Extraction of High Content Toxicity Biomarker Data

Oleg Stroganov, PhD

Bing Zhou, PhD

Tyler Myers, PhD

March 28, 2024

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.

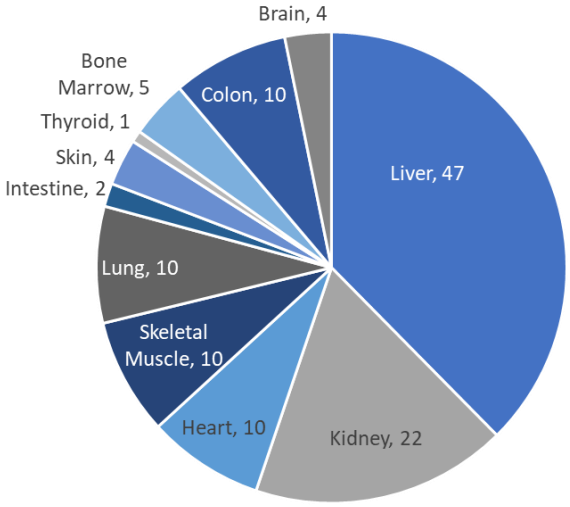
Robust summaries of transcriptional biomarkers of toxicity/disease in different tissues and organs were generated. These summaries will be used to characterize adversity in in vivo toxicity screens where gene expression is measured in a dose response format across multiple tissues. There was specific 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 toxicogenomics 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.

Results

High Content Toxicity Database

Rancho BioSciences provided a set of 125 curated biomarker summaries to NIEHS. Information on toxicity from 11 tissues was extracted, mostly focused on liver, kidney, heart, skeletal muscle, lung, and colon. Prepared summaries contained information on 106 genes. For most genes, a summary was curated for a single tissue, for 18 genes, two or more summaries were curated.



Each report consisted of the following sections:

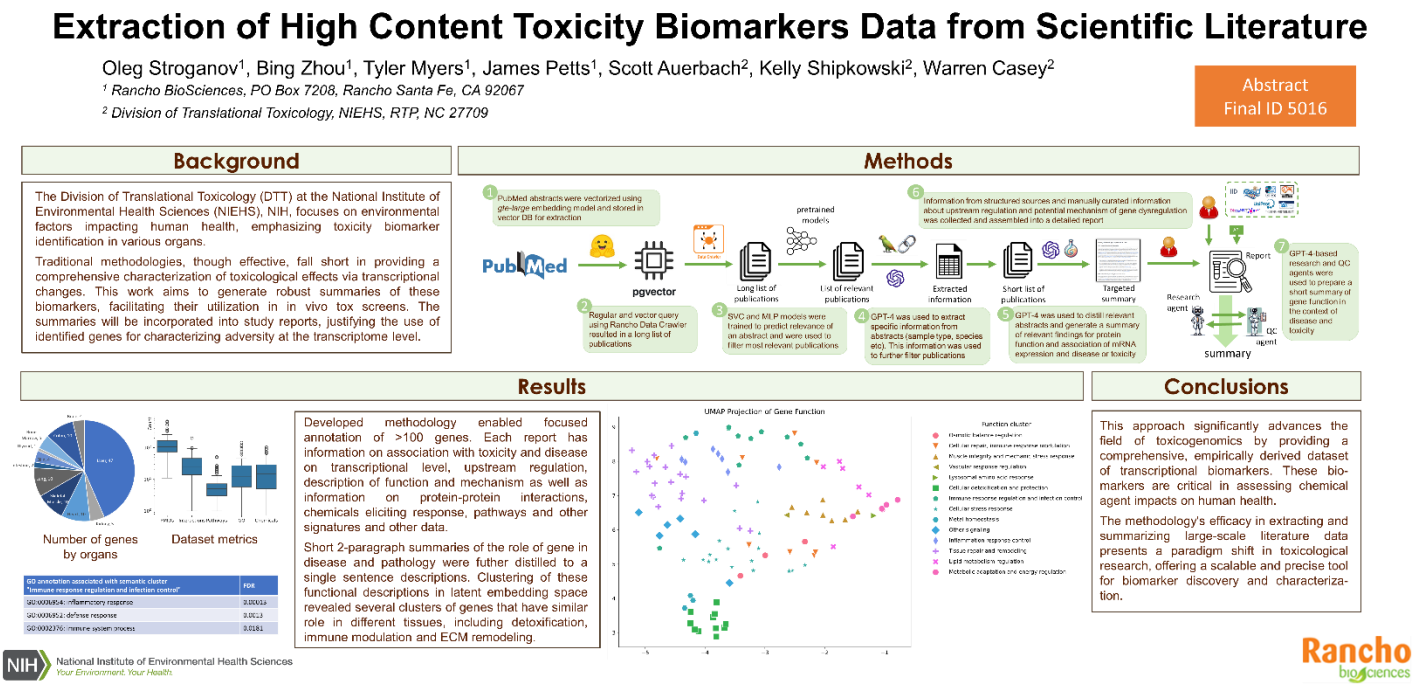
1. *Gene Aliases*: a list of common gene aliases and additional information about gene names if required.
2. *Association with Toxicity and/or Disease at a Transcriptional Level*: a list of manuscripts where the gene was identified as a biomarker of toxicity and/or disease at a transcriptional level. Findings from each manuscript are summarized in 1-2 sentences to provide the evidence of up- and 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 the 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- and organ-specific mechanistic explanations for gene regulation in pathogenesis.
10. *Upstream Regulators*: what is driving the induction and/or reduction upon stress or pathological conditions.
11. *Tissues/Cell Type Where Genes are Overexpressed*: Tissue or cell specific expression as provided by public resources.
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-, html- and markdown documents. In addition to reports, an excel version of information extracted from public sources was provided as described in Appendix B.

Methods

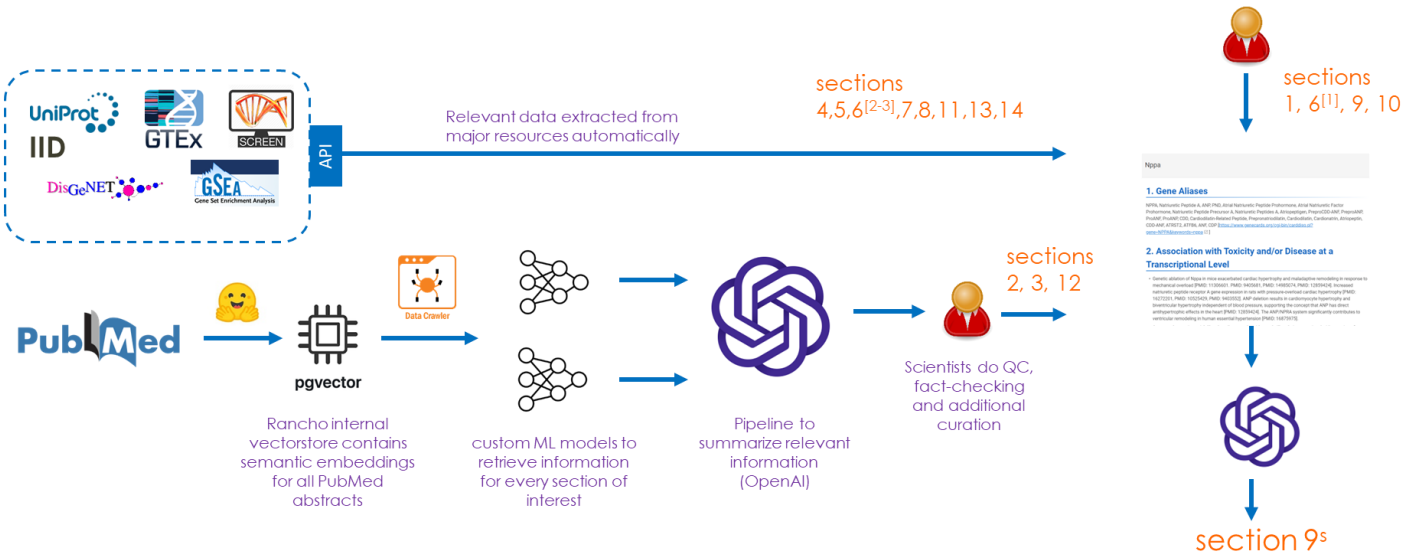
The data extraction workflow was presented at the poster session of 2024 Society of Toxicology meeting at Salt Lake City (Abstract ID 5016).



The poster outlines data extraction methodology, an overview of dataset metrics, mapping to GO pathways, and UMAP representation of gene function description. After the conference additional 17 gene/organ pairs were curated, and updated information is provided in the report below.

Overall workflow

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



Overall, we followed the following process:

1. Sections 2, 3, and 12 were generated automatically and saved in word format into a SharePoint folder
2. Curators created a gene wiki page and populated information for sections 1, 6, 9, and 10. They used automatically generated reports to add sections 2, 3, and 12.
3. 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).
4. 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.
5. The full report was reviewed, checked, and uploaded to the wiki.
6. The full report was downloaded and used to generate the section 9 summary automatically as described above.
7. Different versions of section 9 summaries were reviewed, and the final versions were uploaded to the wiki manually.
8. Additional round of automatic QC and for a subset of genes with manual QC was performed.

Manual curation

Literature Search Strategies for Manual Curation

In the ongoing project, DTT was interested in the creation of comprehensive summaries for a specified set of biomarkers. The primary objective was to enhance their applicability in *in vivo* toxicological screening processes.

The data curation process led by Rancho encompassed a meticulous review of literature sources and integration of data from renowned gene annotation platforms such as Uniprot, MSigDB, RGD, HPA, WikiGenes, DisGeNet, and others. The focal points of these curated reports span diverse dimensions, including insights into gene expression alterations and their correlation with toxicity and disease states. Additionally, an exploration of protein structure in relation to its functional aspects, identification of proteins influencing or being regulated by the gene of interest, and an overview of associated pathways and signatures are provided. Furthermore, these reports shed light on the intricate mechanisms governing gene regulation under pathological conditions. Detailed information on the cellular localization and expression patterns of the gene across various tissues is included. The reports go a step further by incorporating data on chemicals known to induce transcriptional responses of biomarkers within the specific tissue of interest. This approach ensures a thorough and nuanced analysis and understanding.

Complementing the structured extraction from databases, Rancho has developed an in-house Data Crawler App, a web-based tool designed to facilitate targeted data queries from public articles. This innovative tool expedites research endeavors by extracting metadata from articles matching specified search criteria. Leveraging advanced queries, incorporating Boolean operators, and utilizing an AI-assisted search builder powered by OpenAI, the Rancho Data Crawler ensures efficient exploration of numerous publicly available biorepositories, including PubMed (https://pubmed.ncbi.nlm.nih.gov/), a repository of life science literature hosted by NCBI. This sophisticated approach allows for identification of research articles that align with complex requirements, thereby streamlining the data acquisition process for enhanced research outcomes.

For instance, to identify publications pertinent to gene transcription studies in human or rodent models, particularly those associating gene expression levels with diseases or toxicity conditions specific to liver tissues, a refined and targeted query approach was employed. This involved combining various gene synonyms with specific constraints related to tissue type, gene expression, and organisms. For example, the query may encompass a string like:

> ("Trib3" OR "SKIP3" OR "TRB-3" OR "NIPK" OR "SINK" OR "Tribbles Pseudokinase 3") AND ("liver" OR "hepatic") AND ("mRNA" OR "gene expression") AND ("human" OR "mouse" OR "rat").

Additionally, our curators meticulously screened supplementary publications referenced by other reputable gene/protein annotation databases, ensuring a comprehensive coverage of relevant literature for each topic. Examples of such databases include:

- Publications list from UniProt Databases
  (e.g., https://www.uniprot.org/uniprotkb/Q96RU7/publications);
- References from the Rat Genome Databases (RGD)
  (e.g., https://rgd.mcw.edu/rgdweb/report/gene/main.html?id=1345491#pubMedReferences);
- References from the wikigenes webpages
  (e.g., https://www.wikigenes.org/e/gene/e/57761.html).

This multi-pronged approach ensured the inclusion of high-quality and diverse sources, contributing to a thorough and exhaustive compilation of literature relevant to the specified criteria.

Manual Curation Specifications

Ph.D. scientists conducted manual data extraction, aligning their focus with specific topics outlined by the priorities established by DTT. The curation process unfolded within a dedicated wiki system hosted on Rancho's internal server, utilizing a standardized curation template. Within this framework, curation scientists could proficiently tag the various stages of the manual curation process, encompassing primary curation, quality control (QC) review, and subsequent revision processes for gene reports.

Each section of the curation underwent a targeted and strategic approach, guided by pre-established rules that were finely tuned to align with DTT's specific interests. Ensuring the reliability and accuracy

of data extraction, Rancho implemented robust quality control measures spanning both manual and automated curation sections. To uphold the highest standards, 25% of the gene reports underwent an additional layer of review by an independent curator, thereby enhancing accuracy, consistency, and overall data integrity. This comprehensive approach guarantees the reliability of the curated information and attests to our commitment to delivering top-tier data curation services.

Section 1: Gene Aliases

For the extraction of gene aliases, our primary analysis identified the GeneCards database as the most comprehensive source for gene synonyms. Accordingly, Rancho utilized this database as the primary resource for extracting gene aliases.

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

This section focused on elucidating gene expression changes related to pathological conditions or toxicity in the tissue of interest. Transcriptomic changes, representing alterations in gene expression levels at RNA/mRNA levels assessed through techniques like real-time PCR, RNA-seq, and DNA microarray, were prioritized. DTT priorities guided the focus on findings from human or animal model studies, with the exclusion of cell line-based studies. Evidence from in vitro models was considered secondary unless utilizing primary cells from rodent models. Notably, mutagenicity was not regarded as an adverse effect unless the study concentrated on the model of action (MOA) or association mechanisms. Studies solely demonstrating protein expression changes through protein assays such as ELISA, Western blot, mass spectrometry (MS), or SDS-PAGE fell outside the scope of this section.

To assist with annotating section 2, AI-prepared reports were used. These reports were generated according to the curation rules as described below and were used by curators.

Section 3: Summary of Protein Family and Structure

This section provided general information on proteins, encompassing details such as size, molecular mass, domains, and families. The extraction of evidence regarding protein structure in relation to protein functions was derived from pertinent publications. General information about the protein

families and structures was extracted via Uniprot, Protein Atlas, and Gene Cards. Similar to the AI methods for Section 2, machine learning models were trained on embedding representations of relevant publications containing family and structural information about any given protein. The findings from the relevant publications were then filtered for relevancy by OpenAI's GPT-4 and summarized. Models were retrained as curators accumulated more relevant publications and augmented with reinforcement learning from feedback from manual curation.

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

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. Emphasis was placed on Reactome database pathways within GeneCards, with a focus on granular pathway terms. 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." Up to 20 granular-level pathway terms for each gene were included. When Reactome annotations were unavailable in GeneCards, pathways were sourced from publications, and their descriptions were obtained from diverse sources, including WikiPathways (https://www.wikipathways.org/), KEGG (https://www.genome.jp/kegg/pathway.html), and Review papers.

Section 9: Mechanistic Information

This section aimed to uncover mechanistic insights into how the gene of interest experiences dysregulation under various diseased conditions. It explored the intricate relationship between transcriptomic regulation of the gene and toxicology, examining gene signaling, cellular processes, developmental processes, or bioactivities that were affected by the gene's dysregulation in toxic conditions.

Section 10: Upstream Regulators

Focused on extracting information about gene expression upstream regulators, this section explored stimuli or factors, such as coactivators, growth factors, cytokines, hormones, transcriptional factors, and pathway upstream components that trigger or diminish the pathway or signal transduction containing the gene of interest. It delves into how these factors modulate gene transcriptional changes in response to various diseases, pathological conditions, or toxic environments.

Section 12: Role of Gene in Other Tissues

This section involved the extraction of insights regarding the gene of interest and its association with diseases or pathological conditions in tissues beyond the primary focus. Associations may be based on gene or protein dysregulations. Questions addressed include whether the gene serves as a biomarker for disease detection, progression, or prognosis in other tissues, if its dysregulation in other tissues relates to toxic response processes, and whether gene mutations or variations are linked to specific diseases.

Rancho utilized an internally developed tool, the "PubMed Distiller", to streamline data extraction from publication abstracts. Curators utilize this tool by specifying the study focus and incorporating key components like gene names/synonyms, the tissue of interest, and relevant disease conditions. Additional rules or instructions for abstract data extraction can also be specified by the curator. A sample specification for data extraction in Section 2, involving a gene paired with liver tissue, is provided as a reference.

- **Focus:**
  ("Cxcl1" OR “CXCL-1” OR “SCYB1” OR “NAP-3” OR “GRO1”) AND ("mRNA" OR "gene expression") AND ("liver" OR "hepatic") AND ("human" OR "mouse" OR "rat")

- **Additional instructions:**
  Summarize the changes in Cxcl1 gene expression associated with liver diseases, hepatic dysfunctions, or other toxic conditions (This entails investigating associations with specific conditions such as hepatocellular carcinoma, hepatic fibrosis, steatohepatitis, alcoholic hepatitis, hepatitis, cirrhosis, non-alcoholic steatohepatitis, fatty liver disease, Wilson’s disease, hemochromatosis, acute liver failure, ischemia reperfusion injury, and primary biliary cirrhosis). The emphasis is on illuminating the impacted biological processes or signaling

pathways that contribute to the modified expression of Cxcl1 and elucidating their roles in the genesis of pathological conditions.

In instances where the manual curation yielded no relevant information on the above-mentioned topics, a standardized statement was incorporated: “No pertinent information pertaining to this subject matter was identified in the existing body of literature.” This ensured transparency regarding the absence of data on the specified themes within the curated literature.

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 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

GO terms and MSigDB Signatures were extracted automatically. 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-4-turbo* 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.

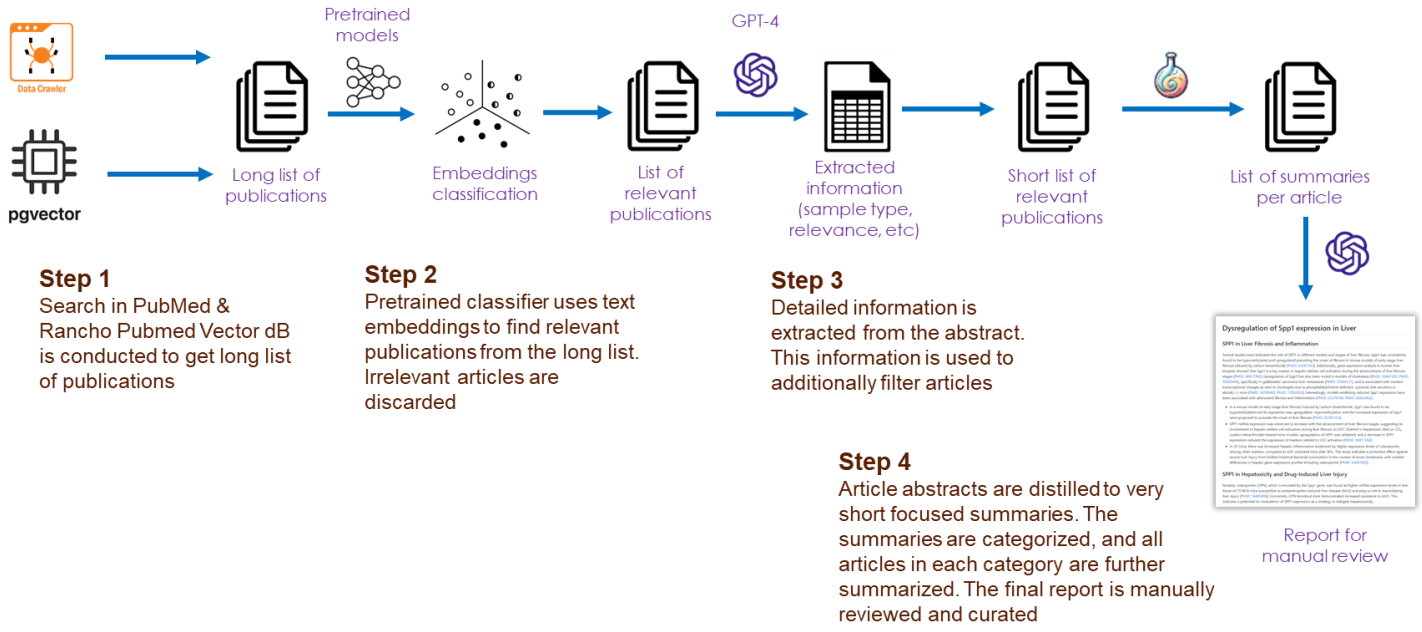
Section 14: DisGeNET Biomarker Associations to Disease in Organ of Interest

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-4-turbo* 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 expression" OR "transcriptional biomarkers") AND (toxicity OR toxicology OR disease) AND ({gene} OR {aliases}) AND (mouse OR rat OR human) AND ({organ})

where {gene} and {aliases} are corresponding gene and its aliases as specified in section 1; {organ} – organ for which the information is retrieved. In the cases where no/little references were found, the query was extended to

> ("gene expression" OR "transcriptional biomarkers") AND ({gene} OR {aliases}) AND (mouse OR rat OR human) AND ({organ} OR {organ_aliases})

In addition to this query, the PubMed vector database was used to augment search results by looking at similar relevant articles.

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-4-turbo*), and only articles following the rules were retained.

Information extracted	Possible values	Rules
Does abstract have information about gene of interest	True, False	True
Was change of mRNA expression of the gene of interest mentioned in the article?	True, False	True
Is study related to the tissue or organ of interest?	True, False	True

Information extracted	Possible values	Rules
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 *human*, *rat* 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 downregulated not significant not mentioned	not "*not significant*"

Step 4 – Distilling and summarization: each selected abstract was distilled using gpt-4-turbo 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.

These short reports were used by curators to prepare section 2 reports for the genes or interest. An example of a fully automatically generated report is provided in Appendix A.

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 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 (Summary)

To generate a summary report for section 9, a multi-agent interaction system was implemented. The “researcher” agent was tasked with providing explanation of why a gene is dysregulated in diseases and toxicities associated with certain tissue. The agent was instructed about the specific format of the summary report (the report contains 2 paragraphs of text, should avoid excessive introductory language, should be function-focused). The “qcer” agent was provided with the same context and rules and was instructed to review the reports generated by “researcher” and make sure that the reports satisfied the constraints and were factually correct. If the case report did not satisfy the requirements, “qcer” agents were to provide feedback on what needed to be changed. Finally, “moderator” agents were moderating conversation between agents and deciding if the process needed to stop. Conversation was also stopped if the number of interactions between qcer and researcher agents exceeded 10 replies by qcer. A gpt-4-turbo LLM was used for qcer and researcher, and gpt-3.5-turbo was used for moderator. Three or more rounds of report generation were conducted for every gene/tissue pair; resulting reports as well as conversation logs were reviewed by human curator and assembled into the final report.

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

You are 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 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 it 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.

Section 12: Role of Gene in Other Tissues

The reports were assembled similarly to section 2, except that instead of a single focused tissue, several reports on several tissues were generated, including liver; kidney; heart; skeletal muscle; brain; lung; bone marrow.

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

77 of 125 reports were manually reviewed by a different curator to ensure consistency. In addition to that, all reports were QCed automatically. Two types of QC checks were performed:

Check type	QC check	Relevant sections
Basic	Report section is present	All
	Report section is not blank	All
	All sub-sections are present	6
	References have correct format (e.g. [PMID: 12343])	All
	No double spaces are present	All
	No non-ascii encoding is used	All
	Each list item has supporting references	2, 9, 10, 12, 13, 14
Deep	References are relevant for corresponding organ	13, 14
	References support statements	2, 9, 10, 12

Basic QC checks were performed automatically by using pattern matching. Deep QC checks were conducted using LLM (gpt-4-turbo), and results were reviewed manually.

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. For example, human REG3A was curated for mouse Reg3b, and human AKR1B10 was curated for rat Akr1b8. In

these cases, a corresponding comment was made in Section 1, and all automatically extracted information was based on human genes.

For the majority of the genes (88 out of 106), the full-scale curation as described above was done. However, 18 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 tissue.

Discussion

On average, each report contained 144 PubMed references, 33 protein-protein interactions, and 7 pathways:

Metric	Average count per report
PubMed references in all sections	146
Protein-protein interactions	33.3
Pathways	8.4
GO Terms	23
MSigDB signatures	21.4
Chemicals, known to elicit transcriptional response	19
DisGenNet diseases	3

The distribution of number of references per section (Figure 1) shows the highest number of references per section 13 (chemicals) and 4 (protein-protein interactions):

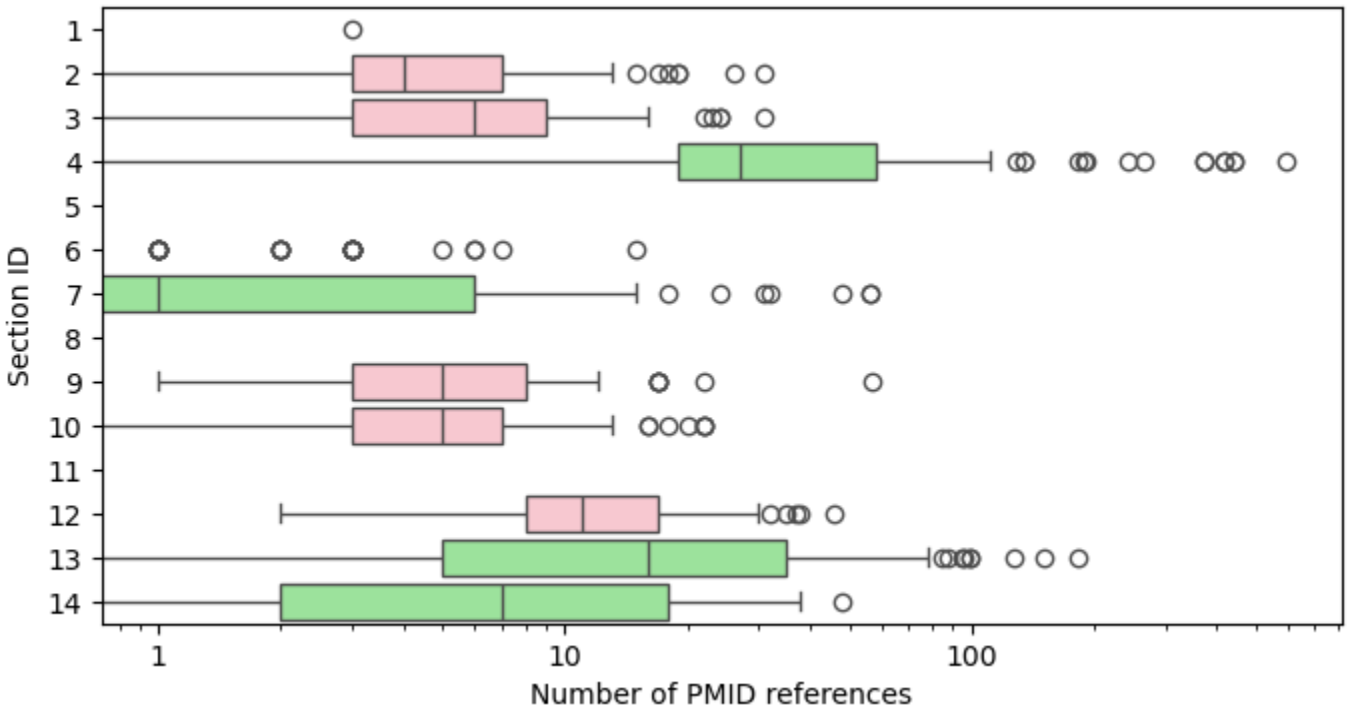


Figure 1 – number of references per section. Manually generated sections are highlighted in red, automatically generated sections are highlighted in green.

Reports that contained the most information are shown in the table below:

Gene	Organ	Refe-rences	Inter-actions	Pathways	GO terms	MsigDB	Chemi-cals	Diseases
Hsp90aa1	Skin	678	67	21	35	72	4	4
Cdkn1a	Liver	647	63	23	59	101	100	3
Myc	Liver	636	66	17	116	101	86	11
Cdk1	Liver	542	95	51	45	101	41	4
Myc	Kidney	535	66	17	116	79	25	9
Cdk1	Kidney	507	95	51	46	31	16	3
Cdkn1a	Bone Marrow	501	63	23	59	101	14	1
Cyp1a1	Liver	384	81	6	44	35	90	1
Jun	Skin	344	49	24	67	101	13	3
Cd44	Kidney	327	198	6	37	15	26	2

Gene	Organ	Refe-rences	Inter-actions	Pathways	GO terms	MsigDB	Chemi-cals	Diseases
Ccl2	Lung	245	66	4	90	47	39	11
S100a4	Kidney	232	97	3	2	8	15	5
S100a4	Liver	228	97	3	2	24	18	5
Il1b	Colon	212	41	19	120	48	34	5
Serpine1	Kidney	200	43	23	77	30	28	12
Igf1	Lung	107	28	5	145	42	13	4
Cp	Kidney	79	27	6	11	101	20	2

Many of the genes that had the most information were either transcriptional regulators (Myc, Jun) or were related to cell cycle checkpoints (Cdk1, Cdkn1a). Reports that contained the least amount of information are shown in the table below:

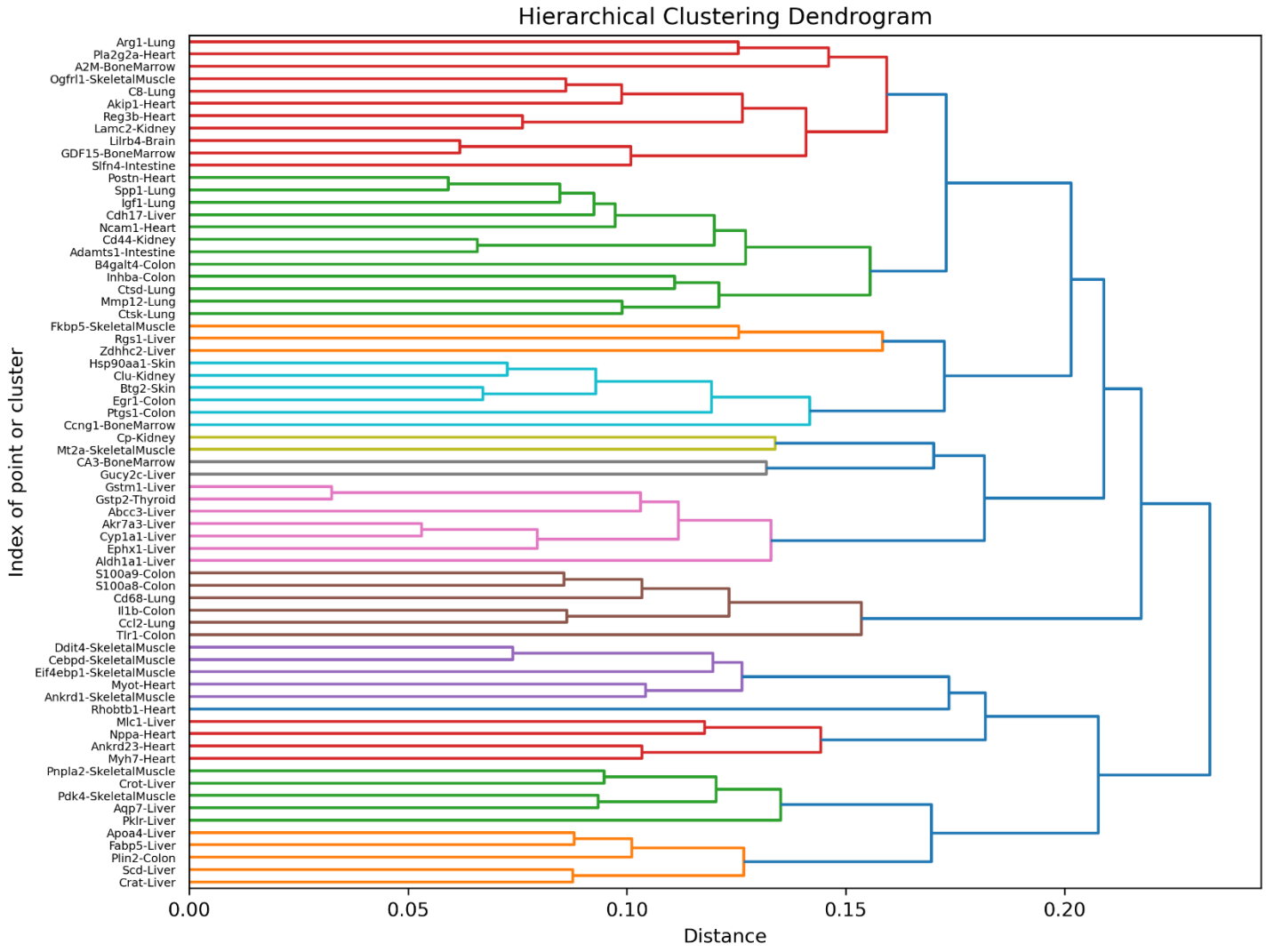
Gene	Organ	Refe-rences	Inter-actions	Pathways	GO terms	MsigDB	Chemi-cals	Diseases
Ogfrl1	Skeletal Muscle	11	1	2	1	2	0	0
Slc66a3	Liver	33	1	1	1	1	16	0
CA3	Bone Marrow	35	8	1	4	2	3	3
Neurl3	Liver	37	2	4	3	3	10	0
Ankrd23	Heart	38	27	2	5	1	1	0

These are relatively understudied genes with < 30 total number of references in PubMed.

Semantic analysis of genes function

To characterize the similarities between roles of genes in diseases and toxicities, a semantic analysis of Section 9 summaries was conducted. These summaries were generated automatically from information in other sections and contain an explanation of why the gene is up/downregulated from

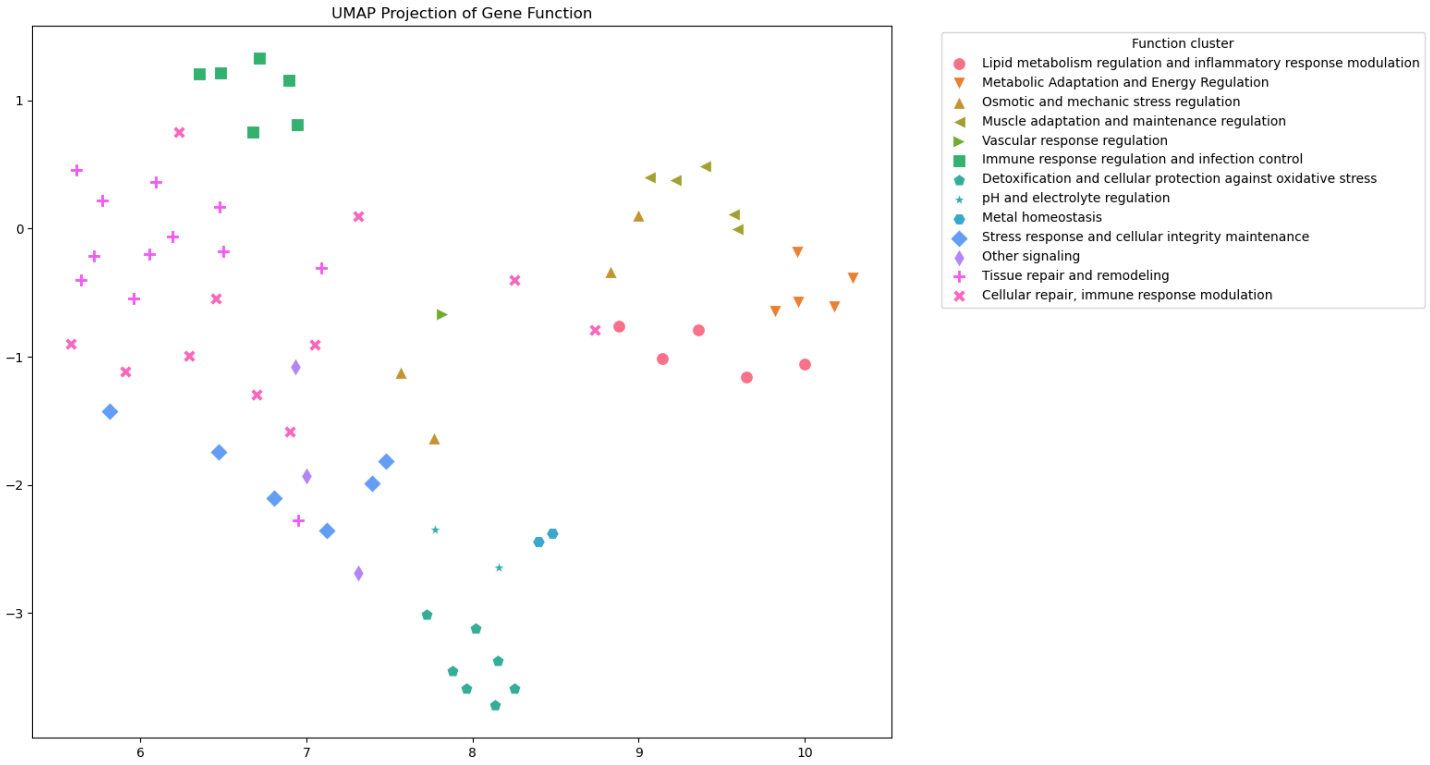
the evolutionary point of view. These summaries were further compressed into 1-sentence statements and mapped to semantic vector space using a *gte-large* model. The resulting vectors were then clustered using Hierarchical Agglomerative Clustering using cosine distance as the measure of dissimilarity between points.



Functionally similar genes were identified which have the same role when upregulated / downregulated in different tissues.

Cluster	Cluster Name	Examples
1	Lipid metabolism regulation and inflammatory response modulation	Colon: Plin2 Liver: Crat, Scd, Fabp5, Apoa4
2	Metabolic adaptation and energy regulation	Liver: Aqp7, Crot, Pklr Skeletal Muscle: Pnpla2, Pdk4
3	Osmotic and mechanic stress regulation	Heart: Nppa, Myh7, Ankrd23 Liver: Mlc1
4	Muscle adaptation and maintenance regulation	Heart: Myot Skeletal Muscle: Ankrd1, Eif4ebp1, Cebpd, Ddit4
5	Vascular response regulation	Heart: Rhobtb1
6	Immune response regulation and infection control	Colon: S100a8, Tlr1, S100a9, Il1b Lung: Cd68, Ccl2
7	Detoxification and cellular protection against oxidative stress	Liver: Abcc3, Cyp1a1, Gstm1, Akr7a3, Ephx1, Aldh1a1 Thyroid: Gstp2
8	pH and electrolyte regulation	Bone Marrow: CA3 Liver: Gucy2c
9	Metal homeostasis	Kidney: Cp Skeletal Muscle: Mt2a
10	Stress response and cellular integrity maintenance	Bone Marrow: Ccng1 Colon: Egr1, Ptgs1 Kidney: Clu Skin: Hsp90aa1, Btg2
11	Other signaling	Liver: Rgs1, Zdhhc2 Skeletal Muscle: Fkbp5
12	Tissue repair and remodeling	Colon: B4galt4, Inhba Heart: Ncam1, Postn Intestine: Adamts1 Kidney: Cd44 Liver: Cdh17 Lung: Spp1, Ctsk, Igf1, Mmp12, Ctsd
13	Cellular repair, immune response modulation	Bone Marrow: GDF15, A2M Brain: Lilrb4 Heart: Akip1, Pla2g2a, Reg3b Intestine: Slfn4 Kidney: Lamc2 Lung: C8, Arg1 Skeletal Muscle: Ogfrl1

Analysis of the functions shows that some are tissue-specific (i.e., cluster 6 – infection control and immune response modulation is relevant for lung and colon; metabolic adaptation and energy regulation genes are specific to liver and muscles), whereas other functions are common across different tissues (cellular repair, tissue repair and remodeling). Function visualization with UMAP projection shows both universal functions (cellular repair), and highly specific and isolated functions (detoxification; infection control).



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, html and md formats. Template for manual curation could be found in html subdirectory.
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.

Appendices

Appendix A – Example of fully automatic report

Dysregulation of Spp1 expression in Lung

Spp1 Expression in Mouse and Rat Models

Spp1 (Secreted Phosphoprotein 1) has been widely studied in various mouse and rat models to elucidate its role during lung development and in response to lung injury and diseases. In various rodent strains, such as C3H/HeJ and JF1/MsJ mice, Spp1 expression levels exhibited variations across different stages of lung development [PMID: 28793908]. Notably, intrauterine growth-restricted Wistar rats that were postnatally hyperalimented demonstrated increased Spp1 expression, suggesting a relationship with lung development disruptions [PMID: 25411031]. Spp1 expression has been implicated in the pathogenesis of fibrosis in scleroderma mouse models [PMID: 34289031] as well as in multiwall carbon nanotube-induced pulmonary granulomas in certain knockout mice [PMID: 30848658]. Contrarily, a diet containing transgenic tomatoes in BALB/c mice led to a decrease in Spp1 gene expression and associated lung tumor reduction [PMID: 29899427]. During crucial stages such as alveologenesis, marked differences in Spp1 expression were observed across different mouse strains, linking its expression to lung morphology and function [PMID: 24816281].

- Spp1 (Secreted Phosphoprotein 1) expression in C3H/HeJ and JF1/MsJ mice was investigated, but specific details on its expression level during the stages of lung development and its impact on lung function were not provided in the abstract [PMID: 28793908].
- In Wistar rats subjected to intrauterine growth restriction (IUGR) and postnatal hyperalimentation (HA), there was greater gene expression of Spp1 (osteopontin) in those with IUGR-HA compared to control and IUGR groups, without HA. This was associated with increased deposition of bronchial subepithelial connective tissue [PMID: 25411031].
- In a mouse model of scleroderma, significant upregulation of Spp1 was observed in fibrotic skin and lung tissue. TGF-β pathway inhibition via oral ALK5 inhibitor SB525334 was assessed, affecting Spp1 expression as part of the transcriptomic changes noted during anti-fibrotic therapy [PMID: 34289031].
- In ABCG1-KO and ABCA1/ABCG1 double-KO mice instilled with multiwall carbon nanotubes, there was increased expression of Spp1 (osteopontin) in bronchoalveolar lavage cells, which is associated with the formation and maintenance of pulmonary granulomas. ABCA1-KO mice did not show this upregulation or an effect on granuloma formation [PMID: 30848658].

- In BALB/c mice injected with CT-26 colon cancer cells, a diet containing 0.06% transgenic tomatoes expressing the apoA-I mimetic peptide 6F (Tg6F) resulted in decreased Spp1 gene expression in both the jejunum and lungs. Additionally, Tg6F-fed mice presented with reduced plasma levels of Spp1 and lower numbers of pro-tumorigenic myeloid-derived suppressor cells, correlating with fewer lung tumors [PMID: 29899427].
- During alveologenesis, JF1/Msf mice displayed reduced Spp1 mRNA and protein levels in the lungs compared to C3H/HeJ mice. Spp1((-/-)) mice exhibited smaller lungs with increased compliance and enlarged airspaces, while a microarray identified disrupted regulation of transcripts associated with lung development and respiratory disease, including decreased levels of specific lung development-related genes in Spp1((-/-)) mice during the peak stage of alveologenesis [PMID: 24816281].

Spp1 in Human Lung Diseases

Spp1 expression is significantly increased in alveolar macrophages of smokers and patients with idiopathic pulmonary fibrosis (IPF), distinguishing it as a key gene associated with the pathology of these diseases [PMID: 16166618, PMID: 33672678, PMID: 34285914, PMID: 37686108, PMID: 36507532]. Similarly, marked upregulation of Spp1 was found in lung adenocarcinoma and non-small cell lung carcinoma (NSCLC), with correlations established between its expression and metastasis, tumor growth, and patient survival [PMID: 26081616, PMID: 22369099, PMID: 30639873, PMID: 32503462, PMID: 20224789, PMID: 31792339, PMID: 18210878, PMID: 30832751, PMID: 34234236, PMID: 33626243]. Notably, Spp1 expression was also higher in lung squamous cell carcinoma (LUSC) than adenocarcinoma (LUAD), suggesting possible subtype-specific roles [PMID: 33244273].

- In human alveolar macrophages from smokers, Spp1 (Osteopontin) expression was increased. This finding differs from changes observed in transgenic mouse models of emphysema [PMID: 16166618].
- Spp1 was identified as one of the 11 key candidate genes upregulated in idiopathic pulmonary fibrosis (IPF) samples when compared to normal samples, and this finding was confirmed with RT-qPCR and immunohistochemical staining [PMID: 33672678].
- In the study, SPP1 was identified as a microRNA (miRNA) target among differentially expressed genes in idiopathic pulmonary fibrosis (IPF) compared to controls. Spp1 gene not explicitly evaluated in the context of IPF cell models or mouse genotypes. Effects of silencing or upregulating SPP1 on IPF were not discussed [PMID: 34285914].
- Spp1 was identified as one of the eight hub genes associated with the pathogenesis of idiopathic pulmonary fibrosis (IPF) and was confirmed in lung tissue from a mouse model. The gene's

involvement in IPF was also analyzed in relation to potential interactions with target microRNAs miR-181b-5p, miR-4262, and miR-155-5p [PMID: 37686108].

- Spp1 mRNA expression was significantly higher in IPF tissue compared to the normal group. Spp1 was associated with monocytes, plasma cells, neutrophils, and regulatory T cells in the immune cell infiltration analysis [PMID: 36507532].
- SPP1 was identified as one of the 20 genes with the largest fold changes in both nonsmoking and smoking patients with lung adenocarcinoma when comparing tumor to normal tissues. The gene was similarly and prominently differentially expressed in tumor samples from both nonsmoker and smoker patients [PMID: 26081616].
- Spp1 was identified as one of the AC diagnostic biomarkers in human lung adenocarcinoma tissues compared to normal lung tissue. The study validated SPP1 using RT-PCR on a tissue array [PMID: 22369099].
- PM exposure increased OPN expression in the bronchial epithelium, serum, and BALF of mice. OPN silencing alleviated PM-induced inflammatory responses in HBECs [PMID: 30639873].
- OCT4A and Spp1 (SPP1C variant) are co-expressed in aggressive human breast, endometrial, and lung adenocarcinoma cell lines, as well as in primary, early-stage lung adenocarcinoma tumors. Ablation of OCT4-positive cells in lung adenocarcinoma cells significantly decreased Spp1C mRNA levels [PMID: 32503462].
- In SCID mouse xenografts, overexpression of OPN-A and OPN-B increased local tumor growth and lung metastasis, while mutating or deleting their RGD domain resulted in decreased tumor growth and metastasis. The RGD domain of OPN-A and OPN-B inhibited apoptosis by triggering NF-kappaB activation and FAK phosphorylation [PMID: 20224789].
- In non-small-cell lung cancer (NSCLC) patients, higher Spp1 (Osteopontin, OPN) mRNA levels were observed in tumor tissues. Knockdown of Spp1 in A549 lung cancer cells reduced their migration and invasion, whereas overexpression in SK-MES-1 cells increased these properties, with accompanying RON phosphorylation activation. RON inhibition tempered OPN-induced cell invasiveness and epithelial-mesenchymal transition. Spp1 is an independent indicator of survival and is implicated in NSCLC progression [PMID: 31792339].
- High mRNA and protein expression of Spp1 (OPN) was detected in non-small cell lung cancer (NSCLC) tissues compared to adjacent normal tissues, with a particularly increased presence in squamous cell carcinomas (SCCs). Over-expression of OPN was considerably correlated with the occurrence of lymph node metastases in SCC [PMID: 18210878].
- SPP1 was significantly increased in afatinib-resistant lung cancer cells and lung cancer tissues as compared with parental cells and adjacent normal tissues, respectively. Knocking down SPP1 in resistant lung cancer cells increased their sensitivity to afatinib and reduced their invasive ability. Gene not mentioned explicitly in any in vivo experiments [PMID: 30832751].

- SPP1 mRNA was significantly increased in ALK-positive lung cancers. SPP1 overexpression was associated with poor outcomes for patients with ALK fusion lung cancer not receiving targeted therapy [PMID: 34234236].
- SPP1 was the only differentially expressed gene found to be up-regulated in both COPD and lung cancer (LC) patients compared to healthy controls. High SPP1 expression was associated with shorter survival time in LC patients [PMID: 33626243].
- SPP1 (Osteopontin) was found to be overexpressed in non-small cell lung carcinoma (NSCLC) stage IIIA samples from patients, more so in lung squamous cell carcinoma (LUSC) compared to lung adenocarcinoma (LUAD) [PMID: 33244273].

Fibrosis, Tumorigenesis, and Immune Responses

Intrinsic to its role in lung tissue pathology is Spp1's involvement in fibrosis. Upregulation of Spp1 after thoracic irradiation or bleomycin instillation is notably attenuated through small-molecule inhibitor treatments [PMID: 27467922, PMID: 11245625]. Additionally, Spp1 expression in SCID mouse xenografts and in the promotion and progression stages of lung tumorigenesis in mice suggests its contribution to lung cancer pathways [PMID: 26141346, PMID: 19925653, PMID: 37794454]. Spp1's role in immune responses is highlighted in smoke-exposed mice, where its high expression level in lung antigen-presenting cells is crucial for pathological T(H)17 responses leading to emphysema [PMID: 22261033].

- In a mouse model treated with 20 Gy thoracic irradiation to induce pulmonary fibrosis, combined small-molecule inhibition of TGFβ and PDGF signaling decreased radiation-induced pulmonary inflammation and fibrosis and extended survival. It was found that SPP1 (Osteopontin) expression was attenuated in the irradiated lung tissue following the treatment [PMID: 27467922].
- In a mouse model of pulmonary fibrosis induced by intratracheal instillation of bleomycin, Spp1 mRNA was notably induced and associated with the fibrotic process in the lung. Treatment with an anti-alphav integrin monoclonal antibody (RMV-7) repressed the extent of pulmonary fibrosis, suggesting the involvement of OPN in fibrosis development [PMID: 11245625].
- In vivo experiments using two lung cancer cell-xenograft mouse models showed that intravenous injection of PSOT/siOPN complexes resulted in reduced OPN mRNA expression and led to the suppression of tumor volume and weight [PMID: 26141346].
- Increases in Spp1 mRNA expression were observed during both the promotion and progression stages of lung tumorigenesis in BALB(Lps-d) mice compared to BALB/c mice. In BALB/c mice, there was significantly reduced expression of Spp1 in both the tumors and uninvolved tissue [PMID: 19925653].

- Spp1 was identified as a hub gene in the pathogenesis of idiopathic pulmonary fibrosis (IPF) and was verified in a pulmonary fibrosis mouse model. Sea Buckthorn and Gnaphalium Affine were predicted as potential traditional Chinese medicines targeting IPF [PMID: 37794454].
- Spp1 was highly expressed in lung antigen-presenting cells (APCs) of smoke-exposed mice and was necessary for T(H)17 responses and the development of emphysema in vivo. This high expression of Spp1 inhibited the expression of the transcription factor Irf7, playing a crucial role in the induction of the pathological T(H)17 responses associated with emphysema [PMID: 22261033].

Spp1 and Other Related Diseases

Beyond lung-specific diseases, Spp1 participation extends to other systemic conditions, such as aortic dissection and bone diseases. For instance, β-aminopropionitrile monofumarate (BAPN) treatment, which induces extracellular matrix softening, led to upregulation of osteopontin in the aortas of mice [PMID: 32500384]. The MEPE gene, which is related to tumor-induced osteomalacia, was identified in the tumor secretome, highlighting a potential systemic impact of Spp1-related genes [PMID: 10945470].

- In mice treated with the lysyl oxidase inhibitor β-aminopropionitrile monofumarate (BAPN), which induces extracellular matrix (ECM) softening, there was upregulation of osteopontin in the aortas. This change suggests a phenotypic switch of vascular smooth muscle cells (VSMCs) to a synthetic phenotype, and these mice subsequently developed severe aortic dissection [PMID: 32500384].
- MEPE gene has been identified as a candidate for the tumor-secreted phosphaturic factor in oncogenic hypophosphatemic osteomalacia (OHO). High-level MEPE mRNA expression was observed in all four OHO tumors examined. The gene not mentioned explicitly [PMID: 10945470].

Biomarker Potential and Therapeutic Impacts

Recognizing the biomarker potential of Spp1, its mRNA expression serves as an independent prognostic indicator across a variety of lung conditions, such as idiopathic pulmonary arterial hypertension (IPAH) and chronic obstructive pulmonary disease (COPD). The association with shorter survival in lung cancer patients underscores Spp1's relevance in clinical prognostics [PMID: 32714978, PMID: 33626243, PMID: 31181055]. Additionally, specific interventions targeting Spp1, such as the use of nimodipine or anti-alphav integrin antibody, illustrate the gene's potential as a therapeutic target in pulmonary hypertension and fibrosis [PMID: 35255963, PMID: 11245625].

In summary, Spp1's expression in the lungs is closely associated with lung development, injury, fibrosis, tumorigenesis, and immune responses. Variation in its expression across different conditions, species, and tissues underlines its complexity and multifaceted role in pulmonary physiology and pathology. Its implication as a biomarker and potential therapeutic target further reinforces the importance of Spp1 as a focal point in lung disease research and treatment strategies.

- In the lung tissue of patients with idiopathic pulmonary arterial hypertension (IPAH), SPP1 was identified as one of the top 10 hub genes in a protein-protein interaction network analysis derived from a DEMI-DEG network comprising 4 differentially expressed miRNAs and 16 differentially expressed genes. Nimodipine was pinpointed as a potential drug targeting hub genes including SPP1 [PMID: 32714978].
- Spp1 was identified as one of the two downregulated hub genes in non-small cell lung cancer (NSCLC) patients. The expression of Spp1 is correlated with the prognosis of NSCLC patients [PMID: 31181055].
- In the pulmonary arteries of hypertrophic pulmonary hypertension (HPH) rats and lung tissues of pulmonary artery hypertension (PAH) patients, Spp1 mRNA expression was found to be dysregulated. The gene was identified as part of a competitive endogenous RNA (ceRNA) regulatory network that potentially plays a significant role in the progression of HPH [PMID: 35255963].

Appendix B – 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 lead 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
entrez_description.xlsx	Gene descriptions extracted from Entrez

File	Description
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 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 needs to be annotated
disgennet_diseases_annotated.tsv	Annotated disease-organ pairs. Each pair relevance is binned into “definitely yes”, “definitely no” and “possible”