Standard Operating Procedures (SOP) for NIEHS Manual Curation of Toxicity Biomarker Data

Bing Zhou, Ph.D.

Oleg Stroganov, Ph.D.

September 19, 2024

Version 1.2

Rancho biosciences

Contents

I. Document History

Version: 1.2

Last Revised Date: Sep. 19 , 2024

Revisions History

Version	Date	Brief description of changes
1.2	Sep. 19, 2024	Updated manual curation of section 6 (pathways)
1.1	July 22, 2024	Added technical specifications of Abstract Distiller
1.0	July 18, 2024	First version of the document

Approval

This SOP is approved by NIEHS and Rancho BioSciences and is effective from July 2024.

II. Project Objectives

The Division of Translational Toxicology (DTT) at the National Institute of Environmental Health Sciences (NIEHS), NIH, focuses on environmental factors impacting human health, emphasizing toxicity biomarker identification in various organs. The purpose of this project is to generate comprehensive summaries of transcriptional biomarkers for toxicity/disease in different tissues. Specific tissues for which biomarkers shall be identified and curated may include aorta, bone marrow, brain, cecum, colon, duodenum, epididymis, esophagus, eye, harderian gland, heart, ileum, jejunum, kidney, liver, lung, lymph node, mammary gland, ovary, pancreas, parathyroid gland, peripheral nerve, pituitary, prostate, salivary gland, seminal vesicle, skeletal muscle, and skin. The DTT is specifically interested in exploring the association of transcriptional up- or down-regulation of the biomarkers with adversity (toxicity or disease) in mammalian models (with an emphasis on rats). These biomarker reports of specific gene/tissue pairs will facilitate DTT with their applicability in *in vivo* toxicological screening processes, specifically to serve as a justification for using the identified genes to characterize adversity at the level of the transcriptome.

Rancho BioSciences previously collaborated with DTT on Phase 1 of this project, successfully delivering 125 biomarker summaries for NIEHS. This encompassed analysis across 106 genes in 11 tissues. Building upon this foundation, Rancho is now tasked with expanding these efforts by preparing an additional 100 reports for DTT. Additionally, Rancho will innovate by developing an AI-driven, agentic pipeline and repeatable workflow using large language models (LLM). This advancement aims to rigorously assess consensus confidence regarding the association of transcriptional biomarkers with target organ toxicity.

The data curation process led by Rancho, involved a meticulous review of public literature and integration of data from renowned gene annotation platforms such as Uniprot, MSigDB, RGD, HPA, WikiGenes, DisGeNet, and others. These curated reports offer comprehensive insights across multiple dimensions:

- Gene Expression Alterations: Detailed analysis of alterations in gene expression and their correlations with toxicity and disease states.
- Protein Structure and Function: Exploration of protein structure and its functional implications, including identification of proteins influencing or being regulated by the gene of interest.
- Pathways and Signatures: Overview of associated pathways and signatures that elucidate the biological contexts of gene function under pathological conditions.
- Regulation Mechanisms: In-depth examination of the intricate mechanisms governing gene regulation in pathological contexts.
- Cellular Localization and Expression Patterns: Comprehensive information on the cellular localization and expression patterns of the gene across various tissues, including responses to chemicals known to induce transcriptional biomarkers in specific tissues.

- Disease and Pathology Insights: Distilled summary of the gene's role in disease and pathology based on curated data, ensuring a nuanced understanding of its implications.

This approach ensures a thorough analysis that meets the rigorous standards required for assessing transcriptional biomarkers in the context of organ toxicity.

III. Scope of the SOP

This SOP governs all personnel engaged in manual data extraction and curation activities within Rancho for the "NIEHS Extraction of High Content Toxicity Biomarker Data" project. It outlines roles and responsibilities across the data curation life cycle, details the data extraction methodology, identifies relevant data sources, and establishes procedures for data curation and quality control (QC). This SOP assumes continuity from Phase 1 reports delivered by Rancho, shaping Phase 2 efforts to maintain consistency and quality. The primary objectives of this SOP are to ensure thorough documentation, training, and adherence to guidelines for transparency and reproducibility; to uphold data integrity and consistency across all toxicity biomarker reports.

IV. Responsibilities

Data Curators

Curators are Ph.D.-level data integrity scientists possessing domain expertise. They are tasked with conducting primary data queries/extractions and performing quality control (QC) checks. Curators initiate literature searches utilizing Rancho's Data Crawler, extracting pertinent data from identified articles and various databases. Organizing extracted information into a curation template in Wiki pages hosted by Rancho's server is also part of their responsibility. Before submission, Curators conduct rigorous quality checks on their work. Curators are also responsible for checking other curators' data extraction upon request.

Senior Curators

Senior Curators serve as subject matter experts responsible for developing and maintaining the curation template and specifications tailored to the project's unique scope and complexity, ensuring alignment with client requirements. They recommend reputable and reliable data sources throughout the data collection process. To maintain the highest standards, Senior Curators are responsible for spot checking the reports to ensure accuracy, consistency, and overall data integrity.

QC Managers

QC Managers collaborate with the Project Manager and Technical Lead to implement a QC Plan that meets project quality and timeline requirements. They conduct periodic reviews of the project

to ensure adherence to the QC Plan and assist in resolving any discrepancies that cannot be resolved by Senior Curators.

V. Materials and Tools

Primary Resources

1) Ph.D.-level Data Integrity Scientists: Highly qualified experts responsible for conducting data extraction and ensuring integrity throughout the process.
2) Computers with Internet Access: Essential for conducting literature search and accessing online databases.
3) Rancho's Data Crawler App: This is an internally developed Data Crawler app, a web-based tool, to facilitate targeted data queries from peer-reviewed literature in scientific journals. 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, a repository of life science literature hosted by NCBI. This sophisticated approach allows us to identify research articles that align with complex requirements, thereby streamlining the data acquisition process for enhanced research outcomes. Curators use this tool to generate lists of relevant publications in Excel format.
4) Rancho's Wiki System: The Wiki system serves as the primary platform for hosting the curation template and organizing extracted reports by tissue types. Curators create wiki pages using the template to perform data extraction tasks based on predefined guidelines to ensure accuracy, consistency and accessibility. The wiki system allows Curators to strategically flag the progress of manual curation on each wiki page. Curators utilize a "Comments" section on each wiki page to communicate revisions or updates related to the curated content.
5) Rancho's Abstract Distiller Tool: Abstract Distiller tool is used to facilitate summarization of relevant findings from publication abstracts, aiding in manual extraction and validation of data. Curators use 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 be specified by Curators. Technical specification for Abstract Distiller is provided in Appendix B.

Reference Database

Leverage reputable sources including the following databases for comprehensive data retrieval.

1) **PubMed**: For literature-based evidence, PubMed is a comprehensive database of biomedical literature, where you can search for articles related to gene function, regulation, diseases, and more. https://www.ncbi.nlm.nih.gov/pubmed/

2) **Google Scholar:** Google Scholar provides a simple way to broadly search for scholarly literature. Search across a wide variety of disciplines and sources. https://scholar.google.com/
3) **GeneCards**: Integrates information from various sources on human genes, proteins, and diseases, including gene function, expression, pathways, and related publications. https://www.genecards.org/
4) **UniProt**: Comprehensive resource for protein sequence and function annotation, including protein-protein interactions, post-translational modifications, and disease associations. https://www.uniprot.org/
5) **Kyoto Encyclopedia of Genes and Genomes (KEGG)**: KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism, and the ecosystem, from molecular-level information. The data-oriented entry points include pathways, genes, diseases, and drugs. https://www.genome.jp/kegg/
6) **Reactome:** REACTOME is an open access, manually curated and peer-reviewed pathway database. https://reactome.org/
7) **WikiPathways:** is an open, collaborative platform dedicated to the curation of biological pathways. https://www.wikipathways.org/
8) **Rat Genome Database (RGD):** a database for genetic, genomic, phenotype, and disease-related data generated from rat research. RGD has expanded to include a large body of structured and standardized data for ten species. RGD stores data about various "objects", including genes, strains, QTLs, orthologs, cell lines, gene promoters. https://rgd.mcw.edu/rgdweb/homepage/
9) **NCBI Gene**: Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide. https://www.ncbi.nlm.nih.gov/gene/
10) **WikiGenes:** WikiGenes is a non-profit initiative to provide a global collaborative knowledge base for the life sciences. https://www.wikigenes.org/
11) **Automatically extracted information**: contains gene/tissue pair-specific information extracted by GPT-4 from abstracts of literature. These documents serve to further filter publications and enhance data relevance. Rancho's SharePoint

VI. Procedure

Curation Procedure Overview

The process of curating biomarker reports combines manual curation with automated data extraction. The manual curation process operates within Rancho's dedicated wiki system. Curators create wiki pages utilizing a standardized template to ensure consistency and clarity in entering biomarker information. Data extraction involves meticulous review of LLM pipeline summaries,

literature searches and information extraction for topics identified and prioritized by DTT. Detailed data extraction requirements will be included in the following sections. Curators utilize Data Crawler App for targeted data queries from peer-reviewed journals. Rancho's Abstract Distiller tool aids in summarizing relevant findings from abstracts efficiently. Curation progress is tracked through strategic flags in the wiki system, covering stages from initial data extraction through QC review and subsequent revisions based on feedback. This SOP emphasizes adherence to the curation template for uniformity, the use of automated tools for efficiency, and rigorous QC checks to uphold data accuracy, supporting comprehensive and reliable biomarker reports.

This SOP outlines the standard curation procedures for curators to effectively utilize the Wiki system for data curation across specific sections of the report:

1) Sections 1, (6), 9, and 10:
   These sections are curated exclusively through manual data query and extraction. The LLM pipeline results are not available for reference in these sections. **For section 6, no manual curation will be performed unless explicitly requested by the Tech Lead**.

2) Sections 2, 3, and 12:
   For these sections, the curation process begins by reviewing the summaries generated by the LLM pipeline for relevant data. If the required information is not adequately covered in the LLM-generated data, curators proceed with manual query and extraction to supplement the dataset.

Curation Setup in Wiki System

1) **Accessing the Wiki System**
   Curators log in to the Wiki system at https://tox-wiki.rbsapp.net/ using their Rancho or external credentials.
2) **Creating a New Wiki Page**
   To initiate a new curation page, curators click on the "Add New Page" icon located at the top right corner of the interface.
3) **Selecting Page Location**
   Curators navigate to the appropriate tissue folder within the wiki system (e.g., /genes/Brain/GeneName) to specify the location for the new page. They then click on the "SELECT" button to confirm the directory.
4) **Choosing the Template**
   From the editor options, curators select "From Template". In the pop-up window, they choose the "Manual gene annotation template" and proceed by clicking "SELECT".
5) **Defining Page Properties**
   In the subsequent dialog box, curators define the page properties as follows: Title = Gene Name, Short Description = Gene Name – Tissue Name, Tags = Data Extraction.
   After confirming the details, curators click "Ok" to proceed. A manual curation template is readily available for immediate use.
6) **Saving Curation**

After curation and self-QC, click “PAGE” tag icon on the top right of the wiki page, update the “CATEGORIZATION” of the page properties as “to qc”. Then, click “OK”. Save the curation by clicking the “SAVED” icon on the top right of the wiki page. Then, “Close” the current page.

Manual Curation for Sections Without LLM Results

Once a wiki page is created under the proper directory for a gene/tissue pair, curators will extract information for each section within the Wiki system based on the data collection rules specified in the template. Please find example curation output in **Appendix A**. The procedures are as follows:

Section 1

Gene Aliases: a list of common gene aliases and additional information about gene names if required.

1) **GeneCards Database Search**
   Navigate to the GeneCards database at https://www.genecards.org/ and search for the Gene Name of interest.
2) **Extracting Aliases from GeneCards**
   Access the "Aliases" section within the GeneCards entry. Copy all listed aliases related to the gene and paste them into a text editor such as Notepad++.
3) **NCBI Gene Database Search (if necessary)**
   If the Gene Name is not found in GeneCards (common for genes identified in non-human species), navigate to the NCBI Gene database at https://www.ncbi.nlm.nih.gov/gene/. Search for the gene name of interest. And select the corresponding gene ID for Homo sapiens.
4) **Extracting Aliases from NCBI Gene**
   From the NCBI Gene report, locate the "Also known as" aliases under the "Summary" section. Copy these aliases to the text editor.
5) **Formatting Aliases**
   Remove citation numbers and comma within the gene names and concatenate the aliases using a comma and space (", "). Ensure all aliases are formatted consistently.
6) **Adding Reference Link**
   Include the reference link to the GeneCards page or NCBI Gene page within square brackets "[ ]" at the end of the alias list. Paste this formatted alias list into Section 1 of the curation template within Wiki system.

Section 6

GO Terms, MSigDB Signatures, Pathways Containing Gene with Descriptions of Gene Sets (i.e., pathways, ontologies, signatures) containing the gene. All gene sets identified should contain textual description.

Within this section, no manual curation will be performed unless explicitly requested by the Tech Lead.

1) **Pathway Term Retrieval from Literature**
   If no pathway information is found for the gene of interest in the GeneCards database, scientific literature will be used to extract pathway names.
   - Curators should initiate queries using PubMed or Google Scholar. Use both the gene name (gene aliases) and "pathway" as keywords.
   - For PubMed, set the filters as follows:
     - Text availability: Abstract, Free full text
     - Article type: Review
     - Species filter: Humans
     - For example, to find pathways related to the Chek2 gene, use the following query:
     - https://pubmed.ncbi.nlm.nih.gov/?term=Chek2+AND+pathway&filter=simsearch1.fha&filter=simsearch2.ffrft&filter=pubt.review&filter=hum_ani.humans
   - Review the abstracts of resulting publications to identify pathways associated with the gene of interest.
2) **Pathway Description Retrieval from Other Sources**
   Curator extract pathway descriptions by querying various reputable pathway annotation sources:
   - Access the Reactome pathway browser at https://reactome.org/PathwayBrowser/ for detailed pathway annotations.
   - Utilize WikiPathways at https://www.wikipathways.org/ to gather additional pathway descriptions.
   - Alternatively, conduct a Google search to obtain comprehensive pathway information related to the identified gene.

Section 9

*Mechanistic Information: A description of the general mechanistic role of the gene product in normal function and disease processes. The summary section provides **tissue- or organ-specific** mechanistic explanations for gene regulation in pathogenesis.*

1) **Relevant Literature Search in Data Crawler**
   To identify literature providing mechanistic insights into gene dysregulation under diseased or pathological conditions, particularly in human or rodent models relevant to specific

tissues, a refined and targeted query approach is employed. This procedure is illustrated using the example of “Anxa2 – kidney” pair.

- Access Rancho’s Data Crawler tool at https://datacrawler.rbsapp.net/.
- Enter a customized search string based on the following example into the “Search…” window (procedures on how to adjust the search string are detailed in item 2 of this section):
- Example search string:
  (Anxa2 OR "Annexin A2" OR Annexin-2) AND (kidney OR renal) 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)
- Choose the "Study-Level" option to ensure detailed examination of relevant studies instead of samples.
- Click on "Select Repo" and choose "PubMed" as the repository for conducting the search.
- Click "Submit" to initiate the query. Wait for the message "All done!" to confirm that the search process is complete.
- Once the search is finished, click on "Download Results" to obtain a list of publications that are relevant to the search query.

2) **Adjusting Query Parameters for Data Crawler**

The general procedure outlined above should be customized for specific gene name, tissue type and/or organisms as follows:

- Replace contents within the first set of parentheses with gene names delimited by "OR". Enclose gene names with spaces in quotation marks. e.g., update the gene name element like (Anxa2 OR "Annexin A2" OR Annexin-2) in the example search string (shown in section 1)) using your gene of interest.
- Avoid including gene aliases that are too short (<= 2 characters), or too long (> 3 terms), or numerical values to minimize irrelevant results.
- Replace contents within the second set of parentheses with specific tissues of interest, e.g., update the tissue name element like (kidney OR renal) in the example search string with your tissue of interest.
- If the resulting Excel file contains fewer than 5 publications, consider broadening the search by removing tissue constraints from the query.
- If the search yields more than 500 publications, refine the query by applying one or more of the following constraints:
  a. Restrict gene names to appear only in the Title or Abstract fields of articles (e.g., **add [Title/Abstract]** after each gene name or gene alias).
  b. Limit results to human studies (e.g., only include "human" in the organism specification part).
  c. Include only Review articles (e.g., add "**AND review[Filter]**" to the end of the query).

d. Narrow down results based on publication date (e.g., query publications from the last 10 years by adding "**AND y_10[Filter]**" to the end).
e. For example: (Anxa2[Title/Abstract] OR "Annexin A2"[Title/Abstract] OR Annexin-2[Title/Abstract]) AND (kidney OR renal) AND (pathogenesis OR toxicology OR pathophysiology OR toxicogenomics OR damage OR toxicity) AND (regulator OR regulation OR pathway OR mechanistic OR mechanism) AND (human) AND y_10[Filter]

3) Screening for Relevant Records in the Data Crawler Search Results

Use the following constrains to refine the screening of publications obtained from DataCrawler or Google Scholar:

- Conditional Formatting: Select columns containing title and abstract. Apply the "Conditional Formatting" function with "Highlight Cells Rules" set to "Text That Contains". Ensure the selected records mention the gene of interest or its aliases as well as the tissue of interest.
- For records meeting the above criteria, prioritize and reorder publications based on impact factor from high to low.
- Priority for Review Articles: Give priority to review articles in the final selection process.
- Ensure to include publications pertaining information on mechanism predefined by the client (please refer to the "Mechanism (per NIEHS)" column of the "Curation_Tracking_Sheet" for predefined mechanisms).

4) Search Literatures Referred by the NCBI Gene

Refer to this source only **when less than 5** relevant articles are identified through Data Crawler.

a. Navigate to the NCBI Gene database at https://www.ncbi.nlm.nih.gov/gene/ and conduct a search for the Gene Name.
b. Identify and select the "Human" gene that matches the gene symbol in the output table to access annotations specific to the gene of interest.
c. Navigate to the "General gene information" section.
d. Click "PubMed" corresponding to Evidence Code = IDA or IEP to access the abstracts of the original articles.
e. Proceed with information extraction facilitated by Abstract Distiller.

5) **Relevant Literature Search in Google Scholar**

Refer to this source only when less than 5 relevant articles are identified from previous steps.

- Open Google Scholar webpage at https://scholar.google.com/schhp?hl=en&as_sdt=0,34&as_rr=1
- Click the three-lined icon on the upper left conner, select the "Advanced search" option on the left panel.

- Fill in the following two fields with the suggested contents and click Search icon
  a. with all of the words = gene name AND tissue name AND mechanism AND review (use gene name and tissue name of the original gene/tissue pair, e.g., “Anxa2 AND kidney AND mechanism AND review”)
  b. with at least one of the words = gene aliases, tissues, toxicity, organisms and mechanism related terms separated by space, e.g., Anxa2 "Annexin A2" "Annexin 2" kidney renal pathogenesis toxicology pathophysiology injury damage toxicity pathway mechanistic mechanism human mouse rat
- Constrain the query by custom range of time (e.g., 2021-2024) to limit the number of results when over 1000 results are found.
- Sort results by relevance and review the top hit records. Click on the star just below citations you'd like to save to "My Library". Repeat as desired.
- Click on the "My Library" start icon to see all saved citations. And proceed with information extraction using Abstract Distiller once up to 10 relevant articles are identified.

6) Extract Mechanistic Information from Literature using Abstract Distiller

Once relevant publications have been identified and sorted, curators can take advantage of Abstract Distiller to facilitate data extraction. Example input and output of Abstract Distiller are included in **Appendix B**.

- Open Abstract Distiller at: https://distiller.rbsapp.net/.
- Copy pastes the abstract from relevant publications from the DataCrawler output to the “Abstract” field or Abstract Distiller.
- Curators may occasionally need to review the full text of an article and copy the relevant section to the Abstract Distiller if the abstract is too general.
- Define “Focus” and “Additional instructions” to extract relevant information
  a. **Focus** = combination of gene names, tissue names and organisms. Curators construct the “Focus” input by replacing the gene names and tissue terms in the following example: ("Annexin A2" OR "Anxa2" OR "Annexin II" OR Annexin-2) AND (kidney OR renal) AND ("human" OR "mouse" OR "rat")
  b. **Additional instructions** = specific requirements relating to the topic of this section. Curators construct the “Additional instructions” input by replacing the gene name in the following example:
  *Summarize the mechanistic insights regarding the dysregulation of the **Anxa2** gene under various diseased conditions. Summarize the relationship between transcriptomic regulation of **Anxa2** and toxicology. Summarize the signaling pathways, biological processes, or bioactivities affected by **Anxa2** dysregulation in diseased or pathological states. Summarize tissue- or organ-specific mechanism for the regulation of **Anxa2** in pathogenesis. Each summary should be limited to 1 to 3 sentences.*

- Once the “Focus” and “Additional instructions” are defined, click the bottle icon to start generating the summary.
- Check the “Do extra QC” box to further validate the summary. A green check sign indicates that extra QC will be conducted on the statement.
- Curators are responsible for verifying statements generated by Abstract Distiller to ensure alignment with the findings presented in the publication.
- Curators should summarize lengthy statements for clarity and conciseness. It is essential to always include the full name before using any shortened form of a term. Put the full name within a parenthesis.
- For data extraction from additional relevant publications, simply copy pastes abstracts to the “Abstract” field and click the bottle icon again.
- Add the citation to the publication at the end of each summary in the format like [PMID: 35413401] to document the source of the information. When findings from multiple publications are integrated, add the citation in the format like [PMID: XXX, PMID: YYY].
- Copy pastes the final summary to section 9 of the wiki page.
- Include the following statement in section 9 if manual curation does not yield any relevant information on targeted topics: “***No pertinent information pertaining to this subject matter was identified in the existing body of literature.***”

Section 10

Upstream Regulators: what is driving the induction and/or reduction upon stress or pathological conditions.

1) Relevant Literature Search in Data Crawler

- **Data Crawler Search String**
  The literature search procedure follows a structured approach as described in section 9, but utilizing the following search string:
  (Anxa2[Title/Abstract] OR "Annexin A2"[Title/Abstract] OR Annexin-2[Title/Abstract]) AND (mRNA OR "gene expression" OR transcription) AND (pathogenesis OR toxicology OR pathophysiology OR damage OR toxicity OR disease OR disorder OR disfunction OR pathologic) AND (Upregulation OR up-regulation OR downregulation OR down-regulation OR increase OR decrease OR enhancer OR repressor OR repression OR activator OR stimulate) AND (human OR mouse OR rat)
  a. Curators are required to customize the gene name of the example search string to match the specific gene-tissue pair under investigation.
  b. Curators can refine tissue constrains by adding a string like “AND (kidney OR renal)” to above search string if >500 articles detected.
  c. Adjust the parameters of the Data Crawler query based on the number of records retrieved, as detailed in item 2 of section 9 of the SOP.

- **Screening for Relevant Records in the Data Crawler Search Results**

To refine the screening of publications obtained from DataCrawler for relevant records, follow these steps:

a. Conditional Formatting: Select columns containing title and abstract. Apply the "Conditional Formatting" function with "Highlight Cells Rules" set to "Text That Contains". Ensure the selected records mention the gene of interest or its aliases.
b. Review for Gene Expression Change: Review the abstract to verify mention of the target gene's expression change (upregulation or downregulation).
c. Check for Toxic Conditions: Review the abstract to confirm mention of a toxic condition (e.g., disease or dysregulated/pathological condition).
d. Identify Upstream Regulators: Review the abstract to ensure it discusses an upstream regulator for the target gene. Upstream regulators include transcription factors, enhancers, repressors, or coactivators that regulate gene expression under diseased conditions. This may also encompass growth factors, cytokines, hormones, or other molecules directly or indirectly influencing the gene's expression under pathological conditions.
e. Prioritization by Impact Factor: Prioritize and reorder publications meeting the above criteria based on impact factor, from high to low.
f. Priority for Review Articles: Give priority to review articles in the final selection process.
g. Proceed to Information extraction once up to 10 relevant articles have been selected. If no relevant articles are identified in Data Crawler, proceed to search for relevant literature in UniProt as detailed below.

2) Extract Upstream Regulators Information from Relevant Literature

Refer to item 5 of section 9 of the SOP for procedures on how to set up Abstract Distiller to extract information from abstracts.

- Define “Focus” and “Additional instructions” as follows to extract relevant information for section 10.
  a. **Focus** = combination of gene names, tissue names and organisms. Curators construct the “Focus” input by replacing the gene names and tissue terms in the following example: ("Annexin A2" OR "Anxa2" OR "Annexin II" OR Annexin-2) AND (mRNA OR "gene expression" OR transcription)
  b. **Additional instructions** =
  Summarize the upstream regulators of ***Anxa2*** *that induce or reduce gene expression during the onset or progression of various diseases, pathological conditions, or toxic environments. These regulators include transcription factors, enhancers, repressors, or coactivators that modulate* ***Anxa2*** *expression under diseased conditions. They may also encompass growth factors, cytokines, hormones, or other molecules directly or indirectly affecting* ***Anxa2*** *expression in pathological contexts.*

*If available, specify the direction of **Anxa2** gene expression change and specify what signaling pathways were affected. Limit the summary to 1 to 3 sentences.*

- Verify accuracy of the summary and add the corresponding reference (formatted as [PMID: xxx]) to the end.
- Paste the summary to section 10 of the wiki page.
- **Example curation**: The tripartite motif-containing 59 (TRIM59) was found to interact with Annexin A2 and induce Annexin A2 expression in ovarian cancer. TRIM59 promotes malignant progression of ovarian cancer by inducing Annexin A2 expression [PMID: 30585270].

3) Relevant Literature Search in UniProt

- Navigate to the GeneCards database at https://www.genecards.org/ and perform a search for the gene of interest.
- Access the "Proteins" section within the GeneCards entry by clicking on "Proteins" in the "Jump to section" panel.
- Copy the "Protein Accession" code, like "P07355".
- Navigate to the UniProt database at https://www.uniprot.org/, paste the code in the "Find your protein" field. Then, click "Search".
- In the resulting protein annotation window, click "Interaction" on the left panel to find literature indicating interactors of target gene.
- Right click on the PubMed id to open the link in a new tab.
- Review the abstract to check if they meet the following criteria:
  a. Review for Gene Expression Change: Review the abstract to verify mention of the target gene's expression change (upregulation or downregulation).
  b. Check for Toxic Conditions: Review the abstract to confirm mention of a toxic condition (e.g., disease or dysregulated/pathological condition).
  c. Identify Upstream Regulators: Review the abstract to ensure it discusses an upstream regulator for the target gene.
- For abstracts that meet the above criteria, conduct data extraction using Abstract Distiller following steps detailed above (item 2 of section 10 of this SOP).

Manual Curation for Sections with LLM Results

For the following sections of the Biomarker report, the curation process begins by reviewing the summaries generated by the LLM pipeline for relevant data. If the required information is not adequately covered in the LLM-generated data, curators proceed with manual query and extraction to supplement the data.

Section 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 **in the specific tissue** 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.

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 tissue of interest, 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.

1) Screen for Relevant Information in the LLM Results

- Navigate to Rancho project SharePoint site containing the LLM pipeline extracted data: https://ranchobiosciences.sharepoint.com/:f:/s/NIH/EntYnzAFSZFIrNQmVpV9W3ABkGH3KptHQDiv_1Tl67qoHg?e=OQvtNc.
- Open the word document named with the specific gene-tissue pair for section 2.
- Open the original Pubmed article by clicking on the PMID number of each record.
- Verify if the LLM summaries meet the following requirements:
  a. Verify the gene name matches the gene of interest.
  b. Verify the summary describes a study using sample materials derived from the tissue of interest in human or animal models. Evidence from in vitro models was considered secondary unless utilizing primary cells from rodent models. LLM summaries describing *in vitro* cell line-based studies should be skipped.
  c. Verify the change of expression is for transcriptional level instead of protein level. Transcriptomic changes, representing alterations in gene expression levels at RNA/mRNA level assessed through techniques like real-time PCR, RNA-seq, and DNA microarray. All LLM results describing studies demonstrating protein expression changes (as assessed by assays such as ELISA, Western blot, mass spectrometry (MS), or SDS-PAGE) should be skipped.
  d. Verify there is an association of target gene with a pathological condition or an adverse effect. Mutagenicity is not regarded as an adverse effect unless the study concentrates on the model of action (MOA) or association mechanisms.
  e. Verify the summary generated by LLM accurately reflects the findings in the original article.
- For LLM summaries that meet the above requirements, copy the verified records to section 2 of the wiki page.
- Curators should further summarize lengthy statements for clarity and conciseness. It is essential to always include the full name before using any shortened form of a term.
- If the LLM Summaries do not provide information that meets the requirements then curators should start the literature query in Data Crawler.

2) Query Relevant Literature Utilize Data Crawler

- **Data Crawler Search String**

a. The procedures on how to set up and perform queries in Data Crawler are specified in the 1st item of section 9 of this SOP.
b. Construct Data Crawler search string by customizing the gene name element and tissue name element based on the following example for "Trib3 – liver" query. The search string contains four main elements including gene name/ alias, tissue of interest, gene expression term, and organisms.

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

c. If the search yields more than 500 publications, refine the query by applying one or more of the following constraints:
   - Restrict gene names to appear only in the Title or Abstract fields of articles (e.g., add [Title/Abstract] after each gene name or gene alias).
   - Limit results to human studies (e.g., only include "human" in the organism specification part).
   - Include only Review articles (e.g., add "AND review[Filter]" to the end of the query).
   - Narrow down results based on publication date (e.g., query publications from the last 10 years by adding "AND y_10[Filter]" to the end).

d. Conduct screening of the Data Crawler output for relevant publications as per the procedures outlined in item 3 of section 9 of this SOP.
e. Extract data from relevant abstracts using the Abstract Distiller tool.
f. If fewer than three relevant abstracts are identified, perform additional literature queries within the GeneCards or UniProt database.

3) Query Relevant Literature Referred by GeneCards

- Navigate to the GeneCards database at https://www.genecards.org/ and perform a search for the gene of interest.
- Access the "Disorders for xx Gene" section within the GeneCards entry by clicking on "Disorders" in the "Jump to section" panel.
- Find Disorders that may exist in tissue of interest from the MalaCards table (Curators might need to click "See all" to view the full table) and click on the corresponding "PubMed IDs" link (s).
- Review the titles and abstracts of the retrieved publications. Ensure that relevant records align with the LLM summary requirements outlined in the first item of section 2.
- No more than 20 relevant articles will be used for further data extraction according to procedures detailed below (item 5 of this section). If there are still fewer than three relevant abstracts identified, perform additional literature query in UniProt.
- If the literature meets all criteria but utilizes sample materials from tissues other than the tissue of interest, proceed with data extraction using Abstract Distiller as per the

configurations detailed in section 12 of this SOP. Save the extracted information under Section 12 in the wiki page.

4) Query Relevant Literature from UniProt

- Navigate to the GeneCards database at https://www.genecards.org/ and perform a search for the gene of interest.
- Access the "Proteins" section within the GeneCards entry by clicking on "Proteins" in the "Jump to section" panel.
- Copy the "Protein Accession" code, like "P07355".
- Navigate to the UniProt database at https://www.uniprot.org/ and paste the code in the "Find your protein" field. Then, click "Search".
- In the resulting protein annotation window, click "Publications" on the horizontal panel, then, click on "Disease & Variants" on the left panel.
- Review the titles of resulting publications and click on "View abstract" for potentially relevant records. Relevant records should meet the LLM summary requirements discussed in the 1st item of section 2.
- For relevant abstracts, proceed with data extraction facilitated by the Abstract Distiller tool.
- If the literature meets all criteria but utilizes sample materials from tissues other than the tissue of interest, proceed with data extraction using Abstract Distiller as per the configurations detailed in section 12 of this SOP. Save the extracted information under Section 12.
- Include the following statement in section 2 if no relevant article on targeted topic is identified: **"No pertinent information pertaining to this subject matter was identified in the existing body of literature."** This statement ensures transparency regarding the absence of data related to specified themes within the curated literature.
- Information from up to 20 relevant literatures will be used to extract information for section 2 using Abstract Distiller.

5) Extract Association of Gene Transcription with Disease or Toxicity in Abstract Distiller

- Refer to item 6 of section 9 of the SOP for procedures on setting up Abstract Distiller to extract information from abstracts.
- Define "Focus" and "Additional instructions" as follows to ensure extraction of pertinent information for section 2.

a. **Focus**: Curators construct the "Focus" and "Additional instructions" input by replacing the gene names and tissue terms in the following example:

("Trib3" OR "SKIP3" OR "TRB-3" OR "NIPK" OR "SINK" OR "Tribbles Pseudokinase 3") AND ("mRNA" OR "gene expression" OR transcription) AND (liver OR hepatic)

b. **Additional instructions**:

*Summarize the alterations in **Trib3** gene expression observed in diseases, dysfunctions, toxicological, or pathological conditions in **liver** tissues. Describe the biological processes or signaling pathways influenced by these changes in **Trib3** expression and their implications in the onset or advancement of pathological conditions. Limit the summary to 2-3 sentences.*

- Validate the accuracy of the summary and append the respective reference (formatted as [PMID: xxx]).
- Copy the summary into section 2 of the wiki page.
- An example of curation is provided: TRIB3 expression is upregulated in hepatocellular carcinoma (HCC) tissue samples compared to normal liver tissues. This upregulation correlates significantly with larger tumor sizes and poor prognosis in postoperative patients [PMID: 32716348].

Section 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.

1) Screening for Relevant Information in the LLM Results

- Navigate to Rancho project SharePoint site containing the LLM pipeline extracted data: https://ranchobiosciences.sharepoint.com/:f:/s/NIH/EntYnzAFSZFlrNQmVpV9W3ABkGH3KptHQDiv_1Tl67qoHg?e=OQvtNc
- Open the word document named with the specific gene-tissue pair for **section 3**.
- Navigate to the GeneCards database at https://www.genecards.org/ and perform a search for the gene of interest.
- Access the "Proteins" section within the GeneCards entry by clicking on “Proteins” in the “Jump to section” panel.
- Verify the following protein attributes for gene of interest from the LLM summary match protein details from the GeneCards database.
  a. Protein Accession
  b. Size
  c. Molecular mass
- Access the "Domains & Families" section within the GeneCards entry by clicking on “Domains” in the “Jump to section” panel.
- Verify the following protein attributes for gene of interest from the LLM summary match protein details from the GeneCards database.
  a. Domains: corresponding to *“Protein Domains for XXX Gene” InterPro* values.
  b. Blocks: If a protein does not have blocks information, remove this item from the template.
  c. Family

- The LLM results for section 3 may also include summaries of protein structure in relation to protein functions sourced from publications. Curators must ensure the following requirements are met by clicking on the PMIDs to refer to the original publications:
    a. The protein corresponds to the gene of interest, with explicit mention of gene names in the publication.
    b. Information on protein structure, encompassing but not limited to protein domains, functional elements, residues, subunits, modifications, and protein complexes, linked to functionality under normal or pathological conditions or bioprocesses.
- Copy all verified information from the LLM results to section 3 of the wiki page. Don't change the order of the protein attributes in the curation template.

2) Query Relevant Literature by Referring to Uniprot Database

- If the LLM results don't contain summaries sourced from publications. Curators should search for relevant literature by referring to Uniprot first. Access protein annotation page for the gene of interest based on the procedures outlined in item 4 of section 2.
- Under the "Function" section, click on "Publications". Then, click on "view abstract" to review the abstracts from articles. Abstracts that align with the requirements specified in item 1, should be coped to Abstract Distiller for further data extraction.
- If relevant articles cannot be detected in both the LLM results and by referring to the Uniprot database, then, curators should query relevant literature in the Data Crawler.

3) Query Relevant Literature in Data Crawler

- The procedures on how to set up and perform queries in Data Crawler are specified in the 1st item of section 9 of this SOP.
- Construct Data Crawler search string by customizing the gene name element (underlined) based on the following example:
(Anxa2[Title/Abstract] OR "Annexin A2"[Title/Abstract] OR Annexin-2[Title/Abstract]) AND (protein) AND ("crystal structure" OR domain OR subunit OR residue OR complex OR motif OR function) AND (human OR mouse OR rat) AND review[Filter]
- If more than 100 publications are found, curators should narrow down results based on publication date (e.g., query publications from the last 5 years by adding "AND y_5[Filter]" to the end).
- Conditional Formatting: Select columns containing title and abstract. Apply the "Conditional Formatting" function with "Highlight Cells Rules" set to "Text That Contains". Ensure the selected records mention the gene of interest or its aliases.
- Reorder publications based on impact factor from high to low.
- Screening for relevant records, where the abstract contains information about the protein structure and functions. For relevant literature, proceed with data extraction.
- Information from up to 10 relevant articles will be used to extract information for section 3 using Abstract Distiller.

4) **Extract Association of Protein Structure with Function in Abstract Distiller**

- Refer to item 5 of section 9 of the SOP for procedures on setting up Abstract Distiller to extract information from abstracts.
- Define "Focus" and "Additional instructions" as follows to ensure extraction of pertinent information for section 3. Curators should replace the gene names (underlined) based on the gene of interest.
  a. **Focus:**
  (Anxa2 OR "Annexin A2" OR Annexin-2) AND (protein structure) AND (function)

  b. **Additional instructions:**
  Summarize the structural characteristics or features of the protein or protein family. Concisely outline the protein's role in fundamental biological processes. Extract evidence linking protein structure to its functional roles. Limit the summary to 1-2 sentences.
- Validate the accuracy of the summary and append the respective reference (formatted as [PMID: xxx]).
- Copy the summary into section 3 of the wiki page.
- An example of curation is provided: Annexin A2 is a Ca2+ regulated protein existing as a monomer or a heterotetramer with the S100A10 dimer, playing a crucial role in virus life cycles and mediating pathogen adhesion and internalization. Its structural capability to bind with pathogens and facilitate their entry into host cells [PMID: 37482693].

Section 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.

- Procedures for data curation in this section follow the guidelines outlined in section 2, with the following modifications:
  a. Update the tissue constraints to exclude the tissue of interest from the search string and data screening requirements in Data Crawler.
  When defining the search string, use the following example for the like "Trib3-liver" pair:
  ("Trib3"[Title/Abstract] OR "SKIP3"[Title/Abstract] OR "TRB-3"[Title/Abstract] OR "Tribbles Pseudokinase 3"[Title/Abstract]) AND ("mRNA" OR "gene expression" OR transcription) AND (human OR mouse OR rat) **NOT (liver OR hepatic)**
  b. Exclude the tissue constraint when specifying the "Focus" in Abstract Distiller.
  When extracting information from literature using Abstract Distiller, define the Focus and Additional requirements as follows:
    - **Focus:**

Curators construct the “Focus” and “Additional instructions” input by replacing the gene names in the following example:
("Trib3" OR "SKIP3" OR "TRB-3" OR "Tribbles Pseudokinase 3") AND ("mRNA" OR "gene expression" OR transcription)

- **Additional instructions:**
  *Summarize the alterations in **Trib3** gene expression observed in diseases, dysfunctions, toxicological, or pathological conditions across different tissues. Describe the biological processes or signaling pathways influenced by these changes in **Trib3** expression and their implications in the onset or advancement of pathological conditions. Limit the summary to 2-3 sentences.*

c. This section may encompass biomarker-disease associations identified based on changes in protein expression, mutations, or other genetic variations.
d. Information from up to 20 literature will be used to extract relevant information for section 12.

VII. Quality Assurance

To ensure the reliability and accuracy of data extraction, Rancho has implemented robust quality control measures spanning both manual and automated curation sections. Each biomarker report undergoes initial manual checking by the primary curator responsible for its creation. Subsequently, an independent curator, not involved in the initial report generation, conducts a thorough review. Curators refer to the original data sources to verify the relevance, accuracy, completeness, and correct formatting of extracted data. QC comments are provided to the original curator for necessary revisions whenever discrepancies are identified during the QC process. In cases where discrepancies persist or are complex, the original curator seeks guidance from a Senior Curator. The Senior Curator conducts further validation using original data or alternative sources to resolve ambiguities.

Self-QC Procedures

Curators will perform self-quality control after completion of data extraction related to a specific gene/tissue pair in both the manual curated sections (1, 6, 9, 10) and the LLM-pipeline facilitated sections (2, 3, 12). The procedures are outlined below:

1) Ensure that Sections 1, 2, 3, 6, 9, 10, and 12 contain content relevant to their respective topics.
2) Verify that all statements include citations to corresponding publications, formatted as [PMID: xxxxx].
3) Include a common default statement in Sections that lack relevant information.
4) Check for and remove extra spaces or unnecessary punctuation marks.

5) Ensure there are no special characters that could affect readability or formatting. Replace Greek symbols with its fully spelled name. For example, “α” to “alpha”, “β” to “beta”.
6) Maintain consistency in formatting throughout the document.
7) Proofread to eliminate typographical errors.
8) Spot-check 10% of citations to confirm they accurately support curated data.
9) Spot-check at least one statement under each section to confirm they accurately match the findings in the original literature.
10) Ensure references are limited to human or animal in vivo studies only (excluding Section 9: Mechanistic Information); identify and remove references to terms such as "cell line" or "in vitro."
11) For every statement ensure curate data relates to the gene of interest meaning the gene names in the extracted data should contain the gene of interest or one of its synonyms.
12) For Section 2, confirm that only studies for tissue of interest are included.
13) In Sections 9 and 12, ensure that tissue information is either explicitly provided or can be inferred for each curated statement.
14) For Section 12, confirm that no studies from tissue of interest are included.
15) Curators should review their own work after data extraction. Once self-QC is done, Click the “PAGE” tag icon on the upper right corner of the wiki page, and set the CATEGORIZATION tag to “to QC”.

Cross-QC Procedures

A Senior Curator will designate some of the gene/tissue pairs for cross-QC. The procedures are outlined below:

1) Based on the gene-tissue pair, navigate to the corresponding tissue folder within the wiki system. Click on the gene name to open the wiki page for the gene of interest.
2) Conduct QC checks following the guidelines outlined for self-QC.
3) Enter QC comments starting with curators’ initials in the “Comments” section at the bottom of the wiki page. Comments should specify detected issues and reference the relevant section of the report. Curators may also suggest pertinent literature or data sources in their comments.
4) Once QC is complete, click on the “Post Comment” button below the Comments section.
5) Click on the pencil icon at the bottom right corner of the wiki page to enter edit mode. Fix minor issues such as typos or extra spaces identified during QC.
6) Click on the “PAGE” tag icon at the top right corner of the wiki page. Change the CATEGORIZATION tag to “to QC Revision”.

Spot Checks by Senior Curators

To enhance consistency and ensure data integrity, a number of gene reports will be checked by a Senior Curator.

1) Senior Curators will thoroughly review QC comments from Cross-QC curators. In cases where conflicting opinions arise between the primary curator and Cross-QC curator, Senior Curators will refer to the original publications to verify that extracted data aligns accurately with the source data. If ambiguity persists in the source data, Senior Curators will consult additional literature or conduct targeted searches to verify or enrich the extracted information.
2) Senior Curators will open multiple wiki pages curated by different curators in parallel. This comparison aims to ensure consistency in the input format of each section. In cases where inconsistencies are identified, the original curator will be notified to revisit their previous curations and make necessary corrections.
3) Senior Curators will select a sample of gene/tissue pairs that have not undergone cross-QC and conduct QC checks. These checks will follow the guidelines outlined for self-QC, ensuring thorough verification of data accuracy and adherence to established standards.

These procedures are designed to maintain high standards of quality control, promoting uniformity and reliability across all biomarker reports. Rancho's commitment to delivering top-tier data curation services is underscored by rigorous validation and QC processes, combining advanced technology with expert oversight to ensure data reliability.

Appendix A: Example Gene Summary Output

Example Gene: Havcr1

Gene Aliases: Kim-1, Tim1, CD365

Association with Toxicity and/or Disease at a Transcriptional Level: Kim-1 is transcriptionally upregulated upon AKI in several species. Corresponding increases in Kim-1 protein are observed in the urine due to proteolytic cleavage of the extracellular domain [*citations*].

Summary of Protein Family and Structure: Describe the protein structure and the family of proteins Havcr1 belongs too. Describe the protein domains and their generally recognized role in protein function. Provide citations.

Proteins Known to Interact with Gene Product: Identify and describe proteins known to interact with Havcr1. Provide citations.

Links to Gene Databases:

- Gene cards (human): https://www.genecards.org/cgi-bin/carddisp.pl?gene=HAVCR1&keywords=HAVCR1
- Harmanizome (human): https://maayanlab.cloud/Harmonizome/gene/HAVCR1
- NCBI (human): https://rgd.mcw.edu/rgdweb/report/gene/main.html?id=708425
- NCBI (rat): https://www.ncbi.nlm.nih.gov/gene/286934
- Ensemble (human): https://useast.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000113249;r=5:157026742-157069396
- Ensemble (rat): https://useast.ensembl.org/Rattus_norvegicus/Gene/Summary?db=core;g=ENSRNOG00000007243;r=10:31119088-31151698
- Rat Genome Database: https://rgd.mcw.edu/rgdweb/report/gene/main.html?id=708425
- Uniprot (human): https://www.uniprot.org/uniprotkb/Q96D42/entry
- Uniprot (rat): https://www.uniprot.org/uniprotkb/O54947/entry
- Wikigenes (human): https://www.wikigenes.org/e/gene/e/26762.html
- Wikigenes (rat): https://www.wikigenes.org/e/gene/e/286934.html
- PDB (human): https://www.rcsb.org/structure/5DZO
- PDB (rat):
- Alphafold (human): https://alphafold.ebi.ac.uk/entry/E9PFX0
- Alphafold (rat): https://alphafold.ebi.ac.uk/entry/G3V6W3

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

- SARS-CoV-1 infection: *[Description of pathway with citations]*
- SARS-CoV-2 infection: *[Description of pathway with citations]*
- NF-kappaB signaling: *The canonical NF-KB pathway is induced by the ligand-dependent activation of a variety of receptors. Stimulus-dependent activation of the IKK complex results in the phosphorylation and subsequent proteasomal degradation of IκBα and IκBε (NFKBIA and NFKBIE). This allows the nuclear translocation of transcriptionally active RelA/p50 (RELA, NFKB1) or c-Rel/p50 heterodimers (REL, NFKB1). [citation/link]*
- Ebola virus infection in host: [*Description of pathway with citations]*

Gene Descriptions:

- *Entrez Gene Summary for HAVCR1 Gene:* The protein encoded by this gene is a membrane receptor for both human hepatitis A virus (HHAV) and TIMD4. The encoded protein may be involved in the moderation of asthma and allergic diseases. The reference genome represents an allele that retains a MTTVP amino acid segment that confers protection against atopy in HHAV seropositive individuals. The protein is a receptor for multiple other viruses, including Ebola virus,

Marburg virus, Dengue virus, and Zika virus and is a possible entry factor for SARS-CoV-2 and other coronaviruses. [provided by RefSeq, Sep 2021]

- *GeneCards Summary for HAVCR1 Gene:* HAVCR1 (Hepatitis A Virus Cellular Receptor 1) is a Protein Coding gene. Diseases associated with HAVCR1 include Hepatitis and Hepatitis A. Among its related pathways are Disease and SARS-CoV-1 Infection. Gene Ontology (GO) annotations related to this gene include virus receptor activity. An important paralog of this gene is TIMD4.
- *UniProtKB/Swiss-Prot Summary for HAVCR1 Gene:* Phosphatidylserine receptor that plays an important functional role in regulatory B-cells homeostasis including generation, expansion and suppressor functions (By similarity). As P-selectin/SELPLG ligand, plays a specialized role in activated but not naive T-cell trafficking during inflammatory responses (PubMed:24703780). Controls thereby T-cell accumulation in the inflamed central nervous system (CNS) and the induction of autoimmune disease (PubMed:24703780). Also regulates expression of various anti-inflammatory cytokines and co-inhibitory ligands including IL10 (By similarity). Acts as regulator of T-cell proliferation (By similarity). May play a role in kidney injury and repair (PubMed:17471468). (HAVR1_HUMAN,Q96D42) (Microbial infection) Acts as a receptor for Hepatitis A virus. (HAVR1_HUMAN,Q96D42) (Microbial infection) Acts as a receptor for Ebolavirus and Marburg virus by binding exposed phosphatidyl-serine at the surface of virion membrane (PubMed:21536871). Serves as a dual receptor for Ebolavirus by also interacting with envelope glycoprotein GP (PubMed:26487564). (HAVR1_HUMAN,Q96D42) (Microbial infection) Acts as a receptor for Dengue virus by binding exposed phosphatidyl-serine at the surface of virion membrane (PubMed:23084921). TIM1 and Dengue virus are co-internalized during virus entry (PubMed:29742433). (HAVR1_HUMAN,Q96D42)(Microbial infection) Acts as a receptor for Zika virus by binding to envelope protein E. (HAVR1_HUMAN,Q96D42)(Microbial infection) Plays a positive role in Chikungunya virus cell entry. (HAVR1_HUMAN,Q96D42)

Cellular Location of Gene Product: Havcr1 is primarily found in the cell membrane but can also be extracellular.

Mechanistic Information (Why Does Expression Change with Toxicity and/or Disease):

- The proposed mechanism of response is that acute renal damage initiates ERK1/2 and signal transducer and activator of transcription 3 (STAT3) phosphorylation. Then, nuclear STAT3 binds to the KIM-1 promoter and increases its mRNA and protein levels [*citations*]
- Acute overexpression of KIM-1 in proximal renal tubular epithelial cells after ischemia, hypoxia, and toxicity promotes transformation of the cells into “semi-professional” phagocytic cells, with the help of KIM-1’s mucin domain. KIM-1 is a phosphatidylserine receptor on the surface of the liposome that can identify the apoptosis body and phosphatidylserine, which mediates further phagocytosis [*citation*].
- KIM-1 plays a role in the removal of apoptotic cells and necrotic tissue fragments. Furthermore, KIM-1 phosphorylation, and its interaction with p85, enhance cell autophagy to degrade KIM-1 phagosomes relying on Unc-51 like autophagy activating kinase 1 (ULK1, also known as ATG1) phosphorylation and maintain self-tolerance by the presentation of antigens on the proximal tubule cell [*citation*].
- Experiments have shown that cells expressing WT KIM-1, but not a variant lacking the mucin domain (KIM-1 Δmucin), exhibit decreased NF-κB activity and inflammatory cytokine production. These findings illustrate KIM-1's role in reducing inflammation and acute injury in the kidney by mediating phagocytosis and downmodulating the inflammatory response [*citation*].

Upstream Regulators: Search transcriptional regulatory and cell signaling databases and describe the upstream regulators and signaling pathways that control expression of HAVCR1. Provide citations.

Tissues/Cell Type Where Genes are Overexpressed: Use GTEX and single cell expression atlases to identify tissues and cell types where the HAVCR1 is expressed the highest. Provide citations.

Role of Gene in Other Tissues: Describe the role of the HAVCR1 in tissue homeostasis outside of the tissue where is it a biomarker of toxicity. Provide citations.

Chemicals Known to Elicit Transcriptional Response of Biomarker in Tissue of Interest: Search resources such as rat genome database, comparative toxicogenomic database and other resources that link chemical exposure to changes in expression of the biomarker in the tissue of interest. Provide citations.

DisGeNet Biomarker Associations to Disease in Organ of Interest: Search each biomarker gene and use association type "Altered Expression" to find data supporting the associations of altered expression with target organ disease. Provide citations.

Summary (LLM): Havcr1, also known as Kidney Injury Molecule-1 (KIM-1), plays a crucial role in maintaining kidney homeostasis, particularly in the context of renal injury and repair mechanisms. In healthy kidneys, Havcr1 expression is minimal, but its upregulation following acute kidney injury (AKI) signifies its importance in tissue repair and regeneration processes. Havcr1 facilitates the clearance of apoptotic cells and debris, thereby mitigating further inflammatory responses and promoting tissue repair (CS=8).

After injury, Havcr1 expression predominantly occurs in the proximal tubule epithelial cells, where it modulates cell adhesion, migration, and differentiation—key processes in renal tubule repair. By promoting the phagocytosis of apoptotic cells, Havcr1 aids in resolving inflammation and restoring tubular integrity, which is vital for the kidney's filtering function (CS=7).

Besides tissue repair, Havcr1 influences the renal immune environment. Its role extends to modulating the immune response, potentially reducing the risk of chronic kidney disease (CKD) following acute injury by limiting ongoing inflammation and fibrosis. This immunomodulatory function is crucial, as unchecked inflammation can exacerbate kidney damage and accelerate the progression to CKD (CS=7).

The unique upregulation pattern of Havcr1 in response to kidney injury has made it a prominent biomarker for detecting and monitoring AKI. Its presence in urine correlates with the extent of tubular injury, offering a non-invasive means to assess kidney health and the effectiveness of therapeutic interventions aimed at mitigating renal damage and promoting recovery (CS=9).

Increased expression of Havcr1 in kidney diseases is primarily driven by inflammation, oxidative stress, and lipid metabolism dysregulation. Factors like TGF-β and Ang II upregulate Havcr1, highlighting its role in CKD pathogenesis. Oxidative stress and inflammation, exacerbated by these conditions, lead to further renal damage. The involvement of signaling pathways such as NF-κB, which regulates inflammation, and the impact of lipotoxicity on renal cells underscore Havcr1's significance in kidney injury responses (CS=8).

The role of Havcr1 in the adaptation and resolution of toxicity or disease within the kidney primarily involves its functions in promoting tissue repair and mitigating inflammation following renal injury. Havcr1, upregulated in response to acute kidney injury, facilitates the clearance of dead cells and debris, thereby reducing the inflammatory response and promoting repair. Its expression in proximal tubule cells after injury supports cell proliferation and differentiation, essential for restoring kidney function. Moreover, Havcr1's interaction with immune cells helps modulate the immune response, potentially preventing chronic kidney disease progression by limiting fibrosis and further damage. These functions underscore Havcr1's critical role in kidney recovery processes, suggesting potential therapeutic targets for enhancing kidney repair and function after injury. (CS=8).

Appendix B: Technical specifications and example Input and Output of Abstract Distiller

Example Input

Abstract (PMID: 25284003)

Objective: Renal cell carcinoma (RCC) is the most lethal genitourinary cancer and intrinsically resistant to chemotherapy, radiotherapy, and hormone therapy. Annexin A2 (Anxa2) is a calcium-dependent phospholipid-binding protein found on various cell types that plays multiple roles in regulating cellular functions. In RCC, Anxa2 expression was correlated with tumor differentiation, clinical outcomes, and the metastatic potential; however, the underlying mechanisms remain obscure. This study investigated the role of Anxa2 in regulating tumorigenesis of RCC.

Materials and methods: Commercial RCC tissue microarray arrays and a kidney cancer quantitative polymerase chain reaction array were used to examine Anxa2 by immunohistochemistry and real-time polymerase chain reaction analysis. Short hairpin (sh)RNA-based lentiviral system technology was used to evaluate the effects of manipulating Anxa2 expression on multiple malignant features of 2 RCC cell lines, A498 and 786-O, and its mechanisms.

Results: (1) The Anxa2 expression level was generally elevated to varying degrees in RCC tissues. In adjacent noncancerous tissues, Anxa2 was mainly expressed in glomeruli and slightly expressed in the cytoplasm of proximal tubules. (2) An increased Anxa2 expression level was found in tissues of clear cell RCC, papillary RCC, and chromophobe RCC, and it was prominently expressed in cancer cell membranes. In addition, the Anxa2 expression level was correlated with poor prognosis. (3) Silencing Anxa2 expression suppressed the abilities of cell migration and invasion, but cell proliferation was less affected. (4) Diminished Anxa2 expression caused alterations in the cell polarity, disrupted the formation of actin filaments, and reduced CXCR4 expression. (5) Inhibition of the Rho/Rock axis restored silencing of Anxa2-mediated suppression of cell motility.

Conclusions: Overall, our study points out the regulatory function of Anxa2 in RCC cell motility and provides a molecular-based mechanism of Anxa2 positivity in the progression of RCC.

Focus

("Annexin A2" OR "Anxa2" OR "Annexin II" OR Annexin-2) AND (kidney OR renal) AND ("human" OR "mouse" OR "rat")

Additional instructions

Summarize the mechanistic insights regarding the dysregulation of the Anxa2 gene under various diseased conditions. Summarize the relationship between transcriptomic regulation of Anxa2 and toxicology. Summarize the signaling pathways, biological processes, or bioactivities affected by Anxa2 dysregulation in diseased or pathological states. Summarize tissue- or organ-specific mechanism for the regulation of Anxa2 in pathogenesis. Each summary should be limited to 1 to 3 sentences.

Example Output Generated by Distiller

Anxa2 expression is elevated in renal cell carcinoma (RCC) tissues and correlates with tumor differentiation and poor prognosis. Dysregulation leads to increased cell migration and invasion but not proliferation, influencing the Rho/Rock signaling pathway and affecting cell polarity and actin filament formation. Anxa2 mainly expressed in cancer cell membranes, with lower levels in adjacent noncancerous tissues, indicates its specific regulatory role in RCC pathogenesis.

Technical specifications for Abstract Distiller:

Version 1.0 (883613b)

LLM model: gpt-4o

System prompt:

> Your job is to summarize abstract, distilling the most relevant information. User will provide the abstract and what to focus on. You read abstract, and prepare very short (2-3 sentences) distillation. You report only what was done and what the main findings are, on the focus item.
>
> The focus could be: specific gene, gene upregulation, tissue, disease and so on.
>
> Rules:
>
> - You NEVER say introductory words such as "This study is about", "The abstract is about" and so on. You just report what was done
>
> - You keep experimental detail (e.g. instead of "transgenic mice" you provide specific strain & other info that is in the abstract)
>
> - You NEVER add conclusion (e.g. this suggests blah, this shows that blah etc)

User prompt:

> Focus: {focus}
>
> {prompt_instructions}
>
> {abstract}

Prompt instructions are empty or the following text block:

> IMPORTANT! ADDITIONAL INSTRUCTIONS: {instructions}\n\n

Where {instructions} are user-provided additional instructions.