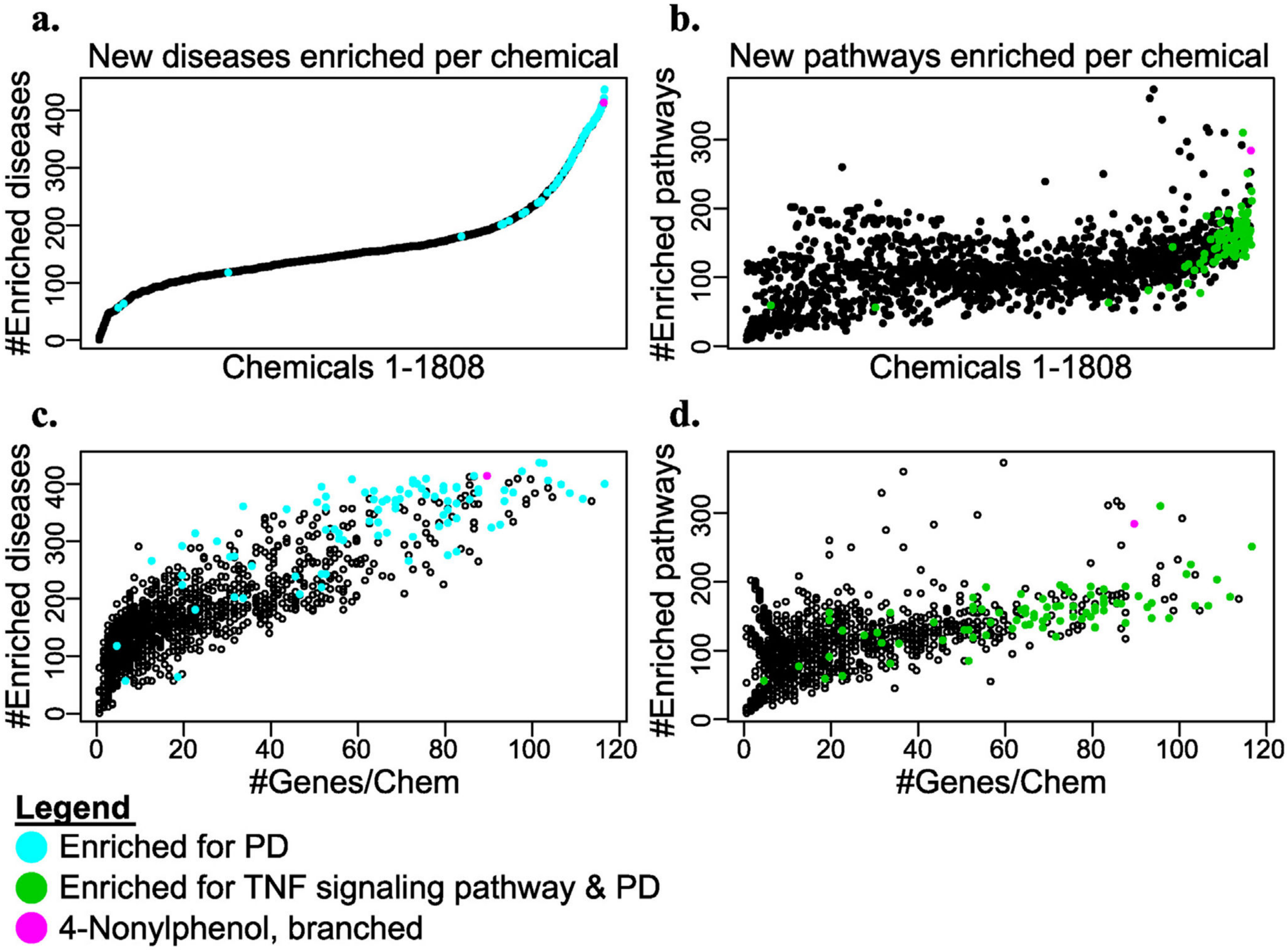

Fig. 5.

Integration of HTS and CTD identifies new chemical-disease and chemical-pathway associations. Left panels = disease associations, Right panels = pathway associations. Top: Enriched disease or pathway associations for chemicals with only HTS gene associations in union dataset. Chemical order on x-axis is the same in both figures. Bottom panels: Number of HTS genes associated with each chemical in union dataset versus the number of new disease or pathway associations.

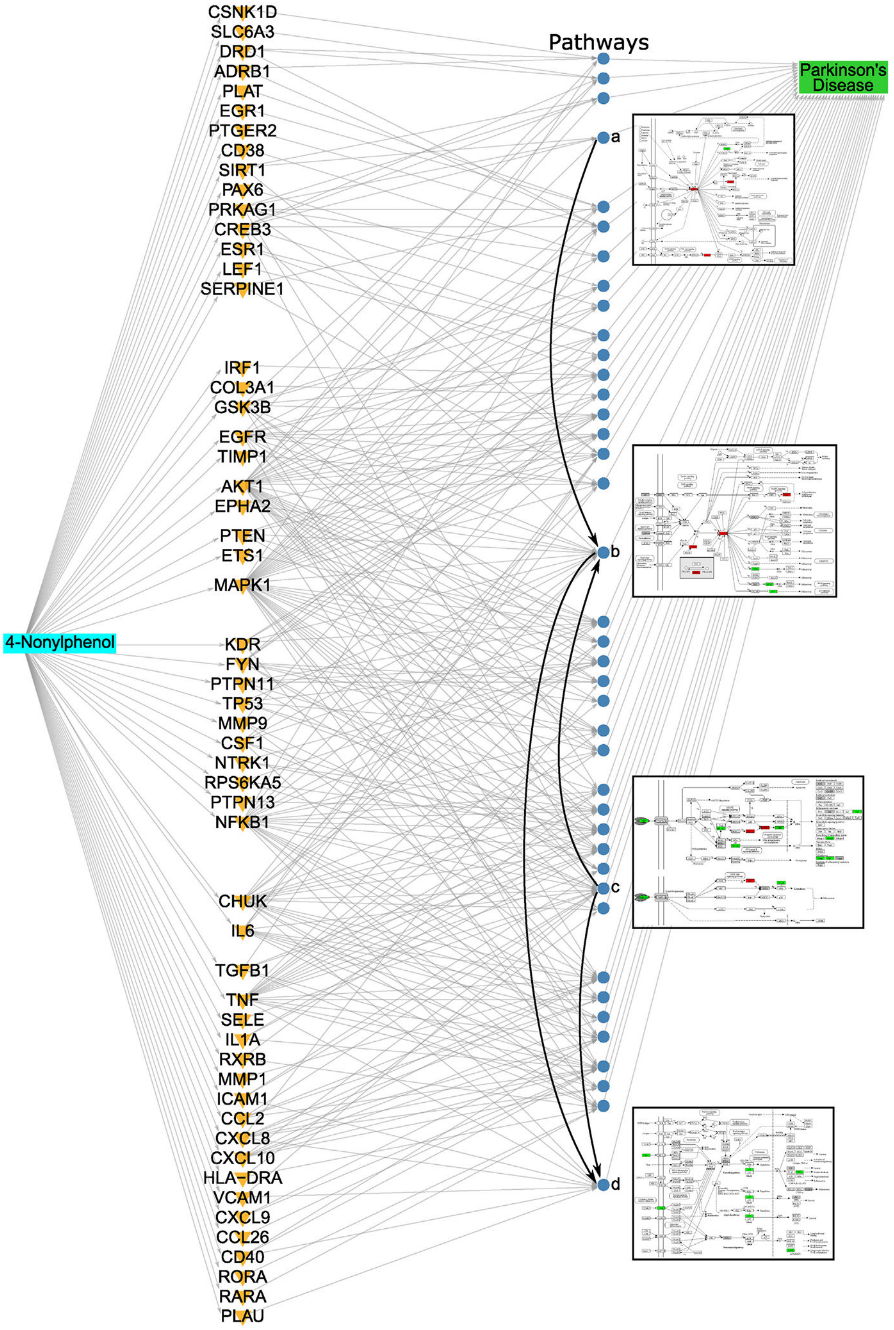

Fig. 6.

Integration of HTS and CTD aids in development of adverse outcome pathways.

Adverse outcome pathway for 4-nonylphenol, branched. Orange triangles = Chemical-gene associations from HTS. Blue circles = Pathways associated with PD in CTD and enriched for 4-Nonylphenol. Black arrows = Connections between pathways. Pathway (A) AMPK signaling pathway, (B) PI3K-Akt signaling pathway, (C) TNF signaling pathway, (D) NF-κB signaling pathway. Direction of gene activity as determined from HTS assays: Green = Increases activity, Red = Decreases activity, Grey = Affects activity.

Table 1

Integration of HTS and CTD augments disease coverage for chemicals. Sources of gene and disease associations for the chemicals benzo(a)pyrene and polymarcine. Gene data: number of genes from HTS, CTD, or both data sources. Disease data: number of diseases originally associated with chemical in CTD, enriched diseases identified using HTS + CTD genes, and how many of these enriched diseases were not already associated with CTD.

	Benzo(a)pyrene	**Polymarcine**
# Genes	**11,894**	**147**
HTS only	0	136
CTD only	11,881	6
Both	13	5
# Diseases	**4,520**	**442**
CTD	4,520	135
Enriched	3,707	411
New, Enriched	0	307