Accelerating Curation of CRISPR Screen Data from Public Repositories

Contact
Shantanu Bafna
Bioinformatics Engineer III
+91 96298 70197
shantanu.bafna@strandls.com
VISIT LINKEDIN

Radhakrishna Bettadapura[1], Aastha Tripathi[1], Manjushree Sahoo[1], Niharika A[1], Suraj Kumar Sharma[1], Anna Zacharia[1], Hima Varghese[1], Charles Lu[2], Neil Kuehnle[2], Swaraj Basu[1], Shantanu Bafna[1]

[1]Strand Life Sciences, Bengaluru, India
[2]AbbVie, North Chicago, IL, USA 600

VISIT OUR WEBSITE

Introduction

CRISPR (clustered regularly interspaced short palindromic repeats) has revolutionized functional genomics by enabling systematic, large-scale studies of gene function.

CRISPR screening approaches

Pooled screens: perturbations in mixed cell populations; gene effects read out via selection/sequencing.
Arrayed screens: perturbations delivered well-by-well; suited for high-resolution phenotypic assays.

Current challenge

Rapidly growing datasets, but finding the right CRISPR data requires manual, expert led curation.
This process is:

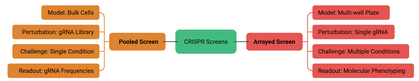
1. Labor-intensive
2. Hard to scale
3. Not easily reproducible

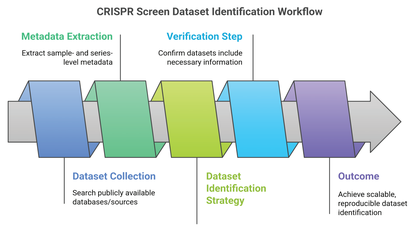
Why curation matters

- Ensures dataset discoverability
- Supports cross-study comparisons
- Accelerates target discovery and therapeutic development

Our approach

We explored methods to streamline curation using NLP, AI-driven agents, and automated pipelines, making CRISPR datasets easier to organize, search, and apply.

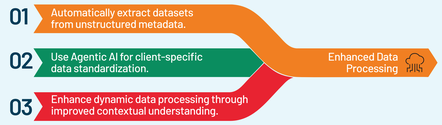




Conclusion

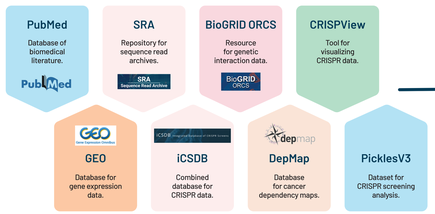
- The identification and curation of CRISPR screen datasets remain a critical yet challenging step for large-scale functional genomics. Our multi-pronged strategy manual review, pattern matching, semantic search, and LLM-based approaches showed that while accurate identification is possible, it often requires significant time, labor, or computational resources. Among these, LLM-based methods emerged as the most scalable and efficient, offering both precision and interpretability.
- Although the accuracy of GPT-4o (~82%) is considered good, it is expected to improve as newer models gain better contextual understanding. Accuracy will also benefit from reinforcement learning and iterative input refinement by data scientists and bioinformaticians, further strengthening identification reliability.
- Looking ahead, the challenge shifts from dataset identification to making them analysis-ready. This requires harmonization on two fronts:
  (1) ontology harmonization to standardize metadata; and
  2) processing harmonization to ensure uniform pipelines for raw data and downstream analysis.
- Establishing such a framework will enable a unified repository of high-quality CRISPR screens, accelerating integrative analyses and downstream applications such as predictive modeling, network inference, and therapeutic discovery.

Streamlining Data Solutions

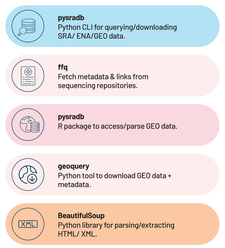


Materials And Methods

Data Sources



Metadata extraction tools



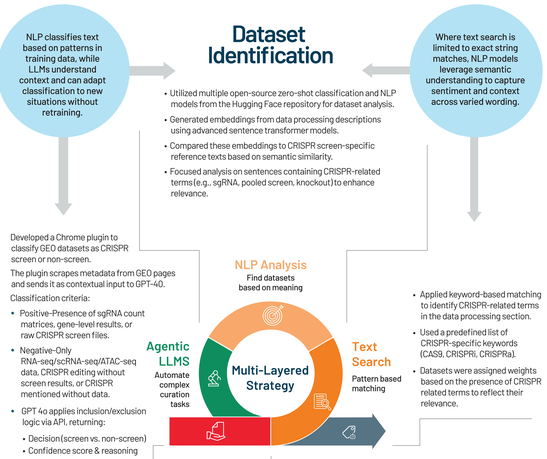
- GEO serves as the main resource: Among all databases surveyed, GEO was the main resource for data extraction and curation, since most other repositories were directly linked or cross-referenced to GEO.
- Metadata extraction tools: Combination of tools was used for reproducible data retrieval, including pysradb, ffq, GEOquery, geofetch, and BeautifulSoup.
- Series-level (GSE): Represents entire studies or experiments, containing high-level details such as library strategy, overall design, and data processing pipelines.
- Sample-level (GSM): Contains detailed metadata for each biological sample, including sample characteristics, library preparation, sequencing platform, and associated processed/raw data.

Series (GSE)

Represents the overall study/experiment, providing design and context. It groups and organizes multiple samples.

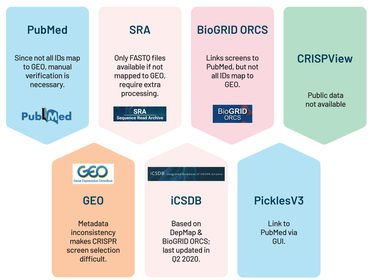
Sample (GSM)

Represents an individual biological sample with specific details. It is linked to raw and processed data files.

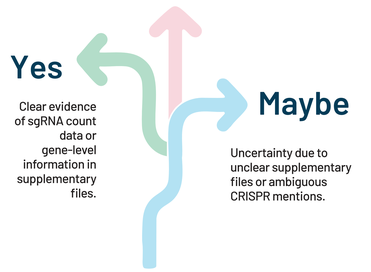


Results

Challenges with Databases



CRISPR screen Dataset Classification



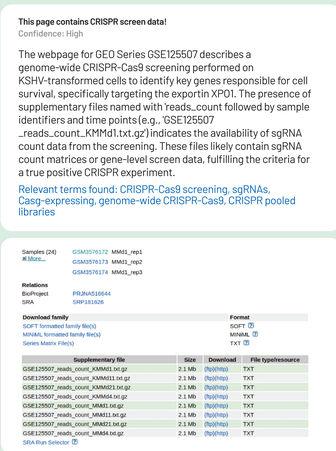
Text Search

- Utilized string-matching approaches to detect CRISPR-related terms within dataset metadata (esp. data processing fields).
- Constructed a predefined keyword list covering multiple CRISPR screen terminologies (e.g., CRISPR screen experiment, sgRNA-based study, CRISPR perturbation, sgRNA counts, raw CRISPR screen data).
- Expanded search to include different forms of labeling within Library Strategy fields (e.g., "CRISPR screen").
- Applied keyword search across both Series (GSE) and Sample (GSM) metadata levels.
- Identified candidate datasets for further validation by manual review, NLP-based semantic search, and AI classification.
- Helped capture datasets that were correctly labeled but otherwise missed due to inconsistent annotation across repositories.

Semantic Search

- Implemented semantic search to identify CRISPR screen studies across multiple metadata fields (e.g., summary, overall design, data processing). FremyCompany/BioLORD-2023
- Sentence Transformer performed relatively better for this task (10-15% increased sensitivity at similar specificity), as it is heavily trained on clinical sentences and biomedical concepts. Generated sentence embeddings from each dataset's processing-related text using the BioLORD-2023 model.
- Compared embeddings against predefined CRISPR-specific reference texts using cosine similarity.
- Flagged datasets exceeding a similarity threshold as potential CRISPR screen studies. This approach focused on conceptual understanding rather than exact keyword presence.
- However, often misclassified datasets where CRISPR-related terms appeared in the description as part of the analysis, but the dataset was not a true CRISPR screen (e.g., different library strategy also used in the experiment).

GEO LLM Plugin



- Developed a Chrome browser plugin to identify whether dataset is a CRISPR screen based on its GEO (GSE/GSM) page.
- The plugin scrapes GEO page content and compiles it as contextual input. Content is passed to GPT-4o , which evaluates datasets using predefined inclusion/exclusion criteria. Furthermore, bulk dataset classification can be performed by executing GPT API calls with the specified context on the backend via a Python script.
- Positive classification: presence of CRISPR screen outputs (e.g., sgRNA count matrices, gene-level results, raw screening files).
- Negative classification: datasets limited to single-cell, RNA-seq, ATAC-seq; CRISPR editing/knockout without screening results; or CRISPR mentioned without relevant data.
- The plugin makes a backend API call, sending scraped content with structured prompts encoding the rules.

  GPT-4o returns:
  1) Classification (CRISPR screen or not)
  2) Confidence score
  3) Reasoning
  4) Extracted keywords supporting the decision

Performance Comparison of Classification Approaches

- The performance of separate approaches was based on 100 randomly selected datasets.
- Manual curation ensures high sensitivity and specificity but requires substantial manual effort and deep understanding of experimental design and biological context to accurately review each series/experiment ID and its samples.
- Pattern matching (text-based search) is highly sensitive but lacks specificity, leading to more false positives.
- Semantic analysis offers medium to low sensitivity and specificity, and requires extensive tuning for optimal performance.
- Large Language Models (LLMs) provide both high sensitivity and high specificity, and are efficient and scalable with proper tuning.

Approach	Sensitivity	Specificity	Notes
Manual Curation	High	High	Ensures accurate data but require significant manual effort to review each series/experiment ID and its associated samples.
Pattern Matching (Text-Based)	High	Low	Sensitive but lacks specificity.
Semantic Analysis	Medium to Low	Medium to Low	Requires extensive tuning for optimal results.
Large Language Models(LLMs)	High	High	Efficient and scalable with tuning

Approach	Precision			Accuracy
	Yes	No	Maybe	
Text search	70.70%	60.00%	15.00%	59%
Semantic search	71.20%	58.30%	33.30%	67%
LLM	87.30%	77.40%	50%	82%