ExpressionGenesis: Automated Disease Annotation and Metadata Generation for the Gene Expression Omnibus Using Large Language Models

Daniel R. Spohn, M.S.
Graduate Program in Bioinformatics, Brandeis University, Waltham, MA, USA
Corresponding Author: dspohn@gmail.com

Abstract

Public repositories, such as the Gene Expression Omnibus (GEO), host hundreds of thousands of transcriptomic datasets. However, inconsistent and unstructured metadata limit their reuse and integration. We developed ExpressionGenesis, an automated platform that generates standardized, ontology-linked metadata for GEO Series (GSE) records using large language models (LLMs) and retrieval-augmented generation (RAG). ExpressionGenesis extracts structured information, including disease names, experimental design, study summaries, and keywords, from free-text GEO metadata and validates disease annotations using the Disease Ontology (DO). The system is implemented on Amazon Web Services (AWS) with a fully serverless architecture combining Lambda, Step Functions, and Athena for data processing, and a Next.js web interface on AWS Amplify for interactive exploration. Evaluation of 200 GEO Series demonstrates that ExpressionGenesis achieves higher accuracy and F1-scores than previous NLP-based annotation methods. The public web application (https://expressiongenesis.com) provides a searchable interface for enriched GEO metadata and submission trends, and a downloadable CSV of disease annotations for all indexed GEO Series. These results show that LLMs, combined with ontology-grounded validation, can generate accurate, standardized disease annotations for GEO Series, thereby supporting improved accessibility and reusability of publicly available gene expression data.

Keywords: Gene Expression Omnibus; metadata curation; disease annotation; large language models; retrieval-augmented generation; Disease Ontology.

Introduction

High-throughput gene expression technologies have generated a large volume of data, creating a significant opportunity for meta-analysis, reproducibility studies, and data reuse. The Gene Expression Omnibus (GEO) database is one of the largest public repositories (Edgar et al., 2002). As of December 2025, ExpressionGenesis has indexed over 268,000 publicly available GEO Series entries (Figure 1). A GEO Series is a researcher-submitted record that provides a summary of a gene expression study and links together samples. GEO serves as a vital

resource for researchers to discover existing gene expression datasets for further analysis and investigation. Reusing these datasets not only accelerates scientific discovery but also maximizes the value of publicly funded research.

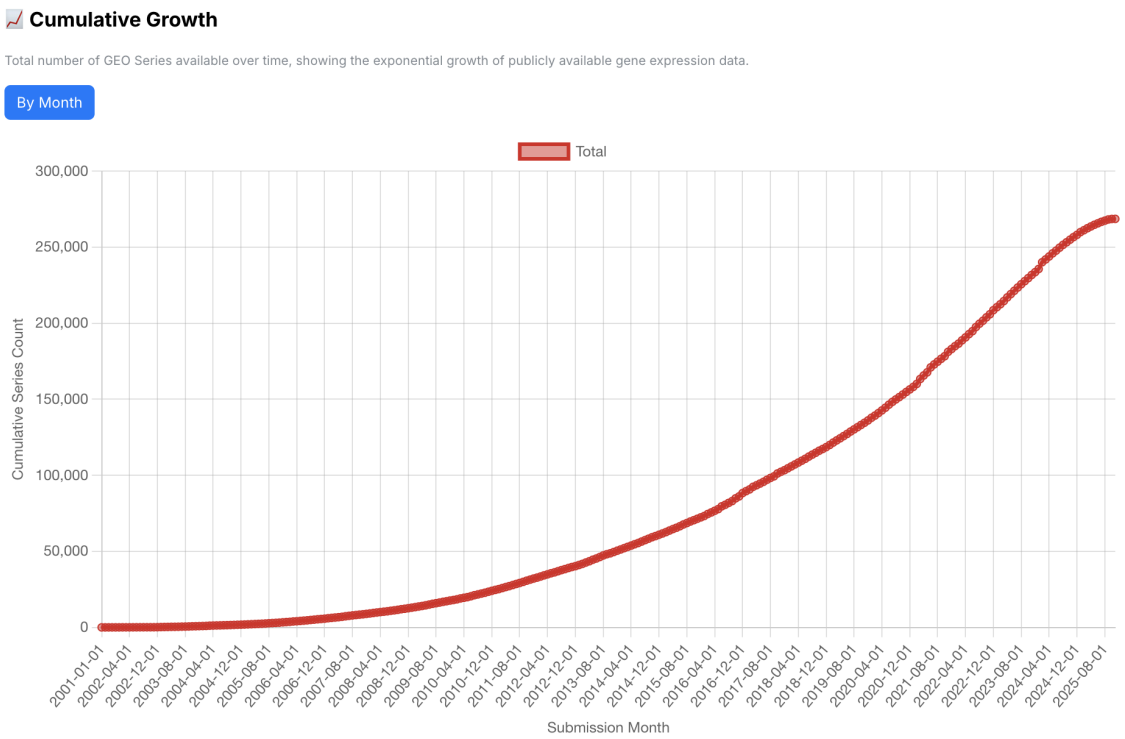

Figure 1. Cumulative number of GEO Series (GSE) submissions as of December 2025.

GEO presents usability challenges because much of its metadata is unstructured, inconsistently populated, and stored as free-text. These inconsistencies can hinder researchers from identifying relevant datasets, leading to an underutilization of GEO's full potential.

Tools and methods that enable researchers to discover and interpret GEO datasets can increase the likelihood of dataset reuse. Several prior tools have aimed to improve GEO dataset discovery. Projects such as GEOMetaCuration (Li et al., 2018), ALE (Giles et al., 2017), CREEDS (Wang et al., 2016), and ReGEO (Chen et al., 2019) have been created to support the discovery of existing GEO datasets and to enrich GEO metadata through manual or automated measures. However, many of these tools are no longer maintained or accessible. For instance, the GEOMetaCuration website is no longer accessible, and ReGEO was last updated in 2018.

Manual curation, while valuable, is not scalable given the vast number of GEO Series entries. For example, the CREEDS project by Wang et al. used a crowdsourcing effort to annotate a few thousand GEO Series (Wang et al., 2016). Automated approaches for generating structured metadata offer an approachable alternative to manually annotating all GEO Series. This annotation can improve the searchability and interpretability of GEO entries. In particular, structured metadata points such as disease names and experimental design descriptions can significantly enhance dataset discoverability.

A challenge in identifying similar transcriptomic datasets is the differing terminology used by researchers. The same medical condition can be described as “non-alcoholic fatty liver disease”, “metabolic dysfunction-associated steatotic liver disease”, “NAFLD”, and “MASLD”, along with many other variations in the researcher-supplied GEO Series descriptions. These variations in terminology can make it difficult to identify all relevant GEO entries of a particular condition. Diseases and conditions are related and can be classified into broader groups. The Disease Ontology (DO) project provides standardized disease names and identifiers (DOIDs) along with hierarchical relationships among diseases (Schriml et al., 2022). Mapping the GEO Series to DOIDs enables clearer identification of related experiments and supports more accurate searches.

ExpressionGenesis is a web-based tool that automates the generation of structured metadata from GEO Series using large language models (LLMs). ExpressionGenesis supports dataset discovery by extracting standardized metadata, including disease annotations, summaries, experimental design, and keywords, making them available within a web user interface. Compared to prior tools, ExpressionGenesis provides a scalable, automated, and continuously updating solution for enriching GEO metadata.

Methods

Implementation Overview

For scalability, ExpressionGenesis was built and deployed on Amazon Web Services (AWS). All processing runs on a serverless infrastructure to reduce maintenance. The processing code is written in Python and executes in AWS Lambda and AWS Step Functions, eliminating the need for server management. Data integration is performed using AWS Athena and SQL. The web application is a Next.js application that runs on AWS Amplify with data served by DynamoDB. Large language models are accessed using AWS Bedrock. Production metadata generation in ExpressionGenesis uses the Meta Llama 3.3 70B instruct model via Bedrock (model identifier: us.meta.llama3-3-70b-instruct-v1:0).

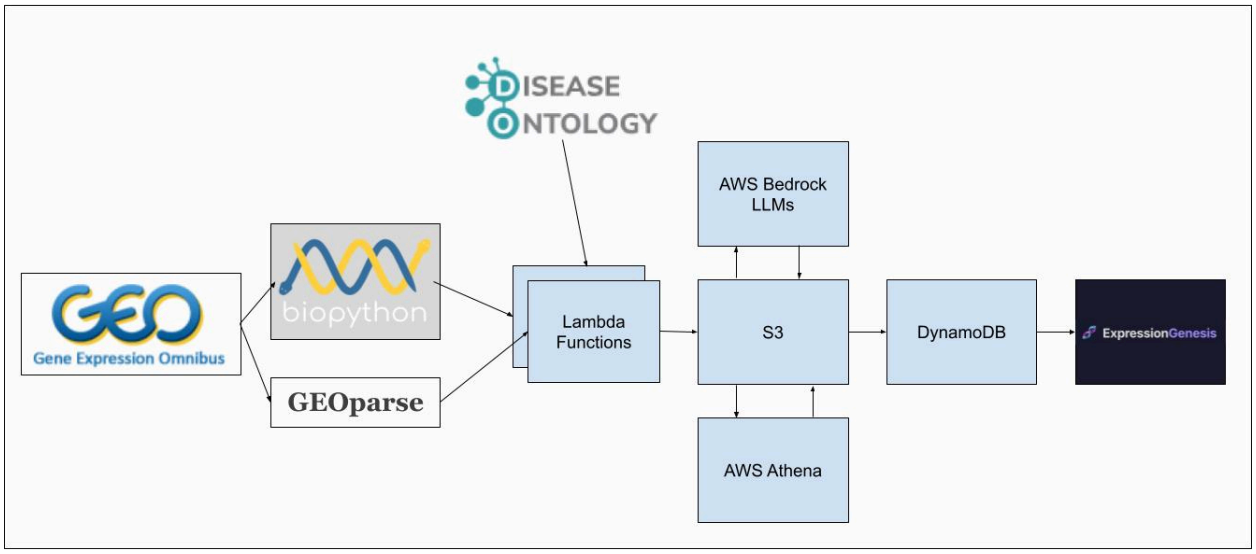

Figure 2. Technical overview of ExpressionGenesis

Data Sources

Gene Expression Omnibus (GEO)

ExpressionGenesis processes GEO Series (GSE) and sample (GSM) data. This data is retrieved using Biopython (Cock et al., 2009) and GEOparse. Biopython accesses the NCBI Entrez API to obtain a list of publicly available GSE entries. GEOparse downloads and parses the associated SOFT files from the NCBI SFTP server to extract detailed metadata for individual GEO Series.

Each GSE entry includes information such as:

- GSE ID
- Title
- Submission and publication dates
- Summary (free text)
- Overall Experimental Design (free text)
- Associated GSM sample metadata
- Associated PubMed article IDs

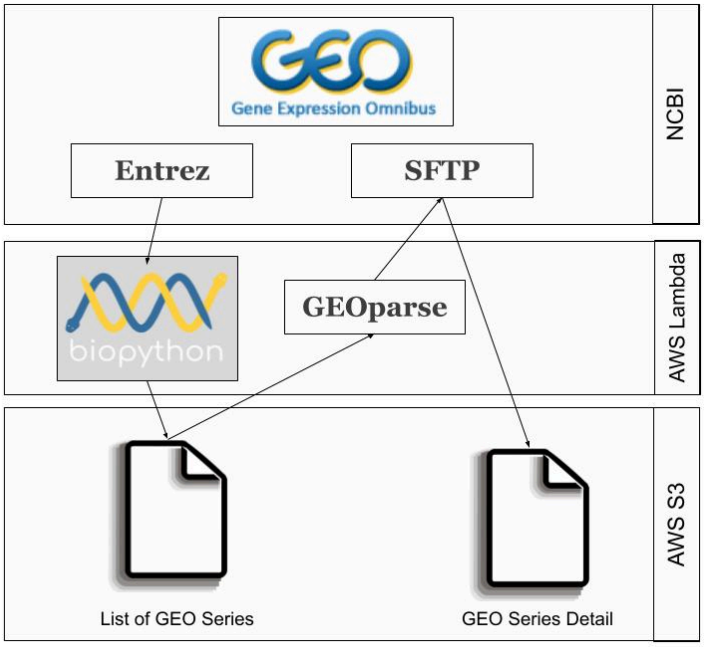

Figure 3. Biopython and GEOparse are used to extract GEO Series data from NCBI

Disease Ontology

The Disease Ontology (DO) database offers standardized human disease terms, unique identifiers (DOIDs), and a hierarchical structure that categorizes diseases. DO terms enable consistent annotation of disease information.

A Python process was written to download the human DO data from the Disease Ontology GitHub repository. The December 2024 release of the Disease Ontology was used for the initial version of ExpressionGenesis, as it was the latest version available at the time of the application's development. New versions of DO will be incorporated in future ExpressionGenesis updates.

Metadata Generation

ExpressionGenesis uses LLMs to generate structured metadata from the free-text fields of GEO entries. These foundational models are accessed via AWS Bedrock and are prompted to generate standardized summaries, disease annotations, experimental design details, and keyword lists in a JSON format.

A standardized prompt is used to ensure consistency across LLM output responses. The prompt specifies the required fields and includes formatting instructions to help ensure a consistent JSON output. The prompt requests:

- Summary of the experiment
- Keywords
- Experimental design (study type, groups, sample size, comparisons)
- Disease information (name, DOID, stage)

Additionally, the prompt requests that fields that cannot be determined should be omitted from the output.

```
You are a bioinformatics metadata expert.
Your task is to analyze the provided GEO (Gene Expression Omnibus) experiment series and extract structured and standardized
metadata fields.

You will provide this structured metadata in a single JSON object.

Return only the JSON object without additional explanation. If a field cannot be determined from the provided series
information, omit it from the JSON rather than including null or placeholder values.

Format: Pure JSON output
Rules:
- Only output valid JSON
- No explanatory text
- No markdown formatting
- No code blocks or backtick

Format dates as YYYY-MM-DD and standardize units to SI format.

For each GEO Series experiment provided, return a JSON object with the following fields:

- metadata:
  - GSE_ID: static value of {gse_id}
  - date_of_entry: set to date of the GEO entry
  - current_date: static value of {current_date}
  - model_name: static value of {model_name}
- summary: 7-12 sentence summary of the experiment that covers:
  - The main goal/purpose
  - Key experimental conditions
  - The type of analysis performed
  - Any notable genes or pathways studied
  - The technology used to generate the dataset (RNA-seq, DNA-seq, ChIP-seq, ATAC-seq, etc.)
  - Mention the species and disease/phenotype studied (if applicable)
- keywords: provide a list (array) of keywords
- experimental_design:
  - type: Categorize as one of [Case-Control, Time Series, Treatment-Control, Longitudinal, Cross-sectional, Dose-response]
  - groups: Array of group descriptions
  - treatment_groups
  - control_groups
  - experimental conditions
  - sample_size: Object with group sizes if specified
  - comparison: String describing the specific comparison
- disease: Array of objects for diseases mentioned in the description (if applicable):
  - name: Standardized human disease name - only if the text explicitly describes or strongly implies a human disease condition
affecting the organism under study.
  - do_ontology_id: Human Disease Ontology ID (DOID) related to name
  - stage: Disease stage/progression if specified
```

Figure 4. Prompt template used in ExpressionGenesis

The prompt template is then followed by GEO Series metadata for an individual GEO entry so that the LLM can produce the requested output.

title: Spatio-temporal interaction of immune and renal cells determines glomerular crescent formation in autoimmune kidney disease
gse_id: GSE294965
publication_date: 2025-05-30
submission_date: 2025-04-17
summary: Rapidly progressive glomerulonephritis (RPGN) is the most aggressive group of autoimmune kidney disease with the worst prognosis. Anti-neutrophil cytoplasmic antibody (ANCA) associated vasculitis, anti-glomerular basement membrane (anti-GBM) and lupus nephritis are the most common causes of RPGN and are characterized by the formation of glomerular crescents and infiltration of leukocytes that eventually lead to glomerulosclerosis and kidney failure. In this work, we used high-resolution spatial transcriptomics of 32 ANCA, 19 lupus nephritis, 6 anti-GBM, and 6 control patients to understand how intercellular signaling between immune and renal tissue cells leads to renal inflammation and glomerular injury. Using 3,218,210 immune and kidney cells, we observed that the biological pathways involved in the sequence of glomerular crescent formation are similar across the diseases. While innate immune cells infiltrated the glomerular compartment relatively early, later increases in adaptive immune cells were largely restricted to the periglomerular regions. These changes in immune cells temporally correlated with increases in glomerular parietal epithelial (PEC) and fibrotic mesangial cells, suggesting disease-relevant functional signaling between these immune and renal cells. Cell communication analysis revealed early disease PDGF signaling from epithelial and mesangial cells to PECs, causing their activation and proliferation. At later stages, TGF-β signaling from macrophages, T cells, epithelial cells, and mesangial cells to PECs triggered the expression of extracellular matrix components resulting in glomerulosclerosis. Our results highlight a spatio-temporally conserved progression into glomerular crescents and sclerosis for ANCA, lupus nephritis, and anti-GBM disease, which is driven by consecutive PDGF and TGF-β signaling to PECs.
overall_design: 8 slides containing 63 samples (biopsies from 32 ANCA, 19 lupus nephritis, 6 anti-GBM, and healthy tissues from tumor nephrectomies of 6 control patients) were analyzed with Xenium to capture RNA. One additional slide (Sample: 0011186) was used for Xenium + phenocycler run.
sample type: RNA
number of samples: 9
organism: Homo sapiens
data processing: Baysor (v0.6.2) was used to re-segment cells with prior being default segmentations (cell_boundaries.parquet). Scanpy (v1.10.1) workflows were used for normalization to the median the downstream analysis.
sample molecule: total RNA
Protocols Used:
treatment:
sample growth:
extract: 5 µm sections were taken for Xenium.
label: Xenium Human Probe Set
scan: Xenium In Situ Analyzer

Figure 5. Example of GEO information provided along with the prompt to the LLM. This example is for GSE ID: GSE294965

Benchmark dataset

Disease-annotation performance was evaluated using a benchmark of 200 GEO Series originally curated by Chen et al. as part of their GEO disease-annotation work. Each Series in this dataset is labeled as either having at least one associated disease term or no disease, and disease-positive Series are annotated with one or more DOID identifiers. From the Chen et al.

dataset, a table was constructed with one row per GEO Series (GSE ID) and the following fields: a Boolean flag indicating whether any disease is present (Disease (TRUE/FALSE)), and a semicolon-separated list of gold-standard DOIDs for disease-positive Series (for example, DOID:3393;DOID:0050828).This dataset was supplemented with additional DOIDs for relevant disease classifications.

In total, the benchmark contains:

- 200 GEO Series,
- 86 disease-positive Series (at least one DOID), and
- 114 disease-negative Series (no disease DOID assigned).

This table is treated as the ground truth for disease presence and DOID assignments in all evaluations.

Model outputs and DOID normalization

For each of the 200 Series five methods were evaluated. The Chen et al. disease-annotation baseline, using the DOIDs and disease labels reported in their work, and the four LLM-based pipelines implemented in ExpressionGenesis (Claude 3.7 Sonnet, DeepSeek R1, Meta Llama 3.3 70B instruct, Meta Llama 3.2 11B instruct).

Each large language model pipeline consumes a structured prompt that includes the GEO Series title, summary, overall design, and related metadata, and returns one or more disease candidates together with candidate Disease Ontology IDs. For the experiments in this section, the focus is on series-level disease classification: whether a method correctly identifies at least one appropriate DOID for a disease-positive Series and avoids assigning disease to disease-negative Series.

For each model and each Series, predictions are represented as a set of DOIDs. A Series with no predicted DOID is treated as a negative prediction (“no disease”). For the gold standard, each Series likewise has either a non-empty set of DOIDs (disease-positive) or an empty set (disease-negative).

Addressing LLM Limitations with Retrieval Augmented Generation (RAG)

While large language models worked well in generating summaries, keywords, and identifying disease names, they occasionally introduced errors. These errors can be referred to as either hallucinations or misclassifications. Errors encountered included inaccurate DOIDs, as well as the LLM proposing diseases unrelated to human disease.

Among the disease-positive Series in the Chen et al. evaluation set, the LLM incorrectly associated (or hallucinated) a DOID that did not match the disease name in approximately 58% of cases. In other words, 58% of the DOIDs initially returned by the model did not correspond to

the disease name identified by the model, based on the Disease Ontology ground truth. To reduce these errors, a retrieval augmented generation (RAG) approach was implemented to look up DOIDs. RAG approaches have been shown to reduce the frequency of LLM hallucinations and provide factual context to LLMs (Lewis et al., 2021).

A two-step process for structured metadata generation was implemented:

1. Initial Pass: The model generates an initial draft of structured metadata using the standardized prompt and GEO entry information.
2. RAG-based Disease Ontology Mapping: Each disease term from the draft is evaluated against the Disease Ontology using a secondary LLM request supported by a RAG knowledge base. This step corrects or confirms DOIDs and helps to ensure that the disease name is standardized and aligns with human biology.

The RAG-based strategy reduced hallucinated DOIDs and improved the standardization of disease names.

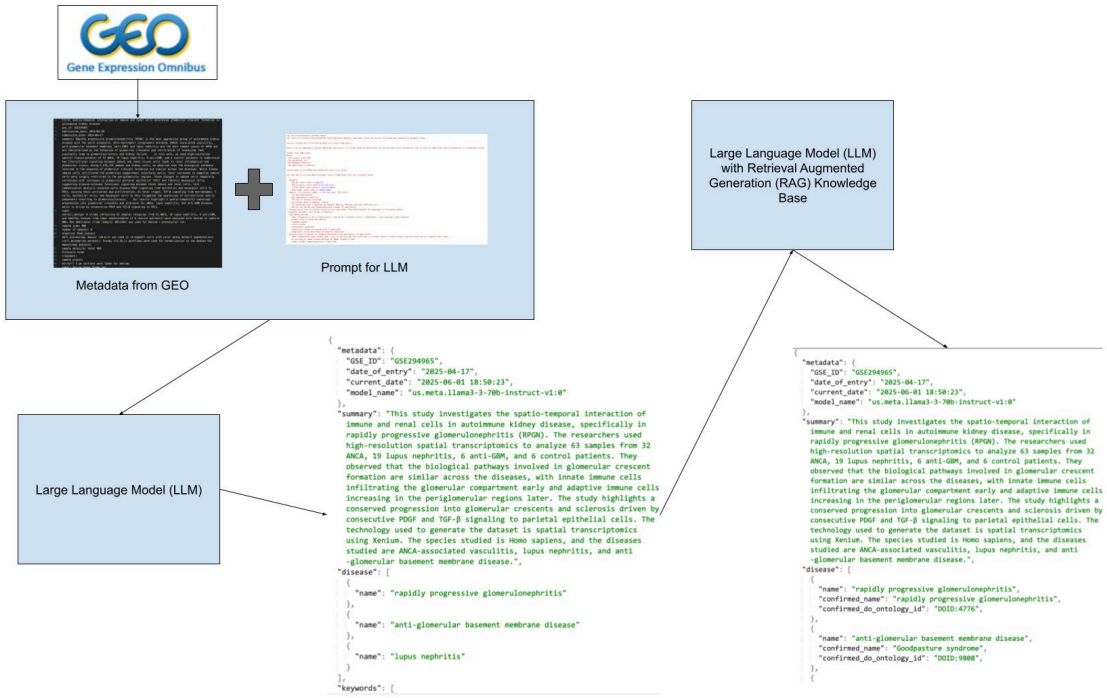

Figure 6. The two-step approach of structured metadata generation using LLMs and RAG

```
{
  "metadata": {
    "GSE_ID": "GSE294965",
    "date_of_entry": "2025-04-17",
    "current_date": "2025-06-01 18:50:23",
    "model_name": "us.meta.llama3-3-70b-instruct-v1:0"
  },
  "summary": "This study investigates the spatio-temporal interaction of immune and renal cells in autoimmune kidney disease, specifically in rapidly progressive glomerulonephritis (RPGN). The researchers used high-resolution spatial transcriptomics to analyze 63 samples from 32 ANCA, 19 lupus nephritis, 6 anti-GBM, and 6 control patients. They observed that the biological pathways involved in glomerular crescent formation are similar across the diseases, with innate immune cells infiltrating the glomerular compartment early and adaptive immune cells increasing in the periglomerular regions later. The study highlights a conserved progression into glomerular crescents and sclerosis driven by consecutive PDGF and TGF-β signaling to parietal epithelial cells. The technology used to generate the dataset is spatial transcriptomics using Xenium. The species studied is Homo sapiens, and the diseases studied are ANCA-associated vasculitis, lupus nephritis, and anti-glomerular basement membrane disease.",
  "disease": [
    {
      "name": "rapidly progressive glomerulonephritis",
      "confirmed_name": "rapidly progressive glomerulonephritis",
      "confirmed_do_ontology_id": "DOID:4776"
    },
    {
      "name": "anti-glomerular basement membrane disease",
      "confirmed_name": "Goodpasture syndrome",
      "confirmed_do_ontology_id": "DOID:9808"
    },
    {
      "name": "lupus nephritis",
      "confirmed_name": "lupus nephritis",
      "confirmed_do_ontology_id": "DOID:0080162"
    }
  ],
  "keywords": [
    "autoimmune kidney disease",
    "rapidly progressive glomerulonephritis",
    "spatial transcriptomics",
    "Xenium",
    "PDGF signaling",
    "TGF-β signaling",
    "parietal epithelial cells"
  ],
  "experimental_design": {
    "type": "Case-Control",
    "groups": [
      "ANCA-associated vasculitis",
      "lupus nephritis",
      "anti-glomerular basement membrane disease",
      "control patients"
    ],
    "treatment_groups": [],
    "control_groups": [
      "control patients"
    ],
    "experimental_conditions": [],
    "sample_size": {
      "ANCA": 32,
      "lupus nephritis": 19,
      "anti-GBM": 6,
      "control": 6
    },
    "comparison": "Disease groups vs. control group"
  }
}
```


Figure 7. Example response after RAG

Results

Disease Information Evaluation

To evaluate disease annotation, ExpressionGenesis was run on the 200 GEO Series benchmark from Chen et al., described earlier. For each Series, the manually curated Disease Ontology identifiers from Chen et al. were treated as the ground truth. Predictions from ExpressionGenesis and from the Chen et al. method were compared at the Series level using the accuracy, precision, recall, and F1 metrics defined above. Results are reported for four LLM models: Claude 3.7 Sonnet, DeepSeek R1, Meta Llama 3.3 (70 billion parameters), and Meta Llama 3.2 (11 billion parameters).

Model Evaluation

All four tested models (Claude 3.7 Sonnet, DeepSeek R1, Meta Llama 3.3 70B, and Meta Llama 3.2 11B) were evaluated using the complete two-step RAG approach described previously. Each model first generated initial structured metadata, then disease terms were validated against the Disease Ontology using the RAG-based correction step. The evaluation results in Figure 9 reflect the performance of each model after RAG validation.

Classification rule and metrics

Disease annotation was evaluated at the Series level using a confusion matrix with two actual classes (disease present vs disease absent) and three prediction outcomes (disease correct, disease incorrect, and no disease), as summarized in Figure 8.

A Series was considered as disease-positive (disease present) in the gold standard if its DOID set was non-empty, and disease-negative (disease absent) otherwise. For each method, predictions for a Series fell into one of three categories:

- Disease correct: the method returned at least one DOID, and at least one of those DOIDs matched a DOID in the gold-standard set for that Series. These cases were counted as true positives (TP).
- Disease incorrect: the method returned one or more DOIDs, but none of the predicted DOIDs appeared in the gold-standard set for that Series. For disease-present Series, these cases were treated as false negatives (FN) for the missed true disease and as false positives (FP) for the incorrect DOID assignment. For disease-absent Series, any predicted DOID was counted as an FP. In other words, a DOID prediction was counted as a false positive whenever it was not one of the DOIDs in the TRUE set.
- No disease: the method did not return any DOID for that Series. For disease-present Series, these cases were counted as false negatives (FN); for disease-absent Series,

they were counted as true negatives (TN).

This rule naturally handles Series annotated with multiple diseases: a Series with multiple gold-standard DOIDs is counted as a true positive if at least one of those DOIDs is returned.

	Predicted		
Actual	***Disease Correct***	***Disease Incorrect***	***No Disease***
Disease Present	TP	FP	FN
Disease Absent	-	FP	TN

Figure 8. Evaluation confusion matrix

From the aggregated counts of TP, FP, FN, and TN, standard classification metrics were computed:

- Accuracy = $\frac{TP + TN}{TP + FP + FN + TN}$
- Precision = $\frac{TP}{TP + FP}$
- Recall = $\frac{TP}{TP + FN}$
- F1 Score = $2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$

All metrics were micro-averaged across the 200 Series.

Evaluation Results

	Chen et al.	claude-3-7-sonnet	deepseek-r1	llama-3-3-70b	llama3-2-11b
Accuracy	0.86000	0.90000	0.89500	0.88500	0.76000
Precision	0.84722	0.88750	0.84615	0.84091	0.57692
Recall	0.78205	0.86585	0.91667	0.89157	0.93750
F1 Score	0.81333	0.87654	0.88000	0.86550	0.71429

Figure 9. Evaluation results of disease annotation

Claude 3.7 Sonnet, DeepSeek R1, and Llama 3.3 (70b) outperformed Chen et al. in terms of accuracy and F1 Score. Among the best performers, Llama 3.3 (70b) was the most cost-effective to scale, and was selected as the production model for metadata generation in ExpressionGenesis.

User Interface

The ExpressionGenesis web interface was developed using Next.js and TypeScript. It currently offers a main dataset browser, individual detailed GEO entry pages, and a submission trends page (Figure 12). Samples of the interface are shown below:

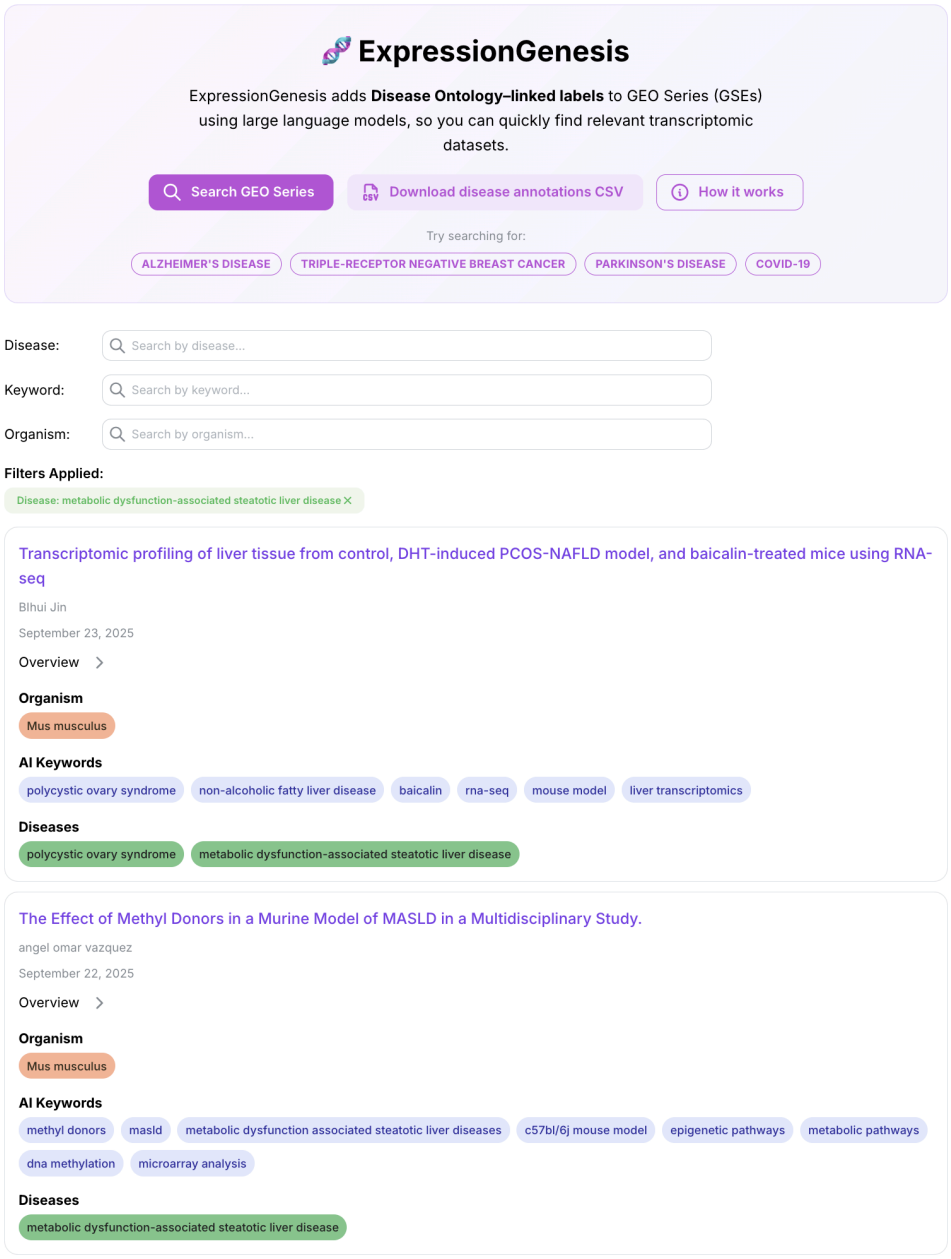

Figure 10. The main dataset browser page lists all current GEO Series records. It is filterable by keyword and disease name.

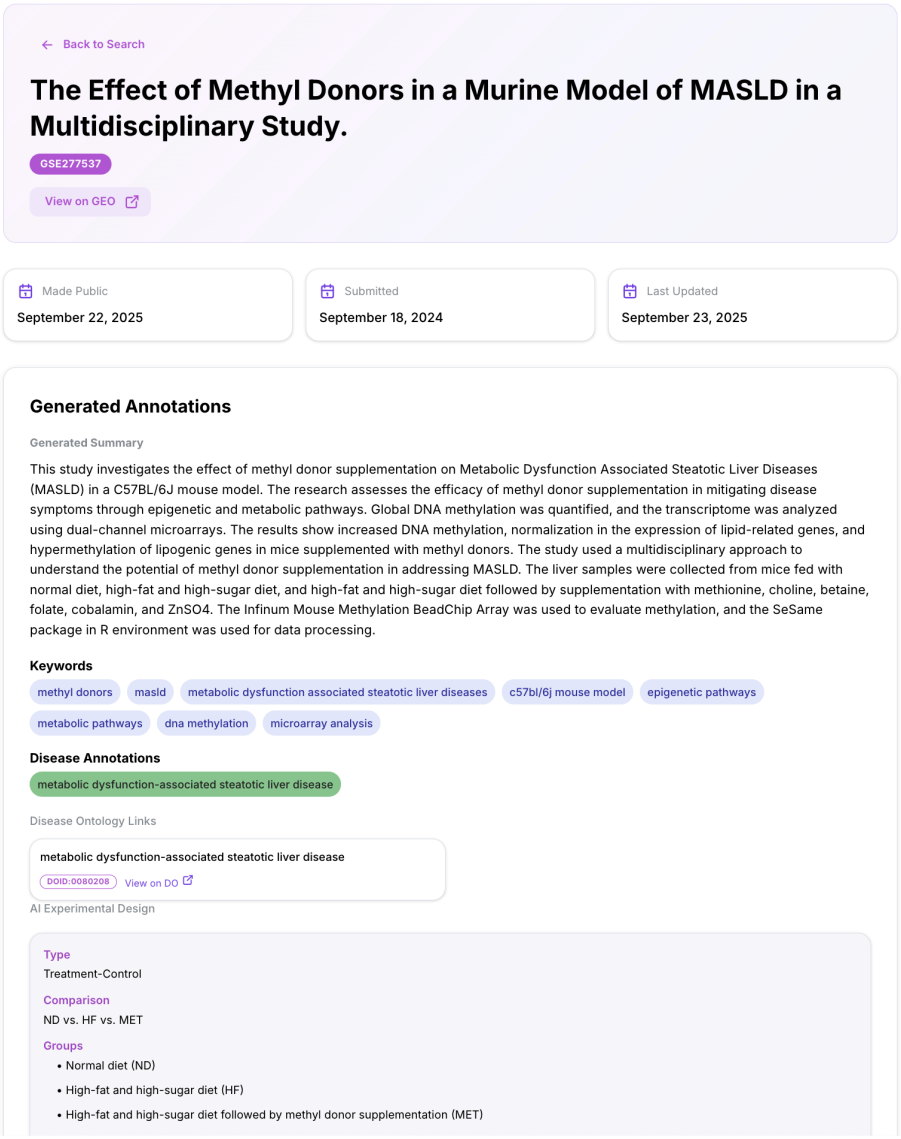

Figure 11. Individual detail pages provide metadata that combines original GEO text with enriched annotations from ExpressionGenesis.

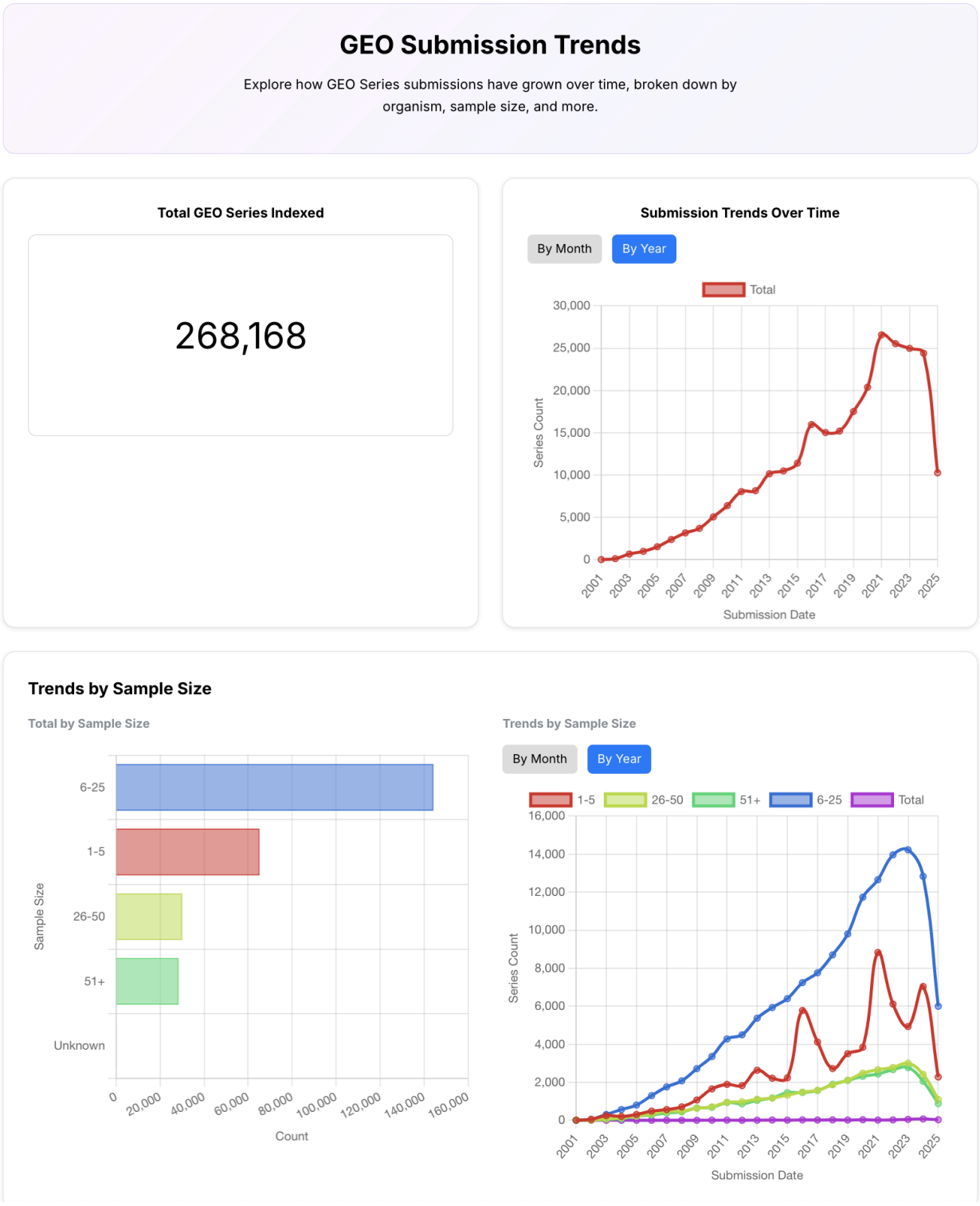

Figure 12. The submission trends page shows aggregated visualizations of trends in GEO Series submissions over time.

In addition to interactive browsing, ExpressionGenesis provides a Download page (https://www.expressiongenesis.com/downloads) that provides the disease annotations as a single CSV file. Each row corresponds to a single disease annotation for a GEO Series and includes three columns: gse_id (GEO Series accession), disease_name (standardized disease label), and DOID (Disease Ontology identifier).

Cost and Scalability

The complete ExpressionGenesis system costs approximately $50 USD per month to maintain on AWS, including data processing, LLM inference, RAG validation, and web hosting. Individual annotations cost approximately $0.0025 USD per GEO Series entry. Annotating the entire current GEO repository of 268,000+ entries (as of Dec 2025) costs about $670 USD in computation.

GEO Series submissions have averaged approximately 25,000 new entries per year over the past three years. At $0.0025 per entry, annotating new submissions costs roughly $5 per month. The serverless architecture automatically scales to handle daily submission volumes without manual server management.

By comparison, manual curation at typical research rates would cost hundreds to thousands of times more and cannot scale to match GEO's growth rate. This cost efficiency addresses one of the limitations faced by prior tools like GEOMetaCuration and ReGEO, which are no longer maintained.

Discussion

ExpressionGenesis demonstrates the practical benefits of combining LLMs with RAG to standardize and enrich metadata for high-throughput gene expression datasets. This two-step approach improves the accuracy of disease annotations by linking predictions to the Disease Ontology database, significantly reducing hallucinations and inconsistent terminology. The use of unstructured free-text metadata in GEO has posed challenges for dataset discovery and reuse. ExpressionGenesis addresses some of these issues by automatically generating structured fields, such as disease names, experimental design, and study summaries. Compared to previous works that relied on manual curation or rule-based methods, ExpressionGenesis provides a scalable, automated pipeline for GEO annotation. Evaluation results confirm that LLMs can outperform prior approaches in both precision and recall, resulting in fewer false positives and false negatives. These improvements are important given the ever-growing size of GEO.

In addition to backend processing, ExpressionGenesis provides an interface for researchers to discover valuable datasets. The interface enables users to search and filter by disease and keyword, enhancing dataset discoverability. By increasing the structured metadata associated

with GEO Series, ExpressionGenesis enhances the use of GEO for researchers and provides new opportunities for meta-analysis and data reuse.

Conclusion

ExpressionGenesis uses large language models, retrieval-augmented generation, and the Disease Ontology to generate standardized metadata for GEO Series. Evaluation on a manually curated benchmark shows that LLM-based pipelines can outperform a prior NLP approach for disease annotation. The public ExpressionGenesis web application makes these enriched annotations available to researchers, improving the ability to search for and reuse GEO datasets.

Next Steps

Future work will focus on expanding the scope of metadata extracted from GEO entries. Specifically, future enhancements will include expanded biological attributes such as tissue type and cell line identification. As with disease annotations, these attributes can be linked to relevant ontological databases to support standardization and semantic interoperability. These enhancements will improve dataset filtering and offer new ways to interpret and explore GEO entries.

On the website, additional search options will be developed that allow users to filter by organism, technological platform, study design, and other new metadata fields. Filtering will also be expanded to support ontology-aware queries. This will enable queries to include related terms within the ontological hierarchy. For example, a search for the disease “breast cancer” could also return results for “progesterone-receptor negative breast cancer”, “breast fibrosarcoma”, and other breast cancer subtypes classified under the same parent term.

In addition, similarity-based scoring and dataset recommendation features will be implemented to assist researchers in discovering related datasets by comparing study summaries, disease terms, and experimental features. Together, these improvements will support exploration and reuse of public gene expression data.

Data Availability

ExpressionGenesis can be accessed at:
https://www.expressiongenesis.com/

The ExpressionGenesis disease annotations can be downloaded from the ExpressionGenesis Downloads page (https://www.expressiongenesis.com/downloads)

Acknowledgments

The author thanks Karol Estrada, Ph.D., Brandeis University, for mentorship and guidance throughout this project.

Funding

This work received no external funding.

Competing Interests

The author declares no competing interests.

References

Chen, G., Ramírez, J. C., Deng, N., Qiu, X., Wu, C., Zheng, W. J., & Wu, H. (2019). Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. *Database: the journal of biological databases and curation*, 2019, bay145. https://doi.org/10.1093/database/bay145

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, *30*(1), 207–210. https://doi.org/10.1093/nar/30.1.207

Giles, C. B., Brown, C. A., Ripperger, M., Dennis, Z., Roopnarinesingh, X., Porter, H., Perz, A., & Wren, J. D. (2017). ALE: automated label extraction from GEO metadata. *BMC bioinformatics*, 18(Suppl 14), 509. https://doi.org/10.1186/s12859-017-1888-1

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (No. arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Li, Z., Li, J., & Yu, P. (2018). GEOMetaCuration: A web-based application for accurate manual curation of Gene Expression Omnibus metadata. Database: *The Journal of Biological Databases and Curation*, 2018, bay019. https://doi.org/10.1093/database/bay019

Schriml, L. M., Munro, J. B., Schor, M., Olley, D., McCracken, C., Felix, V., Baron, J. A., Jackson, R., Bello, S. M., Bearer, C., Lichenstein, R., Bisordi, K., Dialo, N. C., Giglio, M., & Greene, C. (2022). The

Human Disease Ontology 2022 update. *Nucleic acids research*, 50(D1), D1255–D1261. https://doi.org/10.1093/nar/gkab1063

Wang, Z., Monteiro, C. D., Jagodnik, K. M., Fernandez, N. F., Gundersen, G. W., Rouillard, A. D., Jenkins, S. L., Feldmann, A. S., Hu, K. S., McDermott, M. G., Duan, Q., Clark, N. R., Jones, M. R., Kou, Y., Goff, T., Woodland, H., Amaral, F. M. R., Szeto, G. L., Fuchs, O., Schüssler-Fiorenza Rose, S. M., … Ma'ayan, A. (2016). Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature communications*, 7, 12846. https://doi.org/10.1038/ncomms12846