Cluefish: mining the dark matter of transcriptional data series with over-representation analysis enhanced by aggregated biological prior knowledge

AUTHORS

Ellis FRANKLIN[1,2,*], Elise BILLOIR[1], Philippe VEBER[2], Jérémie OHANESSIAN[1], Marie Laure DELIGNETTE-MULLER[2], Sophie Martine PRUD'HOMME[1,*]

[1] Université de Lorraine, CNRS, LIEC, F-57000 Metz, France

[2] Université Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VAS, F-69622 Villeurbanne, France

* To whom correspondence should be addressed. Tel: +33 07 62 32 63 10; Email: ellis.franklin@univ-lorraine.fr. Correspondence may also be addressed to Sophie Martine PRUD'HOMME. Tel: +33 06 49 82 76 27; Email: sophie.prud-homme@univ-lorraine.fr.

Present Address: [Ellis FRANKLIN], Université de Lorraine, CNRS, LIEC, F-57000 Metz, France

GRAPHICAL ABSTRACT

A graphical abstract will be provided at revision.

ABSTRACT

Interpreting transcriptomic data presents significant challenges, particularly in non-targeted approaches. While modern functional enrichment methods are well-suited for experimental designs involving two conditions, they are less applicable to data series. In this context, we developed Cluefish, a free and open-source, semi-automated R workflow designed for untargeted, comprehensive biological interpretation of transcriptomic data series. Cluefish applies over-representation analysis on pre-clustered protein-protein interaction networks, using clusters as anchors to identify smaller, more specific biological functions. Innovative features, including cluster merging and recovery of isolated genes through shared biological contexts, enable a more complete exploration of the data. In our case study with zebrafish embryos exposed to a dose-gradient of dibutyl phthalate, Cluefish—combined with DRomics, a tool for dose-response analysis—identified gene clusters deregulated at low doses and linked to biological functions overlooked by the standard approach. Notably, it revealed that retinoid signalling disruption may be the most sensitive pathway affected by dibutyl phthalate during zebrafish development, potentially leading to morphological

changes. The Cluefish workflow aims to provide valuable clues for biological hypothesis generation and experimental validation. It is freely available at https://github.com/ellfran-7/cluefish.

INTRODUCTION

Advances in omics approaches has made the comprehensive measurement of DNA, RNA, proteins and metabolites in biological samples a standard practice (1). This has led to a massive increase in generated data, providing a goldmine for uncovering novel biological functions, genotype-phenotype relationships, and elucidating molecular mechanisms (2). However, the sheer volume presents significant challenges in both analysis and interpretation (3, 4).

In the context of transcriptomics, analyses yield extensive transcript lists that require impractical manual literature reviews. To address this, functional enrichment analysis, also known as pathway enrichment analysis, has become the standard approach. This method condenses large gene lists into more manageable and interpretable sets of biological functions or pathways (5). Functional enrichment relies on predefined gene sets representing molecular functions, biological processes or pathways, as outlined by databases like the Gene Ontology (GO) (6), pathway databases such as KEGG (7) and Wikipathways (8, 9), or experimentally derived gene sets available in the MSigDB (10, 11).

Functional enrichment methods have evolved over four generations. The first generation consists of Over-Representation Analysis (ORA) methods, which test whether a predefined gene set contains a disproportionately large number of deregulated genes compared to a uniform random sample of genes among all the genes detected in the experiment. In this case, the functional gene set is considered as enriched. The second generation includes Functional Class Scoring (FCS) methods, which evaluate whether genes within a set are concentrated at the top or bottom of a ranked list of all the genes from the experiment, based on expression changes such as fold-change (11). FCS improves on ORA by considering coordinated changes in functionally related gene sets, though it still treats genes as independent (12). The third generation introduced Pathway Topology (PT)-Based methods, which leverage pathway topology information from databases such as KEGG and Reactome (13). These methods utilize knowledge about gene positions, roles and interactions within

pathways (14, 15). While PT-based methods improve on ORA and FCS by incorporating gene dependencies and interactions, they still view pathways as isolated entities and thus overlook pathway crosstalk (16). The most recent and fourth generation, Network Topology (NT)-based approaches, goes a step further by systematically considering pathway crosstalk within a network (17). NT-based methods utilize biological interaction networks, such as those based on protein-protein interactions (PPI) from databases like FunCoup and STRING (18, 19). These approaches examine changes in network structure and correlations to provide a more integrated understanding of pathway interactions, accounting for crosstalk and overlap between pathways—elements neglected by previous generations (20). However, despite their potential, NT-based methods are limited. This is because biological networks are often incomplete, with many missing or unreliable interactions, and these methods tend to focus solely on the network itself, which restricts the effectiveness and completeness of the analyses (17). As a result, earlier generations of functional enrichment methods still remain largely used in practice.

While these earlier methods are commonly applied to transcriptomic data with two or three conditions, dealing with data involving many ordered conditions—referred to as "data series"—poses substantial complexity, even before functional enrichment. Data series capture measurements collected over successive points in time, space, or under varying conditions (e.g., different doses of a contaminant). In dose-response (DR) designs, which explore the relationship between exposure and effect, differential gene expression (DEG) analysis, although often used, is not well-suited. DEG relies on the comparison of the response at each dose to the control (or sometimes on pairwise comparisons across all conditions), multiplying the number of tests, applying each to a subset of the data, and generating multiple fold-change values for each transcript. In contrast, DR omics analysis replaces DEG by leveraging the entire DR relationship for each transcript (21). Tools, like the DRomics R package (22), identify significantly deregulated transcripts and model their DR relationships. These models are used to characterize the response (whether monotonic or biphasic) and to calculate a benchmark dose (BMD) for each transcript. A BMD is a dose or concentration that produces a predetermined change in the response rate of an adverse effect, reflecting the sensitivity of a transcript to a given stressor (22). However, FCS, and PT-based enrichment, which rely on fold-change values, are not adapted for functional

enrichment in a DR context. (12). In contrast, ORA is currently the only suitable method available to complement DR modeling, as it exclusively uses deregulaged gene lists. However, ORA comes with several other limitations. The enrichment results, often derived from Fisher's Exact Tests, tends to highlight broad biological functions. Indeed when using the entire list of deregulated genes, the large test denominator makes it more challenging to achieve the statistical significance needed to identify smaller or lesser-known functions. This tendency is intrinsically tied to the granularity of gene sets in certain databases (e.g., GO). Gene sets may be overly specific, while others are too general, potentially compromising the relevance of the analysis (23, 24). Additionally, the enrichment results represent only a fragmented portion of the deregulated gene list, as the interpretation is based solely on the biological functions and the genes participating in their enrichment. Consequently, relying solely on ORA, as done in the standard approach, is not sufficient. Additional methods are needed in combination to address the shortcomings of the standard approach.

Our main goal was to create a workflow that allows to include a greater part of the deregulated gene list in the interpretation, capturing a more complete picture of the transcriptomic data. We also aimed to improve functional enrichment detection sensitivity to identify more specific and smaller biological functions while minimizing redundancy. Furthermore, we wanted to ensure the workflow is applicable to a wide range of both model and non-model organisms.

To meet these goals, we constructed Cluefish (Clustering, Enrichment and Fishing), a semi-automated workflow designed for comprehensive and untargeted exploration of transcriptomic data series. We developed Cluefish using a DR transcriptomic dataset from zebrafish embryos exposed to varying doses of dibutyl phthalate (DBP), a plasticizer found in human biological samples (25) and environmental surface waters (26). This dataset allowed us to fully test and apply the workflow, conducting an in-depth biological interpretation. Cluefish is open-source and available under the CeCILL license, offering a powerful tool for improving biological interpretation and aiding in the development of targeted interventions.

MATERIAL AND METHODS

Zebrafish dataset

Fish. Zebrafish (*Danio rerio*) embryos (AB strain) were obtained from the Luxembourg Centre for Systems Biomedicine (LCSB) aquatic platform (Luxembourg University, Belval, Luxembourg). They were collected within 1 hour after spawn, placed in standardized fish water at 26°C (NF EN ISO 7346-2 (1998), pre-oxygenated by bubbling for 12h) and transferred to the LIEC laboratory facilities (Metz, France).

DBP solutions preparation and DBP quantification. Commercial DBP (purity 99%, CAS no. 84-74-2) was purchased from Sigma-Aldrich (St. Louis, MO, USA). Solvent-free DBP stock solution was made in fish water at an effective concentration of 5.4 mg/L. Stock solution and fish water effective DBP concentration quantification were performed 1 day before the beginning of exposure (3 technical replicates) on an Ultimate 3000 Rs chromatography system with UV detection at 230 nm (DAD) (Dionex, CA, USA). Quantification was performed on an injected volume of 10 µL, using an Acquity UPLC HSS T3 2.5 µm, 2.1x100 mm column (Waters, Germany) as stationary phase, and 60% acetonitrile and 40% H2O as mobile phase, under a flow of 8 ml/min at 25°C. Fish water DBP quantification were below the limit of detection (LOD) of 3 µg/L. The stock solutions were stored in the dark at 26°C for the all experiment, and were stable for at least the duration of exposure under these conditions (Supplementary Table S1). Tests solutions of DBP were made daily in fish water by serial dilution of the DBP stock solution in oxygenated fish water during the whole experiment to reach a concentration gradient [0 – 5 – 10 – 50 – 100 µg/L] built from environmental concentrations (median to high) up to the NOEC of DBP for chronic exposure in fish (100 µg/L) according to the European CHemicals Agency (ECHA). To avoid any external contamination of test solutions by phthalates or other plasticizers, only glassware was used for fish water, stock and test DBP solutions preparation, handling and storage.

Zebrafish embryos exposure and sample collection. 30 zebrafish embryos (256 cells stage (27)) were randomly distributed in glass beakers with 60 mL of test solution, with 5 replicate beakers of 30 embryos for each exposure condition. Embryos were kept in a climate-controlled chamber (Cooled incubator MIR-554-PE, PHC Europe B.V.) at 26 (±0.5) °C, under 14:10 light/dark photoperiod (20 µE/m^2/s). Exposure was semi-dynamic, with 1/3 of exposure

solution volume being replaced every day by freshly prepared test solutions. After 120h exposure to DBP, pools of 10 embryos ('protruding mouth' stage (27)) were collected as one sample, flash-frozen in liquid nitrogen, and stored at −80 °C before RNA extraction. Remaining embryos were fixed in 4% paraformaldehyde (PFA) in PBS at 4 °C for 1 week and washed several times with PBS before morphological assessment.

Transcriptome sequencing. Total RNA of zebrafish embryos (pools of 10 embryos, 5 replicates per condition) was extracted using the RNeasy mini extraction kit (Qiagen). After quality and integrity assessment of extracted RNA by NanoDrop (Thermo Fisher Scientific, Waltham, MA) and by the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA), the 3 higher quality RNA samples per condition (A260/A280 > 1.7, A260/A230 > 2, RIN >8.8) were selected for high throughput mRNA sequencing. mRNA libraries were prepared from 250 ng total RNA using UltraTMII Directional RNA library Prep kit (New England Biolabs) and library size was checked by the DNA 1000 Assay Kit of the Bioanalyzer 2100 system. Libraries were sequenced on a NextSeq500 Illumina sequencer with 1x 75 bp single reads, providing 40 to 57 million reads for each sample. Library preparation, quality check and sequencing were performed by Helixio, France.

mRNAseq data processing. Raw mRNAseq data processing was performed on the Galaxy french server provided by the 'Institut Français de Bioinformatique' (IFB). Raw reads were filtered according to size (>55 bp) and quality parameters (mean Phred score >20 on 5' and 3' bases and on a 4 pb sliding window) using Trimmomatic (28), leading to keep 87 to 92% of total reads according to the sample. Cleaned reads were mapped to the zebrafish reference transcriptome GCRz11 and quantified using Salmon quant (v. 1.5.1) (29).

Embryos morphological traits. Lateral images of the fixed embryos were taken using a digital microscope (VMX-7000, KEYENCE, France). Body length and left eye area were measured using the "VHX-6000_950F Measurement Data Tabulation Tool" image processing software (KEYENCE, France).

Dose-response analysis

The DRomics R package was used to identify significantly deregulated transcripts and model DR curves. The default parameters in DRomics were used (30). Transcript count datasets were

imported, normalized, and transformed into a logarithmic scale using the rlog method within DRomics (31). Significantly deregulated transcripts were selected using quadratic trend tests with a False Discovery Rate (FDR) of 0.05 (32). DR models (linear, Hill, exponential, Gauss-probit, or log-Gauss-probit) were fitted using non-linear regression and selected based on the lowest second-order Akaike information criterion (33). The modelled DR curves were characterized by their trends (bell-shaped, decreasing, increasing, and U-shaped) and yielded a benchmark dose (BMD-1SD) (22), calculated from a benchmark response (BMR-1SD) using the following equation: $BMR\text{-}1SD = y0 \pm SD$, where y0 is the theoretical level at the control given by the DR model and SD is the residual standard deviation of the modelled DR curve (22, 34). The 95% confidence intervals on the BMD-1SD values were computed by non-parametric bootstrap (5000 iterations). Modelled DR curves were discarded if the point or the interval estimates of the BMD could not be successfully calculated. Body length and left eye area were likewise analysed using the DRomics workflow for continuous anchoring (individual) data.

The Cluefish workflow

Cluefish is composed of 11 main data analysis steps, followed by optional data visualization steps (Figure 1). Each step is as follows:

Step 1: Download transcription (co-)factor annotations. Transcription factors (TFs) and transcriptional co-factors (CoTFs) form multi-protein complexes that mediate cellular signaling and facilitate gene transcription (35, 36). This combinatorial function establishes a crucial link between signaling pathways and gene regulation (37). To incorporate this regulatory information, lists of TF and CoTF genes for the specified organism are obtained from the Animal Transcription Factor Database 4.0 (AnimalTFDB4) (38). If these files are not available for a particular organism, this step can be skipped. The zebrafish TF and CoTF gene lists were successfully downloaded.

Step 2: Load background and deregulated transcript lists. The background transcript list includes all transcript IDs detected in the experiment, whereas the deregulated transcript list contains only the IDs of significantly deregulated transcripts. The zebrafish deregulated transcript list was derived from the DRomics pipeline in our case study.

Step 3: Retrieve gene identifiers. Transcript identifiers (IDs) are converted to gene IDs and other pertinent annotations by linking with the Ensembl database via the biomaRt R package (39, 40). To ensure compatibility with databases and tools used later in Cluefish, specifically the STRING database and g:Profiler (41), the results must include identifiers supporting the organism in both platforms. In most cases, the Ensembl gene ID meets this requirement for well-referenced organisms, significantly reducing the need for repeated ID conversions throughout the workflow. In this step, all detected transcript IDs are mapped to their corresponding Ensembl gene IDs, forming the foundation for the deregulated gene list. In our case, the zebrafish transcript IDs were converted to their Ensembl stable gene ID and external gene name, corresponding to readable gene names.

Step 4: Determine deregulated gene regulatory status. The next step involves determining the regulatory status of the deregulated genes, using the TF and CoTF annotation files acquired earlier. If these files are not available for a particular organism, this step can be skipped. Since the files were successfully downloaded for zebrafish, it was possible to assess the regulatory status of the deregulated genes.

Step 5: Construct and retrieve the clustered protein-protein interaction (PPI) network. The preceding output is then manually imported into the Cytoscape application (42). Using the StringApp plugin (43), the deregulated transcripts' STRING-compatible IDs are entered into the Protein Query search bar. This query process is customizable, offering inclusion of specific interaction types, setting limits for interaction confidence scores, and supplementary interactions. The default value in STRING is set at 0.4 for the confidence score limit to construct the PPI network, representing the probability that the interaction really exists given the available evidence. In our case study, a confidence score of 0.9 was utilized to only consider highly confident interactions, and we included all interaction types. The resulting PPI network was constructed from both known and predicted protein-protein associations from the STRING database. To identify clusters, the Markov Chain Clustering (MCL) algorithm is applied using the clustermaker2 extension (44, 45). The granularity of clusters is controlled by the inflation parameter, with lower values (1-2) yielding fewer, larger clusters, and higher values (5-6) resulting in more, smaller clusters. Very large clusters may lack coherent biological context, while very small clusters may not represent a biological complex or may only be involved in specific cases. In our case study, this parameter was set at 4 which is the

default value in clustermaker2. This setting balances the creation of few large clusters and many small clusters. The node table, containing clustered elements, is manually exported as a CSV file.

Step 6: Filter clusters. Returning to the R environment, clusters are filtered based on their gene set sizes using what is called the lower cluster size filter, under which clusters are removed. By default in Cluefish, this filter is set at 4. The choice of lower cluster size limit may depend on the study, organism, and earlier decisions during PPI network creation and clustering. Higher size limits result in fewer large clusters, whereas smaller limits retain both large and small clusters, removing only very small clusters. In our case study, the default value was used.

Step 7: Perform cluster-wise functional enrichment. To characterize each cluster, ORA can be conducted using the gprofiler2 R package (46), which interfaces with g:Profiler. This tool was chosen firstly for its extensive support of both model and non-model organisms, totaling 984 and the ability to use custom annotation data. The main tool of interest in g:Profiler is g:GOst, which can perform ORA on multiple gene lists and search against a wide range of functional databases, including GO, KEGG and WP. Specifically, Fisher's one-tailed tests can be applied to measure the likelihood that the overlap between a query gene set and biological function gene sets is due to chance. Another important aspect of g:Profiler is its access to previous versions of the software and data, ensuring reproducible analyses. Furthermore, it offers a filtering method for highlighting GO driver terms by considering the underlying topology of annotations, as described in Kolberg et al. 2023 (41). By default in Cluefish, the deregulated gene list is concurrently analyzed against GO Biological Processes (GO:BP), KEGG and WP pathways, using the background gene list comprising all annotated protein-coding genes from the experiment. The FDR multiple testing correction method is applied, with a significance limit set at 0.05. To address the inherent complexity and redundancy within GO term hierarchies, only highlighted GO:BP terms are retained. In our case study, the default settings were chosen. In addition to these, we implemented two new filters.

The first one, called the biological function size filter, are the minimum and maximum gene set sizes of biological functions, i.e., the number of genes associated with each biological

function found within the background gene list. By default in Cluefish, lower and upper biological function size limits are set at 5 and 500 respectively. Very large gene sets, which are dependent on the organism, are often associated with broad biological processes. These can introduce noise and obscure more specific functions. Conversely, very small gene sets can be overly specific and may not capture the full biological context of a pathway. These lower and upper limits can be adjusted based on the study, the organism, and the type of transcriptomic data (e.g., whole transcriptome or tissue-specific). In our case study, the default values were used.

The second filter, called the enrichment gene count filter, is the minimum number of genes participating in the enrichment of a biological function. By default in Cluefish, the enrichment gene count filter is set at 3 in order to avoid identifying biological functions enriched by too few genes, particularly in smaller clusters. Such clusters can lead to a higher rate of false positives, and enrichments based on very few genes may not provide substantial biological insight into the cluster itself. This limit should be proportionate to the cluster sizes, particularly the smallest clusters. In our case study, the default value was used.

Step 8: Merge clusters. After characterizing the clusters, the next step is to identify and merge clusters that share the same enriched biological functions from at least one specified database (e.g. GO, KEGG etc.) in the ORA. By default in Cluefish, the merging process is conducted separately for each chosen database, in descending order of the total number of unique genes present in both the database and the background list. This step combines clusters whose proteins were insufficiently interactive in STRING to form a single cluster but are shown through functional enrichment to participate in the same biological processes. Merged clusters thus form larger, contextually unified groups representing distinct biological functions. In our case study, the default settings were chosen, with clusters first merged based on shared GO Biological Process (GO:BP) functions first, followed by KEGG, and finally WP.

Step 9: Fish lonely genes. Genes not assigned to any cluster are referred to as « *lonely genes* ». These genes are not typically explored because they do not contribute to functional enrichment, and thus, they do not help in characterizing the data. Genes become *lonely* in one of two cases: either they are initially not sufficiently interactive with another protein

complex to participate in or form a cluster, or they were initially part of a cluster but failed to pass the cluster size filtering step, thus falling into the *lonely cluster*. To achieve a more comprehensive analysis and expand clusters, *lonely genes* are incorporated into existing clusters based on shared functional annotations. Specifically, if a biological function annotation of a *lonely gene* matches a biological function enriched in a cluster, the *lonely gene* is integrated into the corresponding cluster. The *friendliness* metric represents the number of clusters a *lonely gene* can be incorporated into.

Step 10: Perform functional enrichment of the lonely cluster genes. The *lonely cluster*, now consisting of the remaining *lonely genes*, can undergo a simple functional enrichment analysis to gain further insight into the biological context of the genes not associated to any existing clusters. The customizations used in the earlier cluster-wise enrichment analysis are fully applicable here as well. Therefore, the parameters were set to match those chosen in the previous cluster-wise functional enrichment analysis.

Step 11: Generate outputs for exploration. To fully explore the Cluefish results, DRomics results are required in order to include DR modeling metrics. While DR analysis is not required for previous steps of Cluefish, it becomes essential for generating the following outputs. A summary table, which captures key details from each step and tool used in the Cluefish workflow, can be generated and exported to CSV for manual exploration. Additionally, plots of fitted DR curves for each cluster's gene set can be generated and exported to PDF, offering a visual exploration of cluster content. Three Quarto (47) report templates are incorporated into the Cluefish compendium, which may be useful to users. Examples can be found in Supplementary Report Examples 1-3. These templates summarize Cluefish results interactively, integrate them with DRomics findings, and enhance data usability for comprehensive biological interpretation. Although these reports can be generated automatically, they may require adjustments to suit various types of cases.

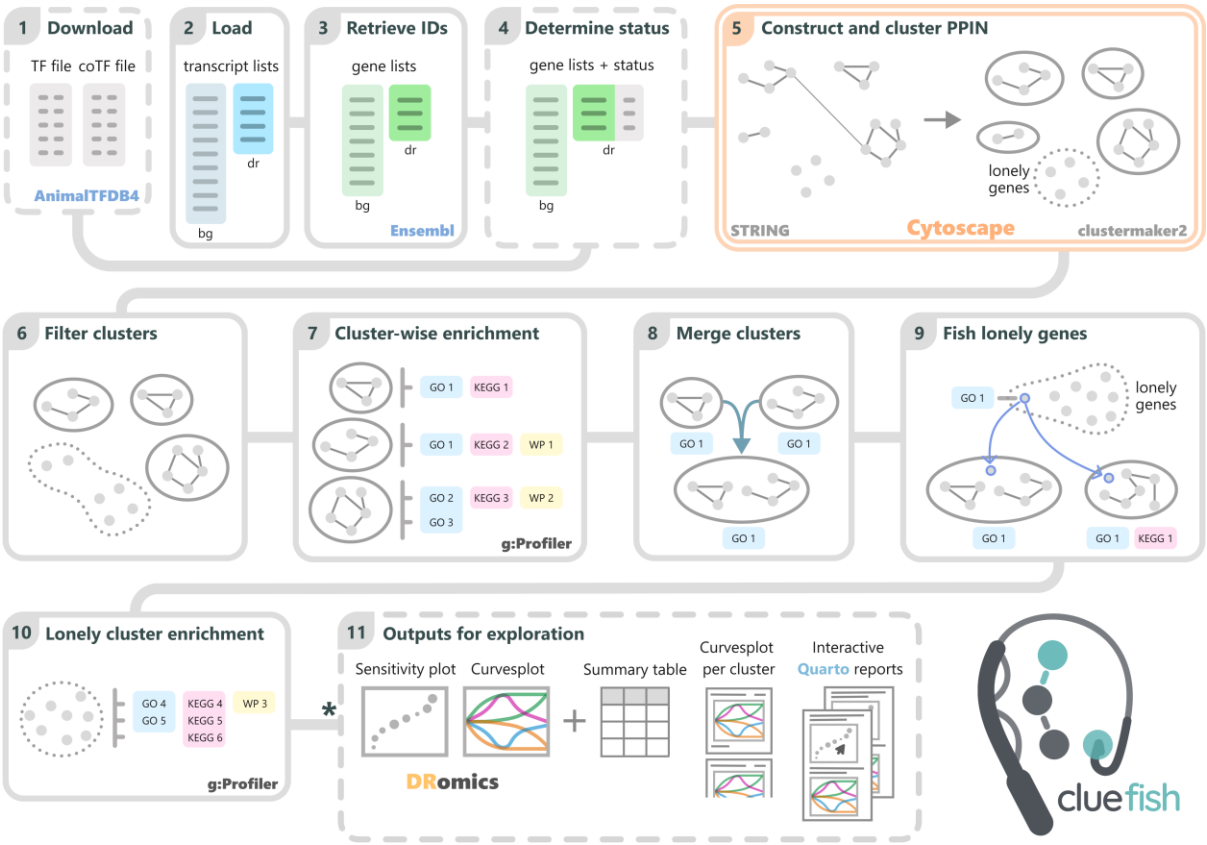

Figure 1. Schematic of Cluefish workflow. Steps performed in R are shown in single-line frames, with dashed frames for optional steps, while the step carried out in Cytoscape is shown within a double lined frame. In step 5, deregulated genes, which code for proteins in STRING, are represented as nodes. Their interactions are illustrated by lines. Clusters are represented by full circles, while isolated genes are categorized as *lonely genes*, forming the *lonely cluster*, represented by a dashed circle. The schematic guides the user through the various steps but does not explicitly depict the output-to-input relationships, as some steps require multiple previous outputs. The full detailed workflow is provided in the vignette (https://ellfran-7.github.io/cluefish/articles/cluefish.html) ***The final step can only be performed if the list of deregulated transcripts loaded in Step 2 is derived from the outputs of DRomics*. Abbreviations: bg: background, dr: deregulated.

The standard workflow

To illustrate Cluefish's capacity to improve biological interpretation of DR transcriptomic data, we compared it with the standard approach using the three same databases (GO:BP, KEGG and WP). This involved applying a simple functional enrichment analysis to the entire deregulated gene list, using the same parameters as those chosen for the cluster-wise enrichment step in Cluefish. Both approaches were compared in terms of number of enriched biological functions and number of deregulated genes involved in them.

Implementation

In our study, we used the zebrafish GRCz11 (GCA_000002035.4) genome assembly from Ensembl (release 111). Details of the versions of tools and databases used are provided in Supplementary Table S2. Cluefish is primarily implemented in the R statistical computing language (48), with a short manual processing step halfway in the Cytoscape application. Dependencies in R, including binary dependencies, are automatically bundled and installed using renv (v. 1.0.7) (49). Cluefish is structured as a research compendium, which is an organized collection of files and directories encompassing all necessary data, code, and documentation required to reproduce the research (50). Git is utilized for version control (v. 2.45.2.windows.1.). The workflow requires two main data inputs: the background and deregulated transcript lists. If these inputs come from DRomics results, the entire Cluefish workflow can be used. However, if the inputs are from a different selection method, Cluefish can be applied up to Step 10. This is because the outputs for visual exploration depend on DRomics figures (e.g., DR curves) and the associated computed metrics. Cluefish is freely available at https://github.com/ellfran-7/cluefish with an associated vignette serving as a guide through the Cluefish workflow with some additional functionalities available, which has not been included in the workflow description for the sake of simplicity (https://ellfran-7.github.io/cluefish/articles/cluefish.html).

RESULTS

Dose-response analysis

A total of 41396 transcripts were detected in at least one sample. Of the 15 initial samples, one of the control samples was removed from the analysis due to its divergence in the principal component analysis (PCA) plot, suggesting it was an outlier (Supplementary Figure S1). Consequently, the analysis proceeded with the remaining 14 samples. As a result, 39,890 transcripts formed the background transcript list for the experiment. From these, DRomics selected 2449 significantly deregulated transcripts, with 2433 having calculated point and interval estimates of the BMD. The majority of these transcripts exhibited a monotonic trend along the dose gradient, with 1132 increasing and 1145 decreasing, corresponding to 46% and 47% of the list, respectively. Biphasic trends were less common, with 64 showing a bell-shaped trend and 92 showing a U-shaped trend, corresponding to 3% and 4% of the list, respectively.

Standard approach

Simple functional enrichment analysis showed that the deregulated gene list, composed of Ensembl stable gene IDs, was significantly enriched in 12 GO:BP, 6 KEGG and 2 WP biological functions. All in all, a total of 320 genes participated in these enrichments, whereas the remaining 2045 genes did not participate in the enrichment of a biological function (Supplementary Table S3).

Cluefish approach

The background and deregulated transcript lists were mapped to 27720 and 2365 unique Ensembl stable gene IDs, respectively. From the AnimalTFDB4 database, 3328 Ensembl genes corresponding to known zebrafish TFs and CoTFs were retrieved. Among the deregulated genes, 286 were identified as coding for these transcriptional regulators. After constructing the PPI network and performing MCL clustering, 204 clusters were formed (Supplementary Table S4). Of these, only 53 clusters passed the size filter, encompassing 399 genes. Cluster-wise ORA identified 41 driver GO:BP, 51 KEGG and 10 WP enriched functions. See Supplementary Table S5 for the initial ORA results prior to filtering, and Supplementary Table S6 for a result summary both before and after filtering. Several clusters were merged based on shared biological function, reducing the total number of clusters to 44. Following *lonely fishing*, the number of genes in the newly expanded clusters increased to 904 (37% of the deregulated gene list). The majority of fished *lonely genes* had a *friendliness* value of one, two, or three, with the highest value reaching eight. The final number of clusters was 45 (44 expanded clusters + *Lonely cluster*) (Supplementary Table S7). The size of the *lonely cluster* was greatly reduced, containing 1461 genes. The functional enrichment of the *lonely cluster* identified 6 driver GO:BP, 4 KEGG and 2 WP biological functions (Supplementary Table S8). These functions did not overlap with any from the cluster-wise enrichment. With this added information, over 1026 deregulated genes enriched at-least one biological function, corresponding to 43% of the deregulated gene list. The sensitivity of the 44 expanded clusters can be illustrated using summarized BMD-1SD values for the transcripts within each cluster (Figure 2). The first quartile represents the value below which 25% of the BMD of the transcripts fall below. According to this summary, the most sensitive clusters #43, #37, #10 and #15 with BMD-1SD first quantile values of approximately 1.95, 2.19, 2.23 and 2.62 µg/L,

respectively. In contrast, clusters #51, #27, #35 and #38 were the least sensitive, with first quartile values of 33.6, 31.7, 31.2 and 29.8 µg/L, respectively (Supplementary Table S9).

Comparison of both approaches

Cluefish identified the same 6 KEGG and 2 WP functions as the standard approach, while also adding 45 KEGG and 8 WP functions to the biological context of the deregulated data (Figure 2). In contrast, only 5 GO:BP terms were shared between both approaches. Cluefish identified 36 unique GO:BP functions, whereas the standard approach identified 7 unique GO:BP functions. These 7 GO:BP terms were higher-level parent terms, representing more general and broader biological concepts. They were redundant with the lower-level child terms found by Cluefish, although they did not share the same GO term names. Cluefish led to the consideration of an additional 733 genes into enriched functions or pathways, accounting for 30% of the deregulated gene list. Of the 320 genes considered by the standard approach, 293 were also considered by Cluefish, leaving 27 genes uniquely considered by the standard approach.

Exploration of the results from Cluefish

In ecotoxicology, the primary focus is on examining the impact of low doses of contaminants in the environment. Prioritizing the exploration of the most sensitive clusters is a reasonable approach in this context. Since the cluster #37 was not characterized, it was excluded from subsequent analysis (Figure 2). We chose to focus on clusters #43, #10, and #15, as they exhibited the lowest 25th BMD quantile values. Cluster #43 comprised 9 distinct transcripts and genes, with some showing increasing and some showing decreasing trends (Figure 2). The majority of the transcripts BMDs ranged between 1.23 and 7.27 µg/L. This cluster solely enriched the "Retinol metabolism" (KEGG:00830) KEGG pathway, with no GO:BP or WP function enrichments. Cluster #10 was composed of 16 distinct transcripts, with the majority showing decreasing trends, also with similar shaped DR curves (Figure 2). Ten BMD values ranged between 0.84 and 5.29 µg/L, while the remaining six ranged between 23.1 and 49.4 µg/L. This cluster enriched sphingolipid metabolism functions in GO:BP and KEGG, specifically "sphingolipid metabolic process" (GO:0006665) and "Sphingolipid metabolism" (KEGG:00600), respectively. Cluster #15 comprised 31 distinct transcripts, mostly showing decreasing trends, also with similar DR curve shapes (Figure 2). In this cluster, 18 BMD values

ranged between 0.49 and 7.37 µg/L, while the remaining 13 BMD values ranged between 19.0 and 57.1 µg/L. This cluster enriched one GO:BP (lipid biosynthetic process - GO:0008610), one KEGG (Steroid biosynthesis - KEGG:00100), and one WP (Cholesterol biosynthesis - WP:WP1387) biological function. None of the 3 clusters contained known transcriptional regulators.

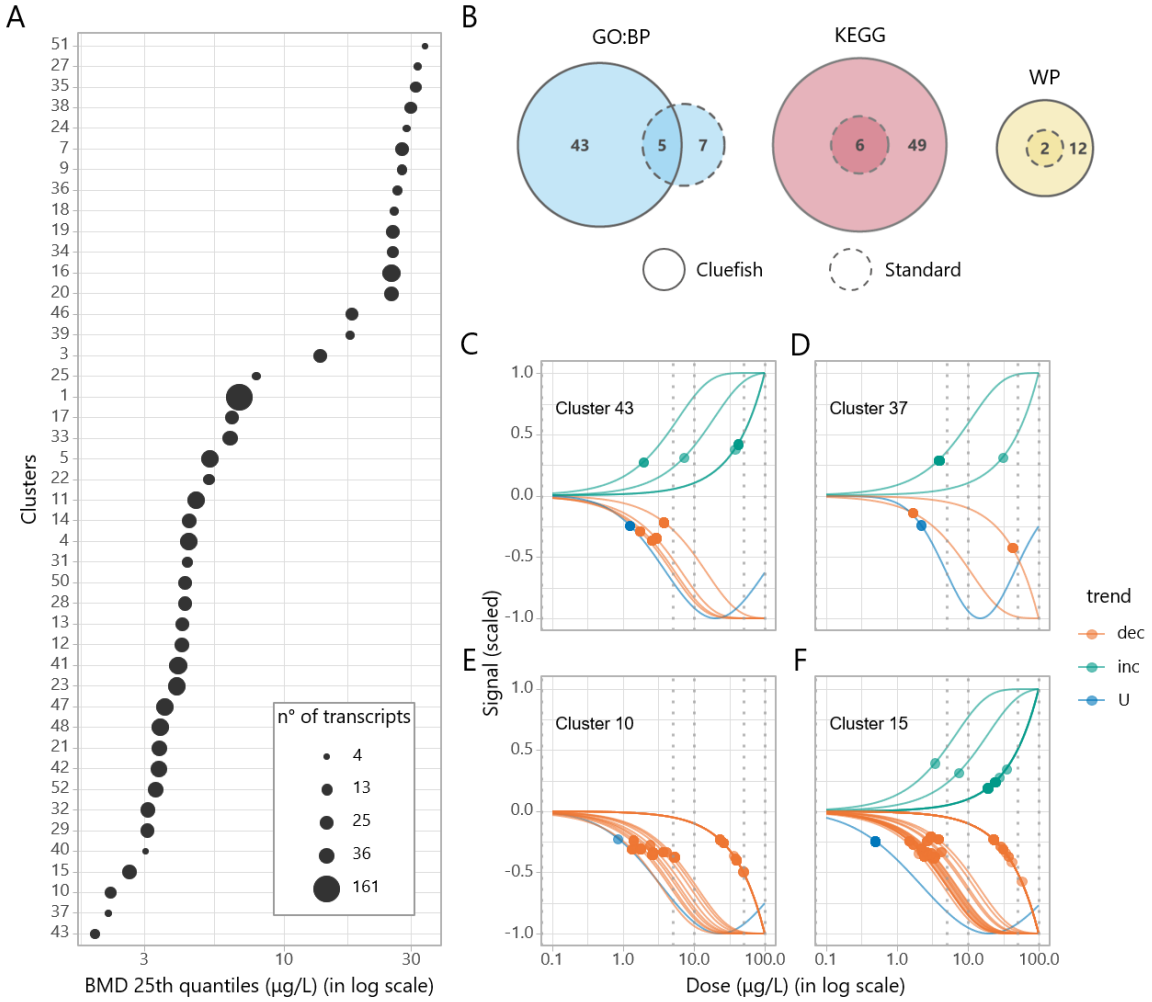

Figure 2. Results of the Cluefish workflow applied to the zebrafish dataset. (A) Summarized BMD 25th quantile values for each cluster, with points ordered from most to least sensitive and sized according to the number of transcripts. (B) Overlap of kept significantly enriched biological functions characterizing deregulated genes, categorized by approach and database. (C, D, E, F) Dose-response fitted curves and computed BMD points for transcripts in (C) cluster 43, (D) cluster 37, (E) cluster 10 and (F) cluster 15, with colors indicating the trend for each curve shape: dec (decreasing), inc (increasing), and U (U-shaped). Grey dashed vertical lines represent the experimental doses.

Investigations towards biological interpretation

Discovering the Retinol metabolism pathway as not only deregulated but also highly sensitive prompted us to investigate the involvement of Hox genes, which are among the primary canonical targets of retinoic acids (51–53). All the Hox genes were located within the *lonely cluster*, where we identified 13 Hox gene transcripts, all showing increasing trends (Figure 3). Eleven of these transcripts had BMD values between 2.51 and 6.80 µg/L and very similar DR curves, while the remaining two had values of 28.9 and 40.6 µg/L (Supplementary Table S10). All of these genes act as transcriptional regulators.

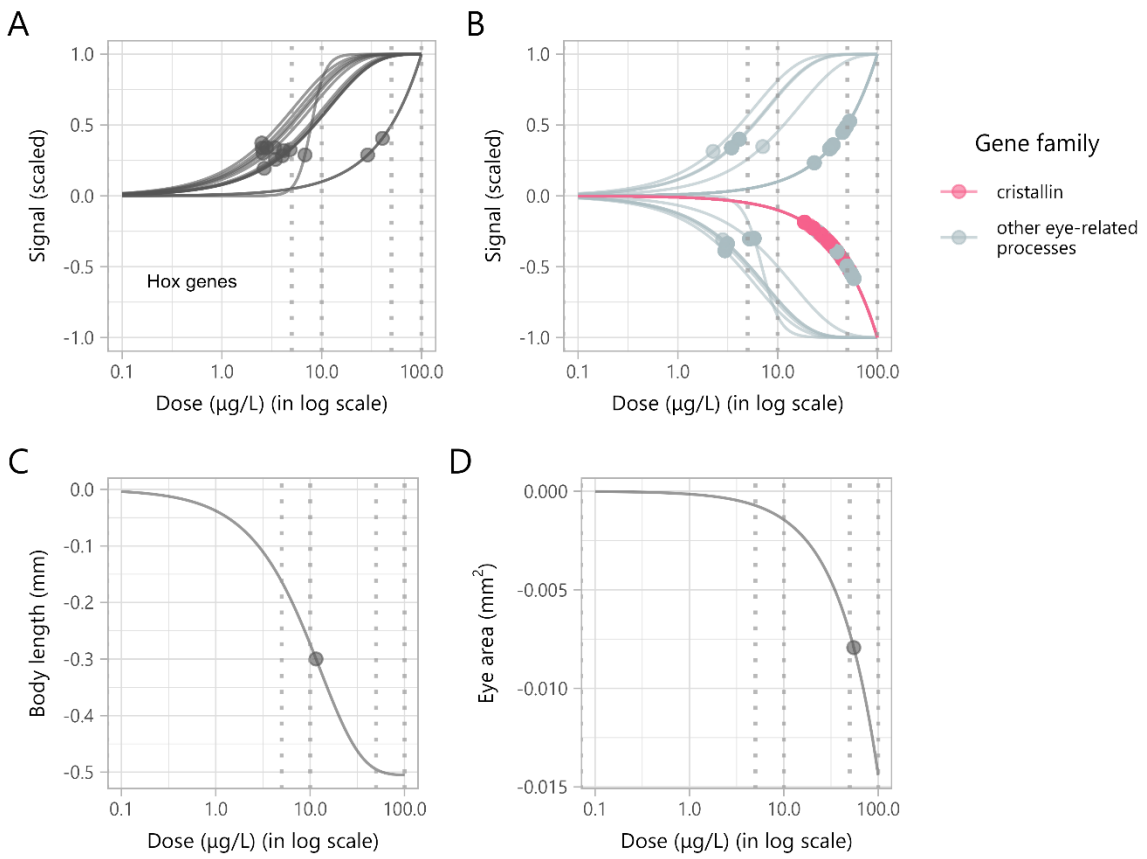

Figure 3. Dose-response curves and computed benchmark dose (BMD) points for deregulated zebrafish transcripts and individual endpoints. (A) Hox gene transcripts. (B) Transcripts associated with eye development and function, with transcripts' genes from the crystallin family highlighted in pink and other eye-related genes in light grey. (C) Body length measurements. (D) Eye surface area measurements.

Further exploration of the *lonely cluster* revealed that the biological functions enriched by the highest number of deregulated genes were "lens development in camera-type eye" (GO:0002088) and "visual perception" (GO:0007601), each enriched by 26 genes. In total, 58

deregulated genes were related to eye development or function (Figure 3). Among these, 27 cristallin gene transcripts were found, all following decreasing trends, with BMD values ranging from 18.6 to 52.6 µg/L (Supplementary Table S11). Our transcriptomic findings led us to investigate individual endpoints. At the individual level, zebrafish embryos exhibited a dose-dependent reduction in body length and eye surface area, with BMD values of 11.5 and 55.1 µg/L, respectively (Figure 3). See Supplementary Table S12 for the raw left eye area and body length measurements.

DISCUSSION

Building the road towards the biological interpretation of transcriptomic data series is a challenging task. To address this, we developed Cluefish, a free and open-source, semi-automated R workflow that offers a comprehensive and untargeted exploration of such data. When used alongside DRomics, Cluefish offers a more complete analysis of dose-response (DR) transcriptomic data. This workflow aims to address the limitations of the standard approach towards biological interpretation by leveraging the power of ORA on pre-clustered networks. These clusters act as anchors for ORA, improving enrichment detection sensitivity. This allows us to identify smaller, more specific biological processes while simultaneously forming exploratory gene groups.

Cluefish integrates valuable information from various biological function/pathway databases (e.g., GO, KEGG, WP), the AnimalTFDB4 database of transcription factors and co-factors for 183 animal genomes, and the STRING database of known and predicted PPIs for 12,535 organisms. Despite drawing from multiple complementary databases, these databases collectively provide a broad coverage of organisms, ensuring that Cluefish is applicable to a wide range of both model and non-model organisms. Innovative features such as cluster merging and the fishing of *lonely genes* based on shared biological context ensure a more comprehensive analysis.

The standard approach identified a limited, overly broad set of enriched functions. This limitation stems from the statistical methods and the coarse granularity of functional gene sets. As a result, both the qualitative and quantitative interpretation of the data suffer. Additionally, focusing solely on enriched functions is risky, as it may overlook valuable insights from deregulated genes that do not contribute to functional enrichment. In fact,

with the standard approach, more than 84% of the deregulated genes did not contribute to the enrichment of any biological function. By analyzing biologically homogeneous gene subsets, Cluefish minimizes the influence of unrelated genes and ensures a more focused analysis. Moreover, by reintegrating *lonely genes* back into these clusters and separately characterizing the list of remaining *lonely genes*, it ensures that no valuable data is overlooked.

The effectiveness of the Cluefish workflow is illustrated in our case study by the identification of specific clusters, such as clusters #43, #10, and #15, which contain transcripts with high sensitivity values. Importantly, the biological functions tied to these clusters, including retinol and lipid metabolism, were missed by the standard approach, as they were not identified as enriched. Cluster #43 includes genes involved in retinol (vitamine A) metabolism. Our findings suggest a mechanistic hypothesis centered on the disruption of retinoic acid (RA) signalling. Disrupted retinol metabolism may lead to altered levels of biologically active retinoids, such as 9-*cis* RA, all-trans RA (ATRA), and 11-*cis* Retinal, causing significant changes in retinoid signalling in zebrafish embryos. This hypothesis is supported by the deregulation of several ATRA target genes abundance. Hox genes, particularly those located in the anterior part of the embryo, are canonical targets of RA signalling (54, 55). Among the 48 Hox genes in the zebrafish genome (56), we detected 47 in our dataset, with 13 upregulated and homogenous in DR curve shape, all located in the anterior positions 1 to 9 of the Hox clusters. Other canonical ATRA target genes are *cyp26a1*, *cyp26b1* and *cyp26c1*. Cyp26 enzymes are key regulators of ATRA levels control in cells, by transforming ATRA into inactive metabolites. Cyp26 are thus responsible for the prevention of activation of RA receptors by RA in specific cells (55). Cyp26 coding transcripts are not identified as deregulated in our dataset. However, *cyp26a1* was selected as significantly deregulated along the dose-gradient but no DR model was successfully fitted to the DR curve of this transcript. Visual inspection of expression data for *cyp26a1* suggests however a bell-shape response of transcript abundance. Alltogether, these responses in abundance of transcripts of canonical ATRA response genes suggest an increased ATRA signalling at the lowest DBP exposure concentrations.

RA plays a central role in early patterning and definition of the spinal cord during zebrafish development (57). Disruption of this pathway has been linked to teratogenic effects,

including spinal deformities such as bent spine and malformed tail, highlighting its crucial role in embryogenesis (58, 59). The dose-dependent reduction in body length observed in zebrafish embryos further strengthens this connection. In addition, the crucial role of RA in ocular development (60) can be linked to the identification of 59 deregulated genes related to eye development. The downregulation of crystallin genes —responsible for maintaining lens transparency and refractive power (61, 62) —points to potential defects in lens function. The dose-dependent reduction in eye surface area aligns with this, suggesting compromised eye development linked to disrupted retinoid signalling.

The sensitive downregulation of genes involved in lipid metabolism and transport (clusters #10 and #15) could also be attributed to the deregulation of retinol metabolism under DBP exposure. 9-*cis* RA serves as a ligand for the retinoid X receptors (RXR), that acts as co-receptors for several nuclear receptors involved in lipid metabolism genes regulation (63). This may otherwise be due to the sensitive downregulation of one of the genes coding for RXRs, *rxraa* (Supplementary Figure S1). While it is currently unclear whether rxr genes expression is under direct regulation of retinoids, both this *rxraa* downregulation at the lowest DBP concentration and retinoid abundance deregulation could directly or indirectly participate to the downregulation of lipid metabolism and transport genes.

This hypothesis of DBP mode of action based on retinoid signalling disruption is consistent with findings from previous studies (59, 64, 65). This has been previously suggested through screening in mouse stem cell lines (65) and observed in transcriptomic data from adult male fish testes (66) as well as in rodent models (67). However, most studies on DBP focus on its interaction with the PPAR signalling pathway (68–70). By applying the Cluefish workflow, our DR transcriptomic data suggests that retinoid signalling and function disruption may be the most sensitive pathway affected by DBP during zebrafish development, potentially leading to significant morphological changes. RA signalling, while still understudied in the context of endocrine disruption, has been increasingly recognized as vulnerable to pollutants in both toxicological and ecotoxicological settings (71, 72). Given the critical role of RA signalling in embryonic development, which is conserved across both vertebrates and invertebrates, the potential impact of DBP on this pathway deserves closer attention (73). These findings illustrate how Cluefish can reveal connections that would have been missed by the standard approach. By investigating clusters, we identified links between them and guided us to delve

deeper into the *lonely cluster*, where valuable insights resided, and target individual endpoints. Cluefish enabled us to formulate hypotheses supported by multiple concordant elements that not only make biological sense but also hold a major scientific interest. These insights primarily arise from exploring three of the four most sensitive clusters. While a more comprehensive analysis could uncover additional roads for biological interpretation, a complete biological interpretation is beyond the scope of this paper.

Clustering is essential in RNA-seq data analysis, grouping genes or samples by shared traits like expression patterns or network connections (74, 75). A widely used approach is weighted correlation network analysis (WGCNA) (76), which identifies co-regulated genes by constructing co-expression networks from their expression profiles across samples (77). The underlying hypothesis behind WGCNA is that coregulated genes are functionally related. However, due to the complexity and multi-faceted interconnectivity of organismal pathways, this can result in clusters of co-regulated but functionally unrelated genes. In our approach, DRomics first identifies a list of deregulated transcripts. Then, using STRING within Cytoscape, we group genes based on known and predicted interactions, directly forming clusters of functionally relevant genes without relying on gene expression patterns. This approach avoids clustering functionally unrelated genes and allows us to characterize the clusters and their genes using the results of DR modelling, within a non-targeted approach.

PPI network construction and clustering are often applied post-enrichment to visualize relationships between gene sets, using tools like GSCluster (78) and NeVOmics (79). Other tools like EnrichmentMap (80) and ClueGO (81), both extensions of Cytoscape, help reduce redundancy in biological functions via PPI networks by accounting for both the overlap and interactions between gene products. Recently, clustering PPI networks before enrichment has gained traction, as seen in tools like PathFindR (82) and ANUBIX (83). This approach improves enrichment sensitivity, uncovering hidden biological mechanisms (84). Though both tools share the pre-clustering concept with Cluefish, their application to lesser-referenced organisms is limited. ANUBIX uses constrained randomization to evaluate subnetwork relevance, making it optimal for highly comprehensive PPI networks. PathFindR, primarily designed for human datasets, is applicable to non-human organisms but enrichment is limited to KEGG, Reactome (85) and MSigDB databases, with the latter two being restricted to very few model organisms. Cluefish, by contrast, seeks to overcome these

limitations by incorporating a functional enrichment tool that supports a broader range of organisms. Nevertheless, challenges still remain for lesser-referenced species.

While Cluefish improves traditional ORA, it has some limitations. Cluefish relies on STRING's PPI data, which can be limiting for transcriptional regulators. For instance, Hox genes, despite being functionally linked, often cluster separately due to a lack of direct interactions (86). This also affects other transcriptional regulators, many of which fall into the *lonely cluster*. Additionally, Cluefish is semi-automated due to a manual Cytoscape step. Full automation could be achieved by integrating this step into R using the RCy3 package (87). However, querying large gene lists and clustering the resulting extensive networks remains challenging.

Despite these limitations, enhancing the methods and databases used in Cluefish could be highly beneficial. Firstly, a more comprehensive set of databases would greatly increase the workflow's versatility and applicability. For instance, expanding PPI data to include lesser-studied organisms would broaden its scope, while extending TFlinks (88) —a database of transcription factor-target gene relationships—to cover a wider range of model organisms would significantly enhance biological interpretation in a mechanistic context. Secondly, a major challenge is the complexity and redundancy of biological functions, particularly in GO terms (89, 90). To reduce redundancy, we applied Alexa et al.'s (2006) method (91), which selects representative terms using the hierarchical structure of the GO tree. This seems to be a more effective approach than semantic similarity methods based solely on gene overlap (91). KEGG pathways also face redundancy issues, and while some methods address this (82, 92), simplifying biological functions remains challenging. Finally, ORA assumes that both genes and pathways are independent, meaning that neither topology nor pathway overlap is considered, which limits its scope. Newer enrichment methods (e.g., topology-based) show promise but are not yet compatible with DR modelling metrics, which Cluefish uses. These methods also require highly complete datasets, limiting their application to well-referenced organisms like humans and mice. To enhance Cluefish's functionality, tools that handle metrics beyond expression values, like BMD and trend metrics, are needed.

To the best of our knowledge, Cluefish is the first workflow to integrate DR modelling with the functional enrichment of pre-clustered networks. This comprehensive approach allows

for the creation of biologically meaningful clusters, characterised by sensitivity and trend derived from DR analysis. With innovative features such as cluster merging and *lonely gene* fishing, Cluefish provides users with a powerful tool for a more thorough interpretation of transcriptomic datasets. By offering clearer insights, it facilitates the generation of biological hypotheses and guiding experimental validation.

To further establish Cluefish's advantages over standard methods and demonstrate its versatility, future work should apply the workflow to additional DR datasets. Testing across different organisms and varying numbers of deregulated transcripts will allow for a deeper evaluation of its robustness and broader applicability.

While Cluefish employs DR modeling to identify deregulated transcripts and visually explore the results, its core functionality extends beyond this specific tool or analysis type. Cluefish can process a background list of transcripts along with a sub-list, typically consisting of deregulated transcripts. These deregulated transcripts can be derived from a wide range of experimental designs, from the simplest to the more complex ones, such as time or spatial data series. Therefore, it would be interesting to explore the application of Cluefish to other types of data series.

DATA AVAILABILITY

The Cluefish project is available in its developmental version on the Github repository (https://github.com/ellfran-7/cluefish). The specific version used to produce the results in this paper is available in the Software Heritage archive at https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/ellfran-7/cluefish, and can be accessed with the following Software Heritage persistent identifier (SWHID): swh:1:dir:e1fc2a9e59d8e2c7fc20cb92ae86b789c07eb303. A detailed step-by-step guide of the workflow is available in the form of a vignette at https://ellfran-7.github.io/cluefish/articles/cluefish.html. The *D.rerio* transcriptomic data underlying this is article is available on NCBI GEO at https://www.ncbi.nlm.nih.gov/geo/, and can be accessed with GSE283957. The *D.rerio* individual data (left eye area and body lengh measurements) are available as supplementary material (Table S12).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

AUTHOR CONTRIBUTIONS

Ellis Franklin: Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing—original draft, Writing—review & editing. Elise Billoir: Conceptualization, Methodology, Supervision, Validation, Writing—review & editing. Philippe Veber: Conceptualization, Methodology, Writing—review & editing. Jérémie Ohanessian: Investigation, Validation, Writing—review & editing. Marie Laure Delignette-Muller: Conceptualization, Methodology, Supervision, Validation, Writing—review & editing. Sophie M. Prud'homme: Conceptualization, Data curation, Investigation, Methodology, Supervision, Validation, Writing—original draft, Writing—review & editing.

ACKNOWLEDGEMENTS

This work was partly done in the "Pôle de compétences en biologie environnementale" and the "Pôle de compétences en chimie analytique environnementale", ANATELo, LIEC laboratory, UMR 7360 CNRS – Université de Lorraine. We would like to express our thanks to Maria Lorena Cordero Maldonado for providing us with the zebrafish embryos, Carole Cossu-Leguille for her assistance with zebrafish larvae sampling, to Bénédicte Sohm for performing RNA extraction and quality checks, and to Claire Caillard for preparing the samples for morphological measurements. Their invaluable contributions significantly enhanced the quality of this research. The raw mRNAseq data processing were performed on the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013).

FUNDING

This work was supported by the Observatoire Terre et Environnement de Lorraine (OTELo) of Université de Lorraine (experimental and sequencing funding). Ellis Franklin was supported by a PhD fellowship from ANR JCJC CHROCO [ANR-21-CE34-0003] and VetAgro Sup. Funding for open access charge: ANR JCJC CHROCO [ANR-21-CE34-0003].

CONFLICT OF INTEREST

Conflict of interest statement. None declared.

REFERENCES

1. Reimand,J., Isserlin,R., Voisin,V., Kucera,M., Tannus-Lopes,C., Rostamianfar,A., Wadi,L., Meyer,M., Wong,J., Xu,C., *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc*, **14**, 482–517.

2. Lander,E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–197.

3. Dong,Z. and Chen,Y. (2013) Transcriptomics: Advances and approaches. *Sci. China Life Sci.*, **56**, 960–967.

4. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X., *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol*, **17**, 13.

5. Das,S., McClain,C.J. and Rai,S.N. (2020) Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy*, **22**, 427.

6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25–29.

7. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27–30.

8. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K., A. Miller,R., Digles,D., Lopes,E.N., Ehrhart,F., *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Research*, **49**, D613–D621.

9. Agrawal,A., Balcı,H., Hanspers,K., Coort,S.L., Martens,M., Slenter,D.N., Ehrhart,F., Digles,D., Waagmeester,A., Wassink,I., *et al.* (2024) WikiPathways 2024: next generation pathway database. *Nucleic Acids Research*, **52**, D679–D689.

10. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

11. Geistlinger,L., Csaba,G., Santarelli,M., Ramos,M., Schiffer,L., Turaga,N., Law,C., Davis,S., Carey,V., Morgan,M., *et al.* (2021) Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*, **22**, 545–556.

12. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol*, **8**, e1002375.

13. Ihnatova,I., Popovici,V. and Budinska,E. (2018) A critical comparison of topology-based pathway analysis methods. *PLOS ONE*, **13**, e0191154.

14. Ma,J., Shojaie,A. and Michailidis,G. (2019) A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics*, **20**, 546.

15. Nguyen,T.-M., Shafi,A., Nguyen,T. and Draghici,S. (2019) Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, **20**, 203.

16. Liu,H., Yuan,M., Mitra,R., Zhou,X., Long,M., Lei,W., Zhou,S., Huang,Y., Hou,F., Eischen,C.M., *et al.* (2022) CTpathway: a CrossTalk-based pathway enrichment analysis method for cancer research. *Genome Medicine*, **14**, 118.

17. Liu,L., Wei,J. and Ruan,J. (2017) Pathway Enrichment Analysis with Networks. *Genes*, **8**, 246.

18. Persson,E., Castresana-Aguirre,M., Buzzao,D., Guala,D. and Sonnhammer,E.L.L. (2021) FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity. *Journal of Molecular Biology*, **433**, 166835.

19. Szklarczyk,D., Kirsch,R., Koutrouli,M., Nastou,K., Mehryary,F., Hachilif,R., Gable,A.L., Fang,T., Doncheva,N.T., Pyysalo,S., *et al.* (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, **51**, D638–D646.

20. Bayerlová,M., Jung,K., Kramer,F., Klemm,F., Bleckmann,A. and Beißbarth,T. (2015) Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics*, **16**, 334.

21. Lovell,D.P. (2016) Chapter 9 - Experimental Design and Statistical Analysis of Threshold Studies. In Nohmi,T., Fukushima,S. (eds), *Thresholds of Genotoxic Carcinogens*. Academic Press, Boston, pp. 129–154.

22. Committee,E.S., Hardy,A., Benford,D., Halldorsson,T., Jeger,M.J., Knutsen,K.H., More,S., Mortensen,A., Naegeli,H., Noteborn,H., *et al.* (2017) Update: use of the benchmark dose approach in risk assessment. *EFSA Journal*, **15**, e04658.

23. Karp,P.D., Midford,P.E., Caspi,R. and Khodursky,A. (2021) Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics*, **22**, 191.

24. Mubeen,S., Tom Kodamullil,A., Hofmann-Apitius,M. and Domingo-Fernández,D. (2022) On the influence of several factors on pathway enrichment analysis. *Briefings in Bioinformatics*, **23**, bbac143.

25. Radke,E.G., Braun,J.M., Meeker,J.D. and Cooper,G.S. (2018) Phthalate exposure and male reproductive outcomes: A systematic review of the human epidemiological evidence. *Environment International*, **121**, 764–793.

26. Baloyi,N.D., Tekere,M., Maphangwa,K.W. and Masindi,V. (2023) Appraisal of the Temporospatial Migration and Potential Ecotoxicity of Phthalic Acid Esters in Municipal Effluents, Rivers and Dam—A Catchment-Wide Assessment. *Water*, **15**, 2061.

27. Kimmel,C.B., Ballard,W.W., Kimmel,S.R., Ullmann,B. and Schilling,T.F. (1995) Stages of embryonic development of the zebrafish. *Developmental Dynamics*, **203**, 253–310.

28. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

29. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, **14**, 417–419.

30. Delignette-Muller,M.L., Siberchicot,A., Larras,F. and Billoir,E. (2023) DRomics, a workflow to exploit dose-response omics data in ecotoxicology. 10.1101/2023.02.09.527852.

31. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

32. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

33. Burnham,K.P. and Anderson,D.R. eds. (2004) Model Selection and Multimodel Inference Springer New York, New York, NY.

34. Larras,F., Billoir,E., Baillard,V., Siberchicot,A., Scholz,S., Wubet,T., Tarkka,M., Schmitt-Jansen,M. and Delignette-Muller,M.-L. (2018) DRomics: A Turnkey Tool to Support the Use of the Dose–Response Framework for Omics Data in Ecological Risk Assessment. *Environ. Sci. Technol.*, **52**, 14461–14468.

35. Stallcup,M.R. and Poulard,C. (2020) Gene-specific actions of transcriptional coregulators facilitate physiological plasticity: Evidence for a physiological coregulator code. *Trends Biochem Sci*, **45**, 497–510.

36. Talukdar,P.D. and Chatterji,U. (2023) Transcriptional co-activators: emerging roles in signaling pathways and potential therapeutic targets for diseases. *Sig Transduct Target Ther*, **8**, 1–41.

37. Reiter,F., Wienerroither,S. and Stark,A. (2017) Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development*, **43**, 73–81.

38. Shen,W.-K., Chen,S.-Y., Gan,Z.-Q., Zhang,Y.-Z., Yue,T., Chen,M.-M., Xue,Y., Hu,H. and Guo,A.-Y. (2022) AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res*, **51**, D39–D45.

39. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

40. Harrison,P.W., Amode,M.R., Austine-Orimoloye,O., Azov,A.G., Barba,M., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2024) Ensembl 2024. *Nucleic Acids Research*, **52**, D891–D899.

41. Kolberg,L., Raudvere,U., Kuzmin,I., Adler,P., Vilo,J. and Peterson,H. (2023) g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, **51**, W207–W212.

42. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498–2504.

43. Doncheva,N.T., Morris,J.H., Gorodkin,J. and Jensen,L.J. (2019) Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.*, **18**, 623–632.

44. Van Dongen,S. (2008) Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.*, **30**, 121–141.

45. Utriainen,M. and Morris,J.H. (2023) clusterMaker2: a major update to clusterMaker, a multi-algorithm clustering app for Cytoscape. *BMC Bioinformatics*, **24**, 134.

46. Kolberg,L., Raudvere,U., Kuzmin,I., Vilo,J. and Peterson,H. (2020) gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res*, **9**, ELIXIR-709.

47. Allaire,J.J., Teague,C., Scheidegger,C., Xie,Y. and Dervieux,C. (2024) Quarto. 10.5281/zenodo.5960048.

48. R Core Team (2024) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

49. Ushey,K. and Wickham,H. (2024) renv: Project Environments.

50. Marwick,B., Boettiger,C. and Mullen,L. (2018) Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician*, **72**, 80–88.

51. Waxman,J.S., Keegan,B.R., Roberts,R.W., Poss,K.D. and Yelon,D. (2008) Hoxb5b Acts Downstream of Retinoic Acid Signaling in the Forelimb Field to Restrict Heart Field Potential in Zebrafish. *Developmental Cell*, **15**, 923–934.

52. Nolte,C., De Kumar,B. and Krumlauf,R. (2019) Hox genes: Downstream "effectors" of retinoic acid signaling in vertebrate embryogenesis. *genesis*, **57**, e23306.

53. Bedois,A.M.H., Parker,H.J. and Krumlauf,R. (2021) Retinoic Acid Signaling in Vertebrate Hindbrain Segmentation: Evolution and Diversification. *Diversity*, **13**, 398.

54. Samarut,E., Fraher,D., Laudet,V. and Gibert,Y. (2015) ZebRA: An overview of retinoic acid signaling during zebrafish development. *Biochim Biophys Acta*, **1849**, 73–83.

55. Samarut,E., Gaudin,C., Hughes,S., Gillet,B., de Bernard,S., Jouve,P.-E., Buffat,L., Allot,A., Lecompte,O., Berekelya,L., *et al.* (2014) Retinoic Acid Receptor Subtype-Specific Transcriptotypes in the Early Zebrafish Embryo. *Mol Endocrinol*, **28**, 260–272.

56. Yamada,K., Maeno,A., Araki,S., Kikuchi,M., Suzuki,M., Ishizaka,M., Satoh,K., Akama,K., Kawabe,Y., Suzuki,K., *et al.* (2021) An atlas of seven zebrafish hox cluster mutants provides insights into sub/neofunctionalization of vertebrate Hox clusters. *Development*, **148**, dev198325.

57. Lara-Ramírez,R., Zieger,E. and Schubert,M. (2013) Retinoic acid signaling in spinal cord development. *Int J Biochem Cell Biol*, **45**, 1302–1313.

58. Dobbs-McAuliffe,B., Zhao,Q. and Linney,E. (2004) Feedback mechanisms regulate retinoic acid production and degradation in the zebrafish embryo. *Mechanisms of Development*, **121**, 339–350.

59. Wang,Y., Chen,J., Du,C., Li,C., Huang,C. and Dong,Q. (2014) Characterization of retinoic acid–induced neurobehavioral effects in developing zebrafish. *Environmental Toxicology and Chemistry*, **33**, 431–437.

60. Thompson,B., Katsanis,N., Apostolopoulos,N., Thompson,D.C., Nebert,D.W. and Vasiliou,V. (2019) Genetics and functions of the retinoic acid pathway, with special emphasis on the eye. *Human Genomics*, **13**, 61.

61. Greiling,T.M.S. and Clark,J.I. (2012) Chapter one - New Insights into the Mechanism of Lens Development Using Zebra Fish. In Jeon,K.W. (ed), *International Review of Cell and Molecular Biology*. Academic Press, Vol. 296, pp. 1–61.

62. Zou,P., Wu,S.-Y., Koteiche,H.A., Mishra,S., Levic,D.S., Knapik,E., Chen,W. and Mchaourab,H.S. (2015) A conserved role of αA-crystallin in the development of the zebrafish embryonic lens. *Experimental Eye Research*, **138**, 104–113.

63. Evans,R.M. and Mangelsdorf,D.J. (2014) Nuclear Receptors, RXR & the Big Bang. *Cell*, **157**, 255.

64. Schlegel,D.K., Ramkumar,S., von Lintig,J. and Neuhauss,S.C. (2021) Disturbed retinoid metabolism upon loss of rlbp1a impairs cone function and leads to subretinal lipid deposits and photoreceptor degeneration in the zebrafish retina. *Elife*, **10**, e71473.

65. Chen,Y. and Reese,D.H. (2013) A Screen for Disruptors of the Retinol (Vitamin A) Signaling Pathway. *Birth Defects Research Part B: Developmental and Reproductive Toxicology*, **98**, 276–282.

66. Prokkola,J.M., Katsiadaki,I., Sebire,M., Elphinstone-Davis,J., Pausio,S., Nikinmaa,M. and Leder,E.H. (2016) Microarray analysis of di-n-butyl phthalate and 17α ethinyl-oestradiol responses in three-spined stickleback testes reveals novel candidate genes for endocrine disruption. *Ecotoxicol Environ Saf*, **124**, 96–104.

67. Alam,M.S., Ohsako,S., Matsuwaki,T., Zhu,X.B., Tsunekawa,N., Kanai,Y., Sone,H., Tohyama,C. and Kurohmaru,M. (2010) Induction of spermatogenic cell apoptosis in prepubertal rat testes irrespective of testicular steroidogenesis: a possible estrogenic effect of di(n-butyl) phthalate. *Reproduction*, **139**, 427–437.

68. Lapinskas,P.J., Brown,S., Leesnitzer,L.M., Blanchard,S., Swanson,C., Cattley,R.C. and Corton,J.C. (2005) Role of PPARα in mediating the effects of phthalates and metabolites in the liver. *Toxicology*, **207**, 149–163.

69. Zhang,W., Li,J., Wei,X., Wang,Q., Yang,J., Hou,H., Du,Z. and Wu,X. (2021) Effects of dibutyl phthalate on lipid metabolism in liver and hepatocytes based on PPARα/SREBP-1c/FAS/GPAT/AMPK signal pathway. *Food and Chemical Toxicology*, **149**, 112029.

70. Fang,H., Wang,H., Zeng,C., Fu,H., Zhao,B., Liu,A. and Yan,J. (2022) A preliminary cumulative risk assessment of Diethylhexyl phthalate and Dibutyl phthalate based on the inhibition of embryonic development via the PPARγ pathway. *Toxicology in Vitro*, **84**, 105430.

71. Martyniuk,C.J., Martínez,R., Navarro-Martín,L., Kamstra,J.H., Schwendt,A., Reynaud,S. and Chalifour,L. (2022) Emerging concepts and opportunities for endocrine disruptor screening of the non-EATS modalities. *Environ Res*, **204**, 111904.

72. Grignard,E., Håkansson,H. and Munn,S. (2020) Regulatory needs and activities to address the retinoid system in the context of endocrine disruption: The European viewpoint. *Reprod Toxicol*, **93**, 250–258.

73. Schubert,M. and Gibert,Y. (2020) Retinoids in Embryonic Development. *Biomolecules*, **10**, 1278.

74. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.-P., Subramanian,A., Ross,K.N., *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

75. Koch,C.M., Chiu,S.F., Akbarpour,M., Bharat,A., Ridge,K.M., Bartom,E.T. and Winter,D.R. (2018) A Beginner's Guide to Analysis of RNA Sequencing Data. *Am J Respir Cell Mol Biol*, **59**, 145–157.

76. Zhang,B. and Horvath,S. (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**.

77. Chung,M., Bruno,V.M., Rasko,D.A., Cuomo,C.A., Muñoz,J.F., Livny,J., Shetty,A.C., Mahurkar,A. and Dunning Hotopp,J.C. (2021) Best practices on the differential expression analysis of multi-species RNA-seq. *Genome Biol*, **22**, 121.

78. Yoon,S., Kim,J., Kim,S.-K., Baik,B., Chi,S.-M., Kim,S.-Y. and Nam,D. (2019) GScluster: network-weighted gene-set clustering analysis. *BMC Genomics*, **20**, 352.

79. Zúñiga-León,E., Carrasco-Navarro,U. and Fierro,F. (2018) NeVOmics: An Enrichment Tool for Gene Ontology and Functional Network Analysis and Visualization of Data from OMICs Technologies. *Genes*, **9**, 569.

80. Isserlin,R., Merico,D., Voisin,V. and Bader,G.D. (2014) Enrichment Map – a Cytoscape app to visualize and explore OMICs pathway enrichment results. *F1000Res*, **3**, 141.

81. Bindea,G., Mlecnik,B., Hackl,H., Charoentong,P., Tosolini,M., Kirilovsky,A., Fridman,W.-H., Pagès,F., Trajanoski,Z. and Galon,J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.

82. Ulgen,E., Ozisik,O. and Sezerman,O.U. (2019) pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. *Front. Genet.*, **10**.

83. Castresana-Aguirre,M. and Sonnhammer,E.L.L. (2020) Pathway-specific model estimation for improved pathway annotation by network crosstalk. *Sci Rep*, **10**, 13585.

84. Chicco,D. and Agapito,G. (2022) Nine quick tips for pathway enrichment analysis. *PLOS Computational Biology*, **18**, e1010348.

85. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R., *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Research*, **48**, D498–D503.

86. Amores,A., Force,A., Yan,Y.-L., Joly,L., Amemiya,C., Fritz,A., Ho,R.K., Langeland,J., Prince,V., Wang,Y.-L., *et al.* (1998) Zebrafish hox Clusters and Vertebrate Genome Evolution. *Science*, **282**, 1711–1714.

87. Gustavsen,J.A., Pai,S., Isserlin,R., Demchak,B. and Pico,A.R. (2019) RCy3: Network biology using Cytoscape from within R. *F1000Res*, **8**, 1774.

88. Liska,O., Bohár,B., Hidas,A., Korcsmáros,T., Papp,B., Fazekas,D. and Ari,E. (2022) TFLink: an integrated gateway to access transcription factor–target gene interactions for multiple species. *Database*, **2022**, baac083.

89. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**, 1–13.

90. Wang,G., Oh,D.-H. and Dassanayake,M. (2020) GOMCL: a toolkit to cluster, evaluate, and extract non-redundant associations of Gene Ontology-based functions. *BMC Bioinformatics*, **21**, 139.

91. Alexa,A., Rahnenführer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

92. Vivar,J.C., Pemu,P., McPherson,R. and Ghosh,S. (2013) Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and "Big Data" Biology. *OMICS*, **17**, 414–422.