GenePrep: Unified Genomic Data Formatter for Statistical Analysis

Yue Xu
Columbia University
yx2876@columbia.edu

Sirihaasa Nallamothu
University of Illinois
Urbana-Champaign
sn37@illinois.edu

Haoyang Liu
University of Illinois
Urbana-Champaign
hl57@illinois.edu

Haohan Wang
University of Illinois
Urbana-Champaign
haohanw@illinois.edu

Abstract

GenePrep is an automated multi-agent system that streamlines preprocessing and analysis of large-scale gene expression data from GEO and TCGA. By simply installing the GenePrep package, users can perform end-to-end workflows including data validation, trait–condition pair selection, statistical testing, and result generation. Given a dataset and trait–condition pairs, GenePrep identifies genes associated with traits while accounting for conditions. Its modular agents enable iterative planning, execution, and debugging with minimal manual scripting. GenePrep reduces preprocessing overhead and improves reproducibility. This work extends on tools to create teams of AI scientists and automate gene expression data analysis, as presented in Liu et al. (2024, 2025).

1 Introduction

The advent of large-scale genomic repositories such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) has dramatically expanded the scope of translational research, enabling integrative analyses that link molecular profiles to clinical outcomes. However, these rich resources pose substantial preprocessing challenges: heterogeneous file formats, inconsistent gene identifiers, missing metadata, and the need to merge clinical and molecular data into analysis-ready tables. Existing toolkits, such as Bioconductor's R packages (`edgeR`, `DESeq2`, `limma`) for expression normalization (Gentleman, 2004), the TCGAbiolinks wrapper for streamlined TCGA access and harmonization (Colaprico et al., 2016), and Python frameworks like Scanpy for high-dimensional omics (Wolf et al., 2018), provide valuable functionality but still require extensive manual scripting and lack unified support across multiple data sources.

To address these limitations, we introduce **GenePrep**, a unified genomic data formatter built on a modular multi-agent architecture. GenePrep automatically orchestrates each preprocessing step—cohort retrieval, gene symbol normalization, clinical–molecular integration, missing-value handling, and quality validation—through specialized software agents that communicate via structured messages. GenePrep is a tool.

The preprocessing of genomic and clinical data is a critical step for any downstream statistical or machine learning analysis. Numerous tools and platforms have emerged for this purpose. One of the most widely used is Bioconductor, which offers extensive R-based packages such as `edgeR`, `DESeq2`, and `limma` to standardize raw gene expression data Gentleman (2004). Tools like TCGAbiolinks facilitate direct access to TCGA data and provide wrappers for data harmonization Colaprico et al. (2016). In Python, packages like Scanpy Wolf et al. (2018) and AnnData Virshup et al. (2021) have been widely adopted for processing high-dimensional single-cell and bulk gene expression matrices.

However, these tools often lack interoperability across datasets from different consortia (e.g., TCGA vs.

GEO) and demand significant manual coding for cohort alignment, gene symbol normalization, and phenotype merging. GenePrep aims to resolve these pain points by automating these preprocessing tasks in a reproducible, agent-based framework.

Multi-agent systems (MAS) have increasingly found applications in biomedical research, including drug discovery, biomedical literature mining, and gene network inference. More recently, AI-based agents like AutoML and GPT-4 powered research assistants have begun to explore hypothesis generation and experiment simulation within biomedical contexts. Agent frameworks such as OpenAgents and CAMEL have shown promise in coordinating multiple LLM-based agents for code generation and scientific reasoning Zhu et al. (2023).

Despite these developments, existing MAS implementations are often domain-agnostic and lack dedicated genomic data preprocessing pipelines. GenePrep uniquely contributes a specialized MAS architecture, where each agent is assigned a clear data transformation or quality-check role within the genomics preprocessing workflow. This builds on earlier studies that introduced AI-powered frameworks for scientific discovery from gene expression data Liu et al. (2024, 2025), this work advances the unified preprocessing pipeline and agent-based system design to facilitate robust genomic data integration and analysis.

2 Methodology

2.1 Multi-agent System

To streamline the complex preprocessing tasks required for large-scale biomedical datasets such as TCGA and GEO, we designed a multi-agent system inspired by Anthropic's modular agent framework Anthropic (2024). Each agent in our system is instantiated as a subclass of a shared `BaseAgent` class, inheriting asynchronous message-passing, prompt construction, and memory management capabilities. Agents are role-specialized—including `TCGAAgent`, `GEOAgent`, `DomainExpertAgent`, and `CodeReviewerAgent`—and coordinated by a principal investigator agent (`PIAgent`).

The programming agents (`TCGAAgent`, `GEOAgent`) follow a multi-step planning architecture defined by the `MultiStepProgrammingAgent`, which leverages a structured `TaskContext` to execute action units, revise code based on reviewer feedback, and recover from execution errors. Domain knowledge injection and review loops are managed dynamically depending on the step type and action requirements. Once the plan is selected, a code writing request is issued and optionally routed through a domain expert or code reviewer. Review feedback may trigger revision requests, which iterate until the step is marked complete or the maximum revision round is reached.

2.2 Auto-adjusted Script

This pipeline integrates clinical and genetic data from The Cancer Genome Atlas (TCGA) or Gene Expression Omnibus (GEO) into a structured and analyzable format. First, it extracts selected clinical features (e.g., phenotype label, age, gender) Next, gene expression data is standardized by normalizing gene symbols to ensure consistent identifiers. To join clinical and genetic features, gene data is transposed to match sample orientation, followed by an inner join. The pipeline then handles missing values and filters out biased features that might skew model training. Finally, the processed dataset undergoes quality validation. If deemed usable, it is saved for future use. This step-by-step pipeline ensures that data integrity is maintained across clinical and genomic features, enabling reproducible and high-quality downstream analysis. During the process, CSV files will be generated.

3 Discussion

GenePrep simplifies a traditionally complex and manual preprocessing workflow by automating data harmonization tasks across TCGA and GEO. Its agent-based design breaks down each step—data retrieval, normalization, integration, and validation—into modular, auditable components.

This structure not only reduces errors but also improves reproducibility and adaptability across projects. While current bioinformatics tools demand extensive manual scripting, GenePrep minimizes user intervention while maintaining flexibility. However, its performance may be influenced by the choice of LLM models for the agents, which is a topic that shall be analyzed by further researches.

References

Anthropic. Building effective agents. https://www.anthropic.com/engineering/building-effective-agents, 2024. Accessed: 2025-05-26.

A. Colaprico, T. C. Silva, C. Olsen, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, 44(8):e71–e71, 2016.

R. e. a. Gentleman. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

H. Liu, Y. Li, J. Jian, Y. Cheng, J. Lu, S. Guo, J. Zhu, M. Zhang, M. Zhang, and H. Wang. Toward a team of ai-made scientists for scientific discovery from gene expression data, 2024. URL https://arxiv.org/abs/2402.12391.

H. Liu, S. Chen, Y. Zhang, and H. Wang. Genotex: An llm agent benchmark for automated gene expression data analysis, 2025. URL https://arxiv.org/abs/2406.15341.

I. Virshup, S. Rybakov, F. J. Theis, et al. anndata: Annotated data. *Zenodo*, 2021.

F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

Y. Zhu, X. Li, T. Wu, et al. Camel: Communicative agents for mind exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023.