Rancho biosciences

# Extraction of High Content Toxicity Biomarker Data

Oleg Stroganov, PhD

Bing Zhou, PhD

March 30, 2025

# Background

## Introduction

The Division of Translational Toxicology (DTT) is an intramural division of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health. DTT uses traditional and cutting-edge approaches to better understand how factors in the environment may impact human health. Such approaches include identifying biomarkers of toxicity in different organs to evaluate the potential impacts of chemical agents of concern.

DTT is increasingly using transcriptional changes to characterize biological and toxicological effects. Current approaches (e.g., gene set enrichment based on annotated pathways and gene ontologies) have limited explanatory power in identifying and characterizing toxicity. As part of this effort, DTT set out to create an alternative collection of gene sets empirically related to toxicological processes in a diagnostic and/or predictive manner. When possible, mechanistic explanation for the response of biomarker to toxicity was curated as well.

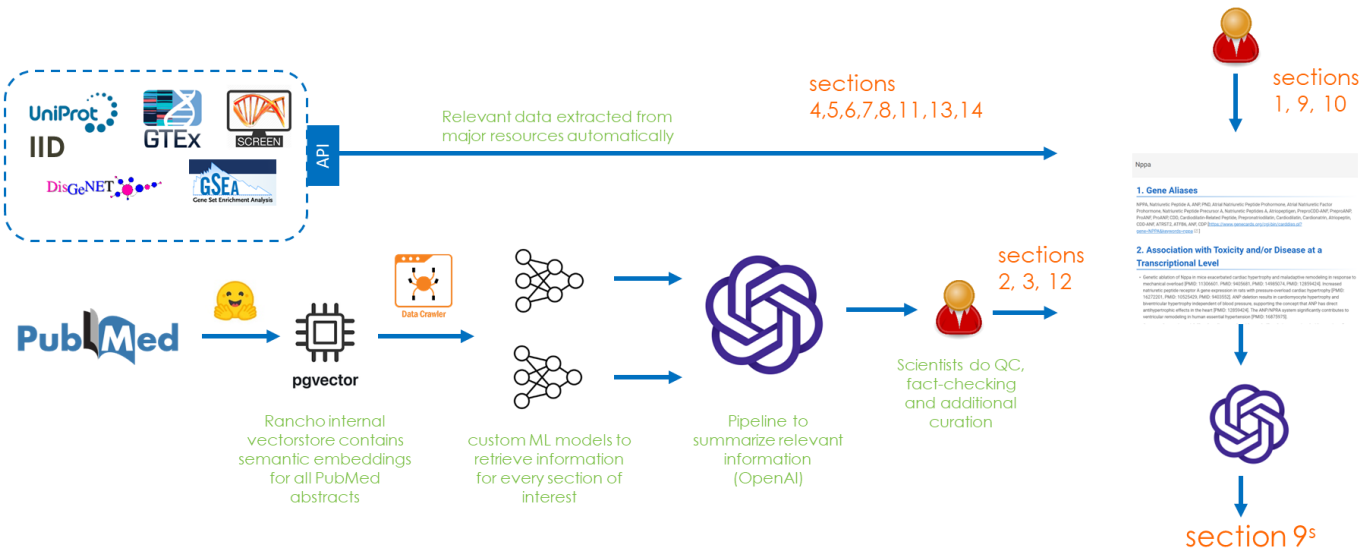
Robust summaries of transcriptional biomarkers of toxicity/disease in different tissues and organs were generated. These summaries will be used to characterize adversity in vivo toxicity screens where gene expression is measured in a dose response format across multiple tissues. There was specific interest in up/down-regulated organ/tissue transcriptional biomarkers of adversity (toxicity or disease) in mammalian models with an emphasis on rats. Specific tissues for which biomarkers were identified and curated included liver, kidney, heart, skeletal muscle, lung, intestine, skin, thyroid, bone marrow, colon, brain. Genes were identified with associated publications cited along with known eliciting agents (i.e., chemical treatments, dietary modifications, genetic models). Once a gene was identified, upstream mechanistic processes were curated to explain the response of each gene. The mechanistic curation ideally was organ/tissue specific but in some cases was also more generalizable. The specific product generated by this effort will be placed in the supplemental material of reports describing the results of DTT toxicogenomic studies, specifically to serve as a justification for using the identified genes to characterize adversity at the level of the transcriptome.

This report describes the High Content Toxicity Database, created by Rancho BioSciences to address DTT needs.

# Methods

## Overall workflow

A combination of manual curation and automatic data extraction was used to prepare reports.




Overall, we used the following process:

1. Sections 2, 3, and 12 were generated automatically and saved in word format into a SharePoint folder
2. Prioritized list of publications was generated for section 9 and saved in Excel format into a SharePoint folder
3. Curators created a gene wiki page and populated information for sections 1 and 10. They used automatically generated reports to add sections 2, 3, and 12, and used prioritized list of publications to add section 9.
4. After manual curation was complete, automatic data crawling for sections 4, 5, 6, 7, 8, 11, 13, and 14 was done. Each section was saved in a markdown format (*.md).
5. Script extracted manually extracted information and merged all sections in a single report. Completeness checks were performed on automatically generated sections. The full report was saved in a markdown format.

6. The full report was reviewed, checked, and uploaded to the wiki.
7. The full report was downloaded and used to generate section 9 summary automatically as described below. The final version of summary report with confidence statements was uploaded to wiki.
8. An additional round of automatic QC and for a subset of genes with manual QC was performed.

## Manual curation

Manual curation of Sections 1, 2, 3, 9, 10, and 12 was performed in accordance with the *Standard Operating Procedures (SOP) for NIEHS Manual Curation of Toxicity Biomarker Data*. The SOP outlines the literature search strategy, tools, and procedures for manual curation—both with and without large language model (LLM) assistance. It also defines the roles and responsibilities of curators and provides quality control guidelines.

## Data crawling

Generated reports contained significant amounts of data that were pulled from structured data sources such as DisGeNET, String DB, UniProt and others using data crawling approaches. Overall, information was extracted from each data source and saved as an intermediate excel file. These files were then converted to a markdown file for each section.

**Section 4: Proteins Known to Interact with Gene Product**

Information about proteins interacting with gene products was extracted from two major sources: Integrated Interactions Database (IID, http://iid.ophid.utoronto.ca/) and String Database (https://string-db.org/).

To get information from IID, a static version of annotated interactions downloaded from http://iid.ophid.utoronto.ca/static/download/human_annotated_PPIs.txt.gz was used. Only interactions where reference type was “experimental” were extracted from PPI. When a gene contained more than 100 interaction partners according to IID, the top 100 interactions based on the number of literature references were preserved.

To get information from String DB, API was used (https://string-db.org/api). Only interactions with a score above 0.7 were retained. Annotations were divided into “experimentally supported” (experimental support score > 0.7), and “supported by text mining” (text mining support score > 0.7 and experimental support score <= 0.7). URL for each String interaction ID was generated automatically.

Results from IID and String were merged and compiled into a list of interactions. Results of text mining were reported only when the total number of experimental interactions was below 100.

**Section 5: Links to Gene Databases**

This section contained references to databases such as:

- GeneCards
- Harmonizome
- NCBI
- Ensembl
- Rat Genome Database
- UniProt
- Wikigenes
- Alphafold
- PDB

Most of the links could be obtained from resources such as UniProt and NCBI; many were automatically generated using gene ids or gene names. For PDB links only proteins that do not contain mutations were used.

**Section 6: GO Terms, MSigDB Signatures, Pathways Containing Gene with Descriptions of Gene Sets**

Pathways, GO terms and MSigDB Signatures were extracted automatically.

Pathways were extracted from Reactome database. Local MySQL dump of Reactome database (https://reactome.org/download-data) was used. Gene/pathway annotations from dababaseobjects table were used, augmented by events summations to retrieve description of the pathway. High-level

pathway branches were excluded in favor of more specific terms, such as "Interleukin-10 signaling" rather than high-hierarchy terms like "Immune System" or "Cytokine Signaling in Immune system."

Within this section, pathway annotations were manually extracted. The sources utilized included the "Pathways & Interactions for Gene" section of the GeneCards database and relevant publications.

MSigDB gene signatures were extracted using web-scraping from https://www.gsea-msigdb.org/gsea/msigdb/human/search.jsp web portal. Only human and rat signatures were used, and only C2: curated gene sets (including CGP: chemical and genetic perturbations and CP: canonical pathways). Extracted MSigDB gene signatures were evaluated by their relevance using semantic similarity towards organs of interest: a *gte-large* semantic embeddings (https://huggingface.co/thenlper/gte-large) for signature descriptions and target organs were calculated, and signatures where cosine similarity of embedding to target organ embedding was above 0.75 were considered relevant. The threshold was calibrated manually by reviewing signatures and their relevance. If more than 100 signatures were extracted, the top 100 signatures by cosine similarity were included in the report.

GO terms associated with a gene were automatically extracted from the Rat Genome Database (https://rest.rgd.mcw.edu/rgdws/annotations/rgdId/) “Gene Ontology Annotation” section. Only “Biological Process” terms were extracted.

**Section 7: Gene Descriptions**

Gene descriptions were taken from the NCBI Gene Summary page and UniProt “Function” section. GeneCards gene descriptions were extracted manually, as GeneCards does not permit automatic scraping of the data.

**Section 8: Cellular Location of Gene Product**

Cellular location of gene product information was taken from HPA (https://www.proteinatlas.org/), from “cellExpression”, “tissueExpression” and “predictedLocation” nodes.

### Section 11: Tissues/Cell Type Where Genes are Overexpressed

Information was extracted from HPA (https://www.proteinatlas.org/) from “rnaExpression” and “cellTypeExpression” nodes.

### Section 13: Chemicals Known to Elicit Transcriptional Response of Biomarker in Tissue of Interest

All chemicals known to elicit transcriptional responses were extracted from the Rat Genome Database (https://rest.rgd.mcw.edu/rgdws/annotations/rgdId/) “Gene-Chemical Interaction Annotations” section. Only chemicals that cause increased or decreased expression, with CTD data source, and at least one referencing publication were kept.

To filter information relevant to the target organ, a LLM classification was used with a *gpt-4o* model. Given the abstract, the model was asked to estimate how confident it was that the article was about the organ of interest. Publications with a confidence score equal or above 1 were kept and up to 5 publications with a confidence score > 0.75 were kept. Compounds increasing expression levels and compounds decreasing expression levels were reported separately.

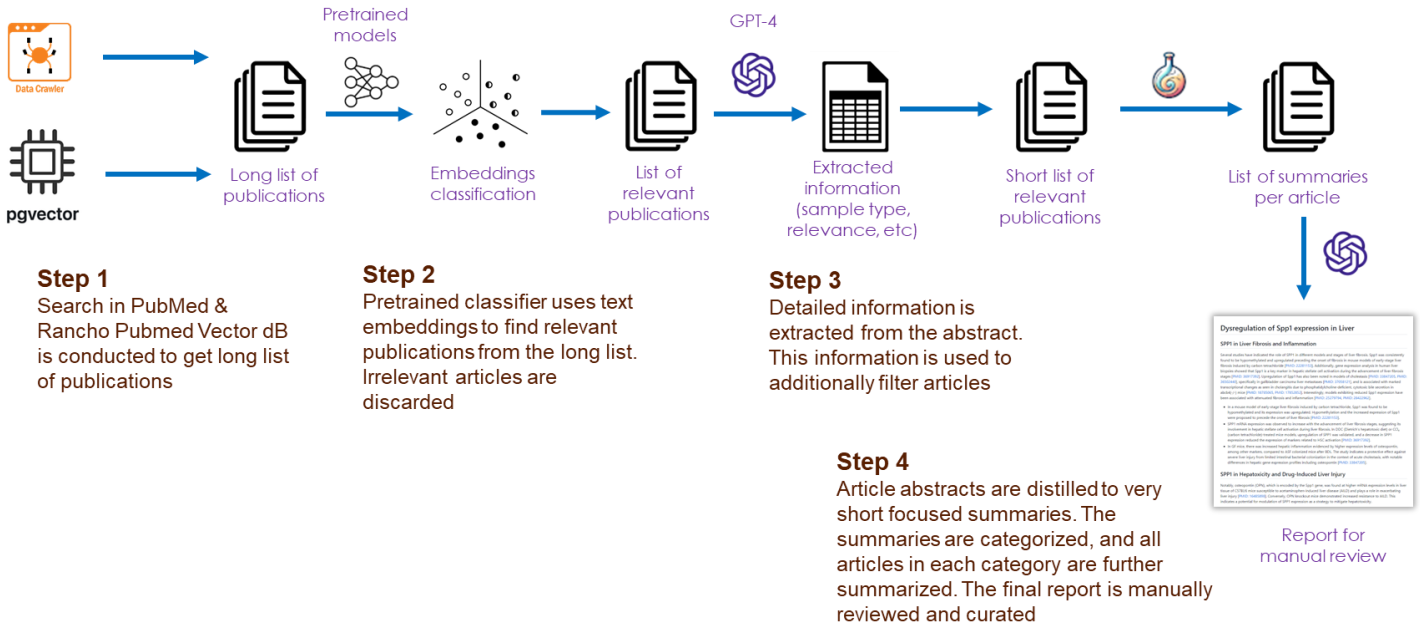
### Section 14: DisGeNET Biomarker Associations to Disease in Organ of Interest

To extract information about Disease – Biomarker associations from DisGeNET, an SQLite dump of DisGeNET database (https://www.disgenet.org/static/disgenet_ap1/files/current/disgenet_2020.db.gz) was used. Only links with “altered expression” association type were retained. The top 5 PubMed references by score were reported for each disease – gene association, and only scores above 0.1 were considered. To evaluate if disease is relevant to the organ of interest, LLM classification was used with a *gpt-4o* model. The LLM was asked to classify the likelihood of disease being associated with organ using “definitely not”, “possibly” and “definitely yes” categories. Only diseases which were classified as “definitely yes” were kept.

# AI-assisted summarization

## Section 2: Association with Toxicity and/or Disease at a Transcriptional Level

To assist with manual curation of section 2 (association with Toxicity and/or Disease at a Transcriptional Level), an automatic data summarization pipeline was built. First, a list of articles was prepared by querying PubMed and the Rancho Semantic Database. Access to the database was provided to curators via a GPT (https://chat.openai.com/g/g-P2BpX2X5G-rancho-biosciences-semantic-search-beta) or used directly through API access. Abstract embeddings were classified into relevant/irrelevant using a pretrained ML model. From the relevant articles information was extracted, and articles were filtered based on pre-defined criteria. Resulting abstracts were distilled using ChatGPT to get the focused summary. All summaries were then further summarized to obtain the final report for manual curation.




*Step 1 – Information search*: to find all articles that could potentially contain information related tosection 2 (association with Toxicity and/or Disease at a Transcriptional Level) a following query was made:

({gene} OR {aliases}) AND ({organ_query}) AND ("mRNA" OR "gene expression")

{gene} and {aliases} are corresponding gene, and its aliases as specified in section 1; {organ_query} – organ and its aliases that helped retrieve relevant information (e.g. liver OR hepatic).

*Step 2 – Filtering relevant information*: text embeddings were calculated for retrieved abstracts using a *gte-large* embedding model. A pretrained classifier (Support Vector Machine) was used to classify embeddings into relevant and irrelevant. To train the classifier, a long list of publications was labeled manually into relevant and irrelevant for 2 genes; after that another list of publications was prepared that contained positive labels for publications which were mentioned in section 2 after initial curation, and negative labels for publications which were retrieved by search query but which were not referenced in manual curation, and the classifier was updated accordingly.

A variety of machine learning models were explored to optimize the classification process, including Support Vector Machine (SVC) with balanced class weight, Random Forest Classifier with balanced class weight, Gradient Boosting Classifier, Neural Network (MLPClassifier) with a maximum iteration of 900, Naive Bayes (GaussianNB), and Logistic Regression with a maximum iteration of 700 and balanced class weight. These models were assessed to determine the most effective approach for distinguishing between relevant and irrelevant publications.

The SVC model emerged as the best performer among the tested models. It was chosen for its ability to handle the imbalanced dataset effectively and the native advantage for classifying vectors and embeddings. The initial evaluation metrics underscored the SVC model's performance, with an Area Under the Curve (AUC) of 0.93 and an accuracy of 0.85. As publications were reviewed by curators, the SCV model was iteratively retrained with the final model AUC of 0.95 and accuracy of 0.89.

*Step 3 – Further filtering*: from a subset of abstracts which were filtered using SVM, further information was extracted using a LLM (model *gpt-4o*), and only articles following the rules were retained.

| Information extracted | Possible values | Rules |
|---|---|---|
| Does abstract have information about gene of interest | True, False | True |

| Information extracted | Possible values | Rules |
|---|---|---|
| Was change of mRNA expression of the gene of interest mentioned in the article? | True, False | True |
| Is study related to tissue or organ of interest? | True, False | True |
| What is the organism that was used for the study? List all organisms using common names (human, rat, mouse) | human, rat, mouse, … | Only studies with *humans*, *rats* or *mouse* are kept |
| What was the sample type on which the study was conducted? Note: if it is unclear whether the experiment was conducted on cell line or patient-derived cells, put "ambiguous". If tissue was extracted from animal, cultured, and then studied, mark it "ex vivo culture" | cell line, organism, patient-derived tissues, ex vivo culture, ambiguous | At least one sample type which is not *cell line*, *ex vivo culture* or *ambiguous* |
| Was gene upregulated, downregulated, no significant difference was reported, or the direction of regulation was not mentioned | upregulated<br>downregulated<br>not significant<br>not mentioned | not "*not significant*" |

*Step 4 – Distilling and summarization*: each selected abstract was distilled using gpt-4o to a 2-sentence summary with the focus on mRNA expression of the gene of interest. A full summary of all information from distilled summaries was then generated. The resulting focused report contains a list of topics (such as e.g., "SPP1 in Liver Fibrosis and Inflammation" or "SPP1 in Hepatoxicity and Drug-Induced Liver Injury" and so on). Each topic has a short summary followed by distilled abstracts with references. To ensure the quality of distilled summary, QC was done using gpt-4o with at most 3 attempts to generate error-free summary that passes the QC.

These short reports were used by curators to prepare section 2 reports for genes of interest.

**Section 3: Summary of Protein Family and Structure**

Protein structure and family information, such as size, length, and synonym, were retrieved using Uniprot, Ensembl, and Gene Cards APIs. Literature pertaining to the proteins was also collected from APIs when available. The collection of publications about protein structure was expanded using PubMed queries and Rancho Semantic Database using keyword queries such as “SPP1 AND (protein structure) AND (human OR mouse OR rat)” for PubMed search, or by embedding the candidate abstract and querying the Rancho Semantic Database using cosine similarity for related publications. The collection of publications was then embedded using the gte-large model and classified using the pretrained classification model for Protein Family and Structure literature. Relevant publications were then passed to a LangChain pipeline to refine and filter publications, as well as surface relevant information about the protein of interest. For section 3 document classification, a multi-layer perceptron (MLP) model performed the best with AUC of 0.9 and an accuracy of 0.92. The model’s performance improved to an AUC of 0.93 and an accuracy of 0.95 with the addition of negative and positive examples provided from curators.

**Section 9: Mechanistic Information**

*Articles prioritization*

To assist curators in identifying relevant literature for Section 9, publications were prioritized using semantic similarity embeddings. Initial article retrieval from PubMed was conducted using the query:

> ({gene}) AND ({organ}) AND (pathogenesis OR toxicology OR pathophysiology OR toxicogenomics OR damage OR toxicity) AND (regulator OR regulation OR pathway OR mechanistic OR mechanism) AND (human OR mouse OR rat)

If this targeted query failed to return relevant articles, a broader query excluding the organ term was used:

> ({gene}) AND (pathogenesis OR toxicology OR pathophysiology OR toxicogenomics OR damage OR toxicity) AND (regulator OR regulation OR pathway OR mechanistic OR mechanism) AND (human OR mouse OR rat)

For instance, each retrieved abstract was segmented into overlapping chunks of 512 tokens (with a 50-token overlap) and embedded using the *gte-large* sentence-transformer model. These embeddings were classified using a pre-trained Support Vector Machine (SVC) to estimate the probability that the abstract contained mechanistic information. Curators were then instructed to review articles in descending order of predicted relevance.

Model selection was based on performance across a training dataset of 7,148 abstracts, of which 250 were manually labeled as relevant. The embeddings for each abstract were generated using the same gte-large model and used as input for several candidate classifiers, including SVC, Random Forest, Gradient Boosting, Neural Networks (MLP), Naive Bayes, and Logistic Regression. Among these, SVC demonstrated the most favorable trade-off between precision and recall for the minority (relevant) class, achieving a cross-validated accuracy of 95.18%, recall of 0.81, and F1 score of 0.48. Although some models exhibited higher recall, such as Logistic Regression and Naive Bayes, they suffered from an unacceptably high false positive rate. Conversely, tree-based methods like Random Forest and Gradient Boosting were overly conservative. Neural networks were moderately effective but less stable and more computationally demanding.

The model training pipeline included vectorization of text chunks, label alignment, and 5-fold cross-validation. In cases where abstracts had conflicting chunk-level labels, only the relevant-labeled chunks were retained. Embedding generation and model training were parallelized to support scalability. All models and intermediate outputs were stored to ensure reproducibility and traceability across experimental runs.

*Generation of summaries*

To generate a summary report for Section 9, a multi-agent system was implemented. The "researcher" agent was responsible for explaining why a gene is dysregulated in diseases and toxicities related to specific tissues. This agent was given explicit instructions on the required structure and tone of the report: it should consist of two paragraphs, avoid excessive introductory language, and focus on functional aspects.

A "QC" agent was tasked with reviewing the reports produced by the researcher, ensuring they adhered to the format and were factually accurate. If a report failed to meet the criteria, the QC agent

provided targeted feedback for revision. A third agent, the "moderator," oversaw the exchange between the researcher and QC agents and decided when the process should be concluded.

Three distinct pipelines were evaluated for report generation:

1. The QC agent reviewed each output without access to gene context. No memory of past interactions was retained.
2. The QC agent had access to the gene report as context, but interactions were still stateless.
3. A conversational thread was maintained, allowing both agents to access prior exchanges and full contextual information during their interaction.

All agents were implemented using gpt-4o. The outputs from the three pipelines were then aggregated using claude-3.7-sonnet, which was tasked with generating a consensus report. In addition, zero-shot reports from claude-3.7-sonnet and the o1 model were generated independently. The final summary report was compiled manually, drawing from the agent-based consensus, sonnet's zero-shot output, and the o1 report.

Once the summary was generated, a gpt-4o LLM model was asked to assign confidence score using the following system prompt:

You are a world class bioinformatician and toxicologist. You worked at NIH to discover molecular pathways that play a role in adverse outcomes, and now you are leading path/tox department of big pharma. One of your responsibilities is to oversee the scientific output of your research team.

Currently they are doing a big project that involves the use of Large Language Models to process scientific literature and extract information related to the response of genes following toxicological challenge. The team wants to evaluate the quality of LLM summaries. With your extensive background in biology, you volunteered to help your scientific team and annotate the statements based on how probable they are. You talked to scientists and came up with the system of Confidence Scores:
- Confidence Score 10 means you are absolutely confident the statement is true.
- Confidence Score 0 means you are uncertain about the statement, and do not know background facts that would be pro or con
- Confidence Score -10 means you know for sure that the statement is false.

You decided that the best way to annotate would be to just add confidence scores to the text in brackets, as if it is an academic citation, e.g. [CS: 9] or [CS: -5]. You want to provide the confidence score for every meaningful statement, but otherwise you do not want to change the format, so as not to cause any problems with downstream processing of the annotation.

Scientific team reached out to you with the first batch of generated summaries. They will provide the summary, and you will be providing the annotated summary in the format laid out above.

Resulting summaries with confidence scores were inspected, and summaries with negative confidence scores were revised. Calibration of confidence statements was assessed as described below.

**Section 12: Role of Gene in Other Tissues**

The reports were assembled similarly to section 2 with the following modifications:

> For article retrieval the following prompt was used:
> ("mRNA" OR "gene expression" OR transcription) AND ({gene} OR {aliases}) AND (mouse OR rat OR human) NOT ({organ_query})

In the prompt we excluded studies where queried organ was mentioned. To filter studies by embedding a different SVC model was used. The model was trained on 56165 abstracts including 612 relevant abstracts and achieved 95.5% accuracy (AUC 0.991). Subsequent filtering of abstracts was done similar to filtering in section 2, except that filters based on organ and expression type were excluded.

## Data Integration

Reports were stored in Rancho's internal wiki server (https://tox-wiki.rbsapp.net/), where a wiki.js system was implemented so that curators could edit information concurrently and leave comments. DTT staff were granted access as well. The wiki implementation supported programmatic access used to merge information generated automatically with manual information and retrieve data from the wiki for QC and export.

## Quality Control

To ensure the reliability and accuracy of data extraction, Rancho has implemented robust quality control measures spanning both manual and automated curation sections. 20 of 113 reports were manually reviewed by a different curator to ensure consistency (Please refer to the NIEHS Manual Curation SOP document to find detailed manual QC specifications). In addition to that, all reports were QCed automatically

| QC check | Relevant sections |
|---|---|
| All sections are present | all |
| References are relevant for corresponding organ | 13, 14 |
| References support statements | 2, 9, 10, 12 |
| Statements are relevant for corresponding section | 2, 9, 10, 12 |

QC checks were conducted using LLM (gpt-4o), and results were reviewed manually.

To ensure the extracted data remained relevant to the specific topics, both AI-assisted and manual reviews were conducted. Data analysis confirmed that both methods maintained high relevance. Additionally, sections curated with LLM assistance demonstrated data relevance comparable to those curated manually.

Furthermore, an LLM-based approach was used to verify the accuracy of extracted data by referencing publication abstracts. This method effectively facilitated precise data accuracy assessments.

### Validation of Manual Curation

To assess the consistency of manual curation, we selected a set of genes for repeated annotation. First, three genes were curated by three PhD-level curators who, while experienced in curation generally, had not previously worked on this specific project. Then, three additional genes were curated by more experienced curators who were already familiar with the project-specific curation framework. Each set of three genes was selected based on the number of associated publications. Specifically, one gene (Tier 1) was chosen from the top third of the list sorted by publication count,

another (Tier 2) from the middle third, and the third (Tier 3) from the bottom third, representing genes with relatively few publications.

Manual curation covered four specific sections: Section 2 (Association with Toxicity and/or Disease at a Transcriptional Level), Section 9 (Mechanistic Information), Section 10 (Upstream Regulators), and Section 12 (Role of Gene in Other Tissues). For Section 2, curators were provided with an LLM-generated summary to assist in their evaluation.

To measure agreement between annotators within each section, we used two complementary methods: a *reference-based approach* and a *statement-based approach*.

In the reference-based approach, annotations were considered identical if they cited the same publications.

In the statement-based approach, annotations were first decomposed into individual factual statements using an LLM (claude-3.6-sonnet). These statements were then evaluated for relevance to the specific section (2, 9, 10, or 12) using GPT-4o.

To assess semantic equivalence between statements, we first computed sentence embeddings using the gte-large model. Pairs of statements with cosine similarity below 0.94 were considered distinct. For pairs above this threshold, semantic equivalence was further evaluated using GPT-4o-mini. Clusters of semantically equivalent and section-relevant statements were then formed.

Annotations were deemed identical if they contained statements that mapped to the same clusters of relevant facts.

Preliminary analysis revealed that curators often produced statements at varying levels of generality. For example, one annotation might list specific chemicals that induce gene expression, while another might refer to a broader chemical class. Under the original statement-based approach, such differences led these statements to be classified as distinct.

To address this, we introduced an additional generalization step: for each factual statement, a more general version was generated using claude-3.6-sonnet. These generalized statements were then processed through the same pipeline—relevance filtering via GPT-4o, semantic similarity using gte-

large, and equivalence confirmation with GPT-4o-mini. This step improved the method's sensitivity to conceptual overlap across differently phrased but substantively similar annotations.

## Calibration of Confidence Scores

To assess how well calibrated LLM-generated confidence scores were, the following procedure was used. First, each summary was split into individual statements corresponding to specific confidence score. This was done with the assistance of claude-3.7-sonnet, with instructions to split summary into individual statements while retaining the context of each statement. 10 statements of each confidence score (2 through 10) were randomly selected and assessed by curator using the following criteria:

**Groundedness**: how well a statement is supported by the evidence in the full report. Curators were asked to ignore truthfulness or falsehood of statement, and just assess groundedness using the metric:

| Score | Meaning | Description |
|---|---|---|
| 3 | grounded | statement is fully supported by the report |
| 2 | speculative | statement is partially supported by the report (only one part of multi-part statement is supported, or statement could be inferred from report) |
| 1 | ungrounded | statement is not supported by the report |

**Truthfulness**: curators were asked to evaluate if statement was correct or incorrect based on general scientific knowledge and available literature. If statement was ungrounded, curators were instructed to do a brief literature search to check the correctness of statement:

| Value | Description |
|---|---|
| correct | Statement is clearly correct |
| incorrect | Statement is clearly incorrect |
| arguable | Correctness of statement is hard to determine, or it depends on unspecified conditions |

**Specificity**: curators were asked to assess how general or how specific the statement was:

| Score | Meaning | Description |
|---|---|---|
| 3 | specific | statement is specific: a specific disease or condition is mentioned, specific pathway is named, specific transcriptional activators are mentioned |
| 2 | moderate | disease classes, general pathways, partner classes are mentioned |
| 1 | general | general statement – general conditions are mentioned, no information about particular or pathways is provided |

## Additional information and decisions

Some of the genes initially planned for curation did not have human orthologs or had many possible orthologs. For these genes, a single human gene was selected for curation by DTT. In particular:

- Human CYP2B6 was curated for rat Cyp2b1
- Human CYP4A22 was curated for rat Cyp4a1
- Human CYP3A5 was curated for rat Cyp3a9
- Human UGT2B7 was curated for rat Ugt2b1

For the genes where curation should be focused on mechanistic biomarkers, very little information was available on specific biomarkers. Such genes were excluded from curation list. Examples are: Klhl15, Shroom2, Fam92a (Androgen receptor), Slco2a1, Cited4, Rgs3 (Estrogen receptor), Mybl1, Atp6v1d, Lama5, Jam3, Enc1, Gria3 (DNA damage).

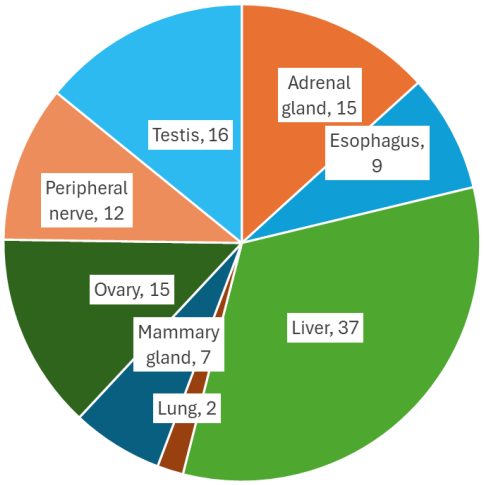
For the majority of the genes (58 out of 79), the full-scale curation as described above was done. However, 21 genes were curated for more than one tissue. In such situations when curating additional tissues, we utilized manually curated data that was prepared for a first tissue. Section 2 (Association with Toxicity and/or Disease at a Transcriptional Level) was curated from scratch, and

Section 12 (Role of Gene in Other Tissues) was assembled based on information from reports on other tissues. All other manually curated sections were taken from the report prepared on the first tissues, and automatically generated sections were updated with the new tissues.

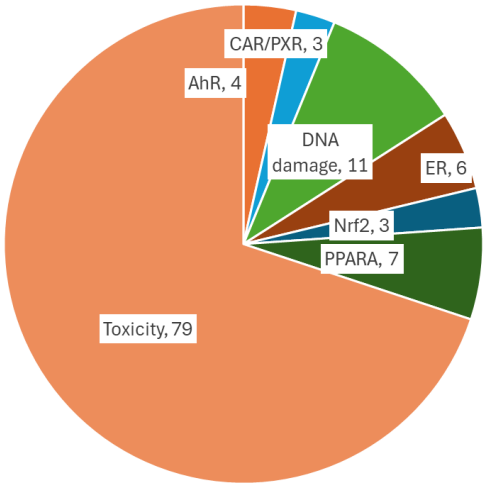
# Results and Discussion

## High Content Toxicity Database

Rancho BioSciences provided a set of 113 curated biomarker summaries to NIEHS. Information on toxicity from 8 tissues was extracted, mostly focused on liver, adrenal gland, testis, ovary, peripheral nerve, and esophagus. Prepared summaries contained information on 79 genes. For most genes, a summary was curated for a single tissue, for 17 genes, two or more summaries were curated.




There are 34 reports focused on Mechanistic Biomarkers (DNA damage, AhR, CAR/PXR, PPARA etc), and 79 reports focused on Toxicity Biomarkers.




Each report consists of the following sections:

1. *Gene Aliases*: a list of common gene aliases and additional information about gene names if required.
2. *Association with Toxicity and/or Disease at a Transcriptional Level*: a list of manuscripts where the gene was identified as a biomarker of toxicity and/or disease at a transcriptional level. Findings from each manuscript are summarized in 1-2 sentences to provide evidence of up- or downregulation within the context of toxicity and/or disease.
3. *Summary of Protein Family and Structure:* a summary of the protein domain structure, protein family, and description of the general role of the protein family. The section also contains information on the general mechanistic role for the gene product in normal function.
4. *Proteins Known to Interact with Gene Product:* a list of proteins known to interact with gene product through direct interaction or through co-expression.

5. *Links to Gene Databases*: links to NCBI gene, Ensembl, UniProt, Rat genome database and other resources.
6. *GO Terms, MSigDB Signatures, Pathways Containing Gene with Descriptions of Gene Sets:* gene sets (i.e., pathways, ontologies, signatures) containing the gene. All gene sets identified should contain textual description.
7. *Gene Descriptions*: the gene function as provided by resources such as Entrez Gene, GeneCards, Uniprot.
8. *Cellular Location of Gene Product*: the cellular location of the gene product (e.g., nuclear, cytosolic, membrane, specific organelle).
9. *Mechanistic Information*: A description of the general mechanistic role of the gene product in normal function and disease processes. The summary section provides a tissue- or organ-specific mechanistic explanations for gene regulation in pathogenesis.
10. *Upstream Regulators*: what is driving the induction and/or reduction of a biomarker upon stress or pathological conditions.
11. *Tissues/Cell Type Where Genes are Overexpressed*: Tissue or cell specific expression as provided by public data sources.
12. *Role of Gene in Other Tissues:* the evidence of up- and downregulation within the context of toxicity and/or diseases other than tissue the report is focused on.
13. *Chemicals Known to Elicit Transcriptional Response of Biomarker in Tissue of Interest*
14. *DisGeNet Biomarker Associations to Disease in Organ of Interest*

The reports were provided to NIEHS through a wiki-based portal, and as word documents. In addition to reports, an excel version of information extracted from public sources was provided as described in Appendix A.

## Validation of Manual Curation

For validation we selected Mgmt (Tier 1), Sclo2a1, Cpt1b (Tier 2) and Phlda3 (Tier 3) genes.

| Gene | Tissue | Tier | Number references (gene only) | Number of references (gene & tissue) | Selected for curation |
|---|---|---|---|---|---|
| Mgmt | Liver | Tier 1 | 5703 | 179 | Project-naïve, Experienced |
| Sclo2a1 | Liver | Tier 2 | 288 | 19 | Project-naive |
| Cpt1b | Liver | Tier 2 | 530 | 132 | Experienced |
| Phlda3 | Liver | Tier 3 | 63 | 14 | Project-naïve, Experienced |

Reference-based recall was calculated as follows:

$$R_{ref} = \frac{N_{ref}}{N_{ref,total}} \times 100\%$$

Where $N_{ref}$ – is number of unique references made by a curator in a specific section of a specific gene, and $N_{ref,total}$ – total number of unique references by all curators. Similarly, fact-based recall was calculated as

$$R_{fact} = \frac{N_{fact}}{N_{fact,total}} \times 100\%$$

Here, $N_{fact}$ is total of unique factual clusters identified by curator, and $N_{fact,total}$ – total number of unique factual clusters identified by all curators.

From the comparison of average reference recall by different sections below it is clear that the recall rate was much higher for section 2 (Association with Toxicity and/or Disease at a Transcriptional Level), which has low topic complexity, and where LLM-generated summaries were used by curators. Sections 9, 10 and 12 showed a lower recall rate. Overall, we did not observe large differences between project-naïve and experienced curators: a slight increase in recall was observed for curation of section 9 (Mechanistic Information), potentially reflecting standardization in information retrieval, while at the same time, we observed a decrease in fact-based recall for section 2.

| Report Section | Topic complexity | LLM-supported | Refence recall, averaged over curators | |
|---|---|---|---|---|
| | | | Project-naïve curators | Experienced curators |
| 2 | Low | yes | 66.2 | 63.0 |
| 9 | High | no | 36.2 | 38.9 |
| 10 | Medium | no | 40.6 | 40.2 |
| 12 | Low | no | 39.8 | 39.2 |

| Report Section | Topic complexity | LLM-supported | Fact-based recall, averaged over curators | |
|---|---|---|---|---|
| | | | Project-naïve curators | Experienced curators |
| 2 | Low | yes | 51.7 | 44.2 |
| 9 | High | no | 34.4 | 40.8 |
| 10 | Medium | no | 37.6 | 37.0 |
| 12 | Low | no | 36.8 | 38.1 |

Comparison of reference recall for genes from different tiers shows the effect of the volume of information: Mgmt (tier 1, a lot of information) has lower recall rate than Phlda3 (tier 3, not much information), probably due to the different justification for inclusion of specific facts into the report.

| Gene tier | Gene | Reference recall, averaged over curators | |
|---|---|---|---|
| | | Project-naïve curators | Experienced curators |
| 1 | Mgmt | 40.8 | 43.3 |
| 2 | Sclo2a1 / Cpt1b | 44.3 | 44.3 |
| 3 | Phlda3 | 52.1 | 48.3 |

Analysis of differences in curated statements showed that curators used different levels of generality. For example, section 2 reports for Phlda3 contained the following statements:

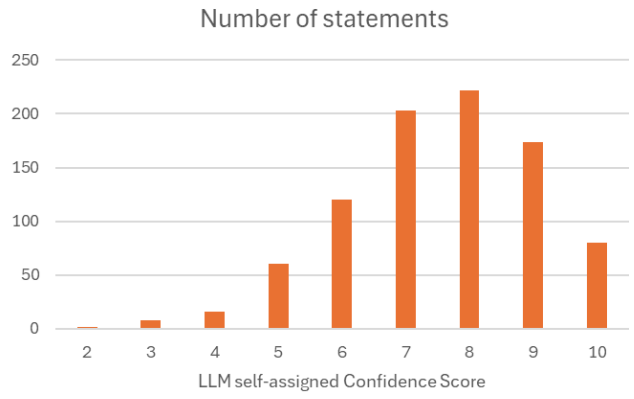
- Phlda3 mRNA expression was analyzed in rats exposed to non-genotoxic hepatocarcinogens (NGTHC)
- Phlda3 gene expression was significantly changed in the liver of F344 rats after treatment with 1,4-dioxane (DO)
- phlda3 gene expression in rat liver tissues showed significant changes when exposed to di(2-ethylhexyl) phthalate (a non-genotoxic hepatocarcinogen).

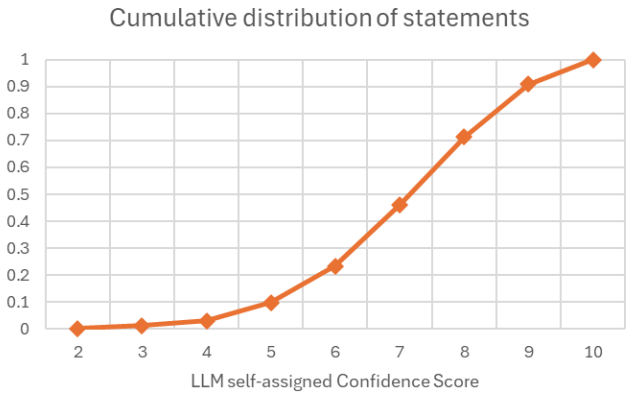
These statements are describing the same fact – increase in Phlda3 gene expression – on the different levels of generality, and for the purpose of information collection could be classified as the same fact. However, these facts were classified as different and thus contributed to decrease in recall. To mitigate this, we generalized facts and conducted recall analysis using these more general statements.

As shown in the figure below, there is a significant increase in recall rate across project-naïve and experienced curators. This increase is more prominent for low-complexity sections (section 2 and 12) and less prominent for high-complexity sections (section 9 and 10).

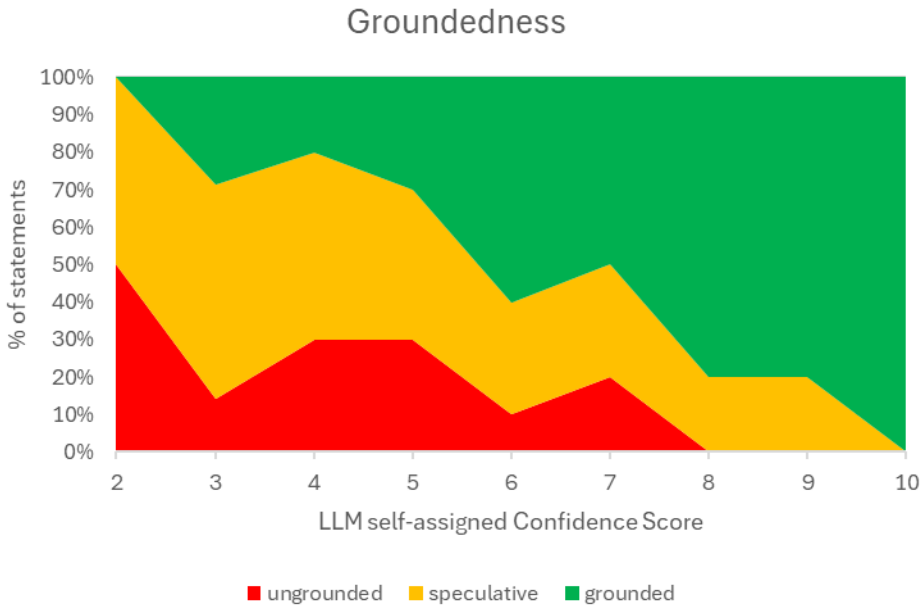

## Calibration of Confidence Scores

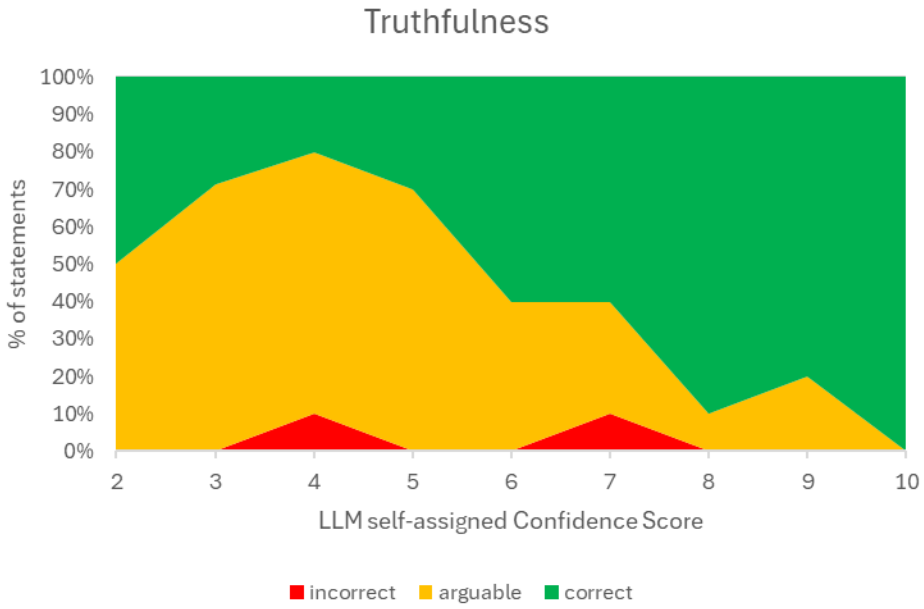
Overall, 885 individual statements with a confidence score (CS) were generated by LLM. The median statement has a confidence score of 8, with a mean value of 7.54. About 10% of statements had confidence scores of 5 and below; 10% of statements had confidence scores of 10.

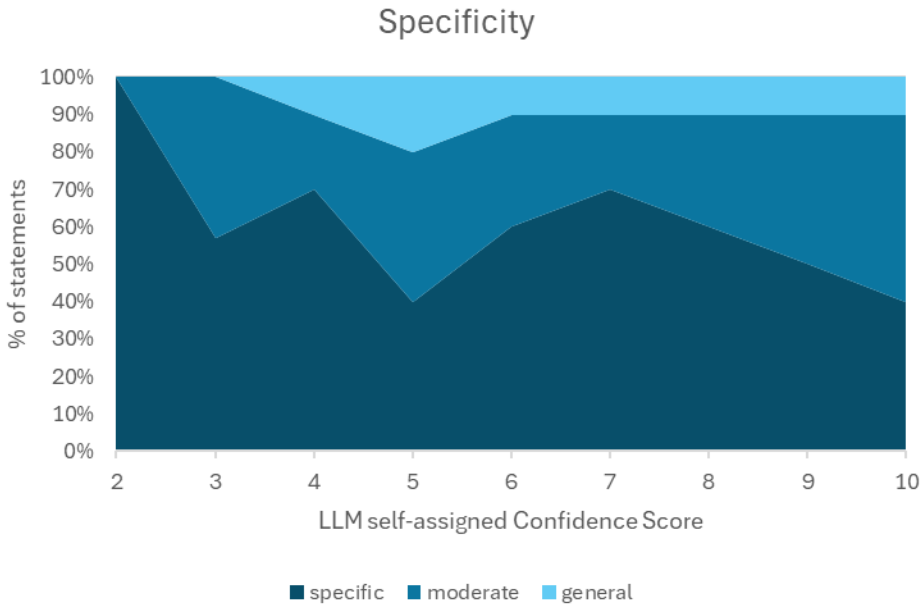


Curators evaluated 79 of the 885 statements based on the matrics of groundedness, truthfulness and specificity.

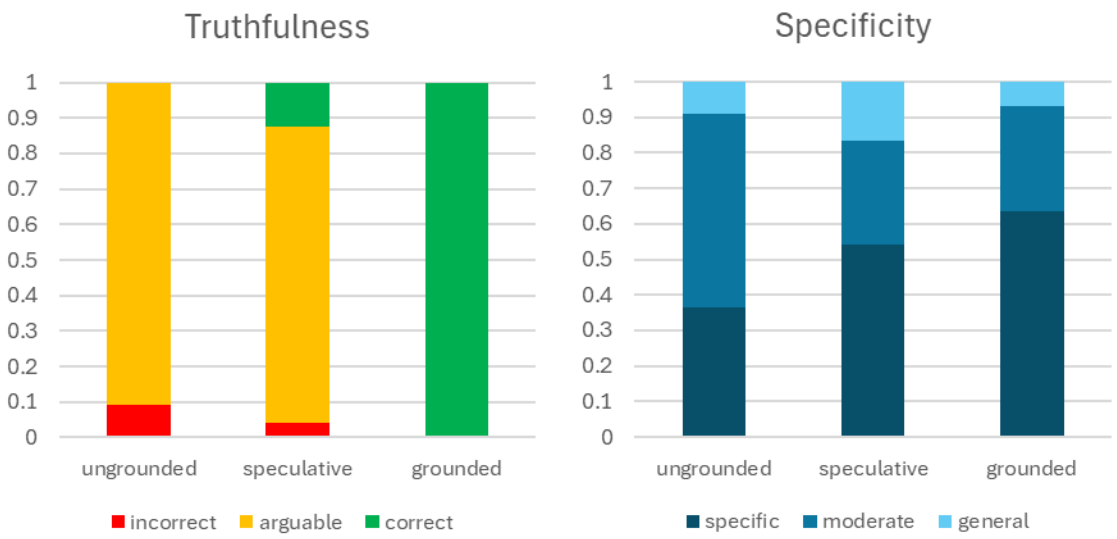





We observed the correlation between groundedness and truthfulness and confidence scores. For confidence scores ≤ 5, 76% of statements were ungrounded or speculative, and 72% were incorrect or arguable, whereas for confidence scores ≥ 8, only 13% statements were speculative (none ungrounded), and 10% of statements were arguable (none incorrect).

Distribution of specificity was less pronounced, with a slight decrease in the fraction of highly specific statements as confidence score is increased.




Analysis of subsets of statements with the same groundedness showed that both truthfulness and specificity are correlated with groundedness: grounded statements are likely to be correct and specific, whereas speculative statements are more likely to be arguable and of moderate specificity.

## Conclusions

This innovative approach significantly advances the field of toxicogenomics by providing a comprehensive, empirically derived dataset of transcriptional biomarkers. These biomarkers are critical in assessing chemical agent impacts on human health. The methodology's efficacy in extracting and summarizing large-scale literature data presents a paradigm shift in toxicological research, offering a scalable and precise tool for biomarker discovery and characterization. This contributes immensely to the predictive and diagnostic capabilities in toxicology, aligning with DTT's mission to incorporate advanced methods for environmental health research.

# Deliverables

| File | Description |
|---|---|
| reports.zip | An archive with gene reports in docx format. |
| extracted_data.zip | An archive with excel files containing automatically extracted information, alongside a description of files. |
| NIEHS report.docx | Final report on the project. |
| NIEHS SOP for Manual Curation_v1.2.pdf | Standard operating procedures for manual curation and QC |

# Appendices

## Appendix A – List and description of files obtained during automatic data extraction

| File | Description |
|---|---|
| gene_list.xlsx | Input list of genes and tissues used for processing |
| gene_ids_curated.xlsx | Curated list of gene ids that contains references to NCBI, Ensembl, Rat Genome DB and UniProt. The list is used for data extraction at subsequent steps. The list was manually curated as automatic extraction of ids is unreliable and has led to several errors (see comments) |
| ppi_iid.xlsx | Information about protein-protein interactions extracted from Integrated Interactions Database (IID, http://iid.ophid.utoronto.ca/) |
| ppi_string.xlsx | Information about protein-protein interactions extracted from string database (https://string-db.org/api) |
| pdb_ids_human.xlsx | Structures from Protein DataBank (PDB) corresponding to human proteins |
| pdb_ids_rat.xlsx | Structures from Protein DataBank (PDB) corresponding to rat proteins |
| pdb_ids_mouse.xlsx | Structures from Protein DataBank (PDB) corresponding to mouse proteins |
| msigdb.xlsx | Information about gene signatures from MSigDB |
| msigdb_scored.xlsx | Information about gene signatures where each signature is scored for its relevance for corresponding tissue |
| rgdb_go.xlsx | GO biological process terms associated with genes and extracted from Rat Genome Database (https://rest.rgd.mcw.edu/rgdws/annotations/rgdId/) |
| uniprot_description.xlsx | Gene function descriptions extracted from UniProt |

| File | Description |
|---|---|
| entrez_description.xlsx | Gene descriptions extracted from Entrez |
| genecards_description.xlsx | Gene descriptions manually extracted from GeneCards |
| hpa_subcellular.xlsx | Information about subcellular localization extracted from HPA (https://www.proteinatlas.org/) |
| hpa_tissue_celltype.xlsx | Information about tissue and cell type where gene is expressed, according to HPA (https://www.proteinatlas.org/) |
| rgdb_compounds.xlsx | Compounds known to elicit transcriptional response of genes, extracted from Rat Genome Database. |
| rgdb_compounds_pubmeds.xlsx | List of combination publications referencing compounds eliciting transcriptional response of genes. Relevant organs are listed for each publication. This list is used to score the relevance compound/gene relationship in the context of tissue. |
| rgdb_pubmed_organs_annotated.tsv | Scores that show how relevant a particular publication is for the organ of interest. |
| disgennet_associations.xlsx | List of gene-disease associations extracted from DisGeNet |
| disgennet_diseases.xlsx | List of disease-organ pairs that need to be annotated |
| disgennet_diseases_annotated.tsv | Annotated disease-organ pairs. Each pair relevance is binned into “definitely yes”, “definitely no” and “possible” |
| reactome_summaries.xlsx | Annotation of low-level Reactome pathways |