Bioinformatics, 2025, 11–9
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 09 November 2025
Manuscript Category

Computational Genomics Final

Fine-tuning a Biomedical Language Model for Leukemia Subtype Prediction from Gene Expression Data

Charlotte Cheung, Claire Cui

[1]Department of XXXXXXX, Address XXXX etc., [2]Department of XXXXXXX, Address XXXX etc.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Leukemia comprises more than twenty biologically distinct subtypes that differ in progression, treatment response, and patient demographics. These subtypes are characterized by distinct patterns of gene expression, reflecting underlying molecular mechanisms. Acute myelogenous leukemia (AML) is both the most common adult acute leukemia and among the most aggressive forms, making accurate subtype identification critical for diagnosis and treatment planning. Recent advances in RNA-based representation learning provide an opportunity to leverage large-scale pretrained models to improve disease classification from transcriptomic data.

Results: The expected outcome is improved subtype prediction accuracy, particularly for AML, with reduced model training time and improved generalization across datasets.

Availability: The quick brown fox jumps over the lazy dog.

Contact: ccui8@jh.edu, ccheung28@jh.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Leukemia is a heterogeneous group of blood cancers with more than twenty clinically recognized subtypes that differ in their cellular origin, aggressiveness, and therapeutic response. These subtypes are characterized by distinct patterns of gene dysregulation, and gene expression profiling has played a central role in both disease classification and biomarker discovery.

Recent advances in foundation models for gene expression data provide new opportunities for improving disease subtype prediction. Geneformer is a transformer-based biomedical language model pretrained on tens of millions of single-cell transcriptomes, enabling it to learn generalizable patterns of gene co-expression and regulatory structure. By applying transfer learning, we aim to adapt this broadly trained model to the task of leukemia subtype classification using bulk expression data. This approach has the potential to improve subtype prediction accuracy while requiring substantially fewer labeled samples than training a model from scratch, providing a powerful framework for translational applications in cancer genomics.

2 Methods

We will use bulk leukemia gene expression data from the Gene Expression Omnibus (GEO) dataset GSE13159, which contains approximately 2,000 patient samples representing multiple leukemia subtypes. The series matrix file will be parsed to retrieve the gene expression matrix and associated sample metadata. Probe identifiers will be mapped to official gene symbols using the corresponding microarray platform annotation file, and when multiple probes correspond to the same gene, their expression values will be aggregated into a single value. The resulting dataset will consist of a gene-by-sample matrix along with harmonized subtype annotations for each sample.

We will apply a multi-step normalization strategy. First, expression values will be log-transformed to stabilize variance. Then, we will apply quantile normalization to reduce technical variation arising from differences in array intensity distributions across samples. The dataset will be split into training and testing subsets while preserving subtype distributions. Model training will involve optimization procedures appropriate for high-dimensional gene expression data, and learning rate and regularization settings will be selected to prevent overfitting.

To evaluate model generalizability, we will test the fine-tuned model on independent AML datasets (GSE122505, GSE122511, and GSE122515), which will be processed using the same normalization and rank-based representation steps.

3 Results

The primary training dataset for this study will be GSE13159, which contains 2,096 samples profiled on the Affymetrix Human Genome U133 Plus 2.0 Array. This dataset spans a wide spectrum of leukemia subtypes, including multiple genetically defined forms of acute lymphoblastic leukemia (ALL), acute myelogenous leukemia (AML), chronic lymphocytic and myeloid leukemias (CLL and CML), myelodysplastic syndrome (MDS), and healthy or non-leukemic bone marrow samples. The diversity and scale of this dataset make it suitable for learning subtype-specific expression signatures.

Expression values and metadata from GSE13159 have been successfully extracted and aligned. Probe identifiers have been mapped to unique gene symbols, and samples have been assigned standardized subtype labels. The resulting expression matrix consists of gene-level expression values for 2,096 samples, with no missing or mismatched identifiers. These preprocessing steps establish a consistent and clean training set for model fine-tuning.

To evaluate generalization, we will test the model on three independent leukemia datasets, GSE9476 (64 samples; AML vs. healthy controls), GSE51082 (141 samples; AML and multiple ALL/CLL/CML subtypes), and GSE37642 (562 AML samples).

Name	Year Submitted	Last Updated	Microarray Template	Number of Samples	Sample Types
GSE13159	2008	March 2020	Affymetrix Human Genome U133 Plus 2.0 Array	2096	mature B-ALL with t(8;14); Pro-B-ALL with t(11q23)/MLL; c-ALL/Pre-B-ALL with t(9;22); T-ALL; ALL with t(12;21); ALL with t(1;19); ALL with hyperdiploid karyotype;c-ALL/Pre-B-ALL without t(9;22); AML with t(8;21); AML with t(15;17); AML with inv(16)/t(16;16); AML with t(11q23)/MLL; AML with normal karyotype + other abnormalities; AML complex aberrant karyotype; CLL; CML; MDS; Non-leukemia and healthy bone marrow
GSE9476	2007	November 2018	Affymetrix Human Genome U133A Array	64	Healthy donors; AML
GSE51082	2013	November 2018	Affymetrix Human Genome U133A Array	141	AML; precursor B-cell acute lymphoblastic leukemia; T-cell acute lymphoblastic leukemia; chronic lymphocytic leukemia; chronic myeloid leukemia;myelodysplastic syndrome
GSE37642	2012	January 2021	Affymetrix Human Genome U133A Array	562	AML

4 Discussion

Because the training dataset (GSE13159) contains many distinct leukemia subtypes with clear cytogenetic and molecular differences, we expect the fine-tuned Geneformer model to learn stable expression patterns characteristic of each subtype. However, true test performance will depend on how well these subtype-defining features transfer to external patient cohorts not seen during training.

The independent testing datasets vary in composition, disease focus, and cohort size. For example, GSE9476 provides a relatively simple AML-vs-normal classification scenario, while GSE51082 introduces multiple acute and chronic leukemia subtypes. GSE37642, which contains only AML cases, will allow us to assess performance on intra-AML stratification, where subtype expression differences are often more subtle. Therefore, we expect performance to be higher for subtype groups with strong molecular signals (e.g., t(15;17) PML-RARA AML and t(9;22) BCR-ABL ALL), and lower for highly heterogeneous subtypes (e.g., AML with normal karyotype).

Conflict of Interest: none declared.

References

Chennamadhavuni, A., Iyengar, V., Mukkamalla, S.K.R., et al. (2025) Leukemia. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK560490/

Selvaraj, S., Alsayed, A.O., Ismail, N.A., Kavin, B.P., Onyema, E.M., Seng, G.H., Uchechi, A.Q. (2024) Super learner model for classifying leukemia through gene expression monitoring. Discover Oncology, 15, 1337. https://doi.org/10.1007/s12672-024-01337-x

Theodoris, C.V., Xiao, L., Chopra, A., et al. (2023) Transfer learning enables predictions in network biology. Nature, 618, 616–624. https://doi.org/10.1038/s41586-023-06139-9

Warnat-Herresthal, S., Ulas, T., Schultze, J.L., et al. (2009) Microarray Innovations in Leukemia (MILE) Study: Stage 1 Data. Gene Expression Omnibus (GEO), GSE13159. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13159

Warnat-Herresthal, S., Ulas, T., Schultze, J.L., et al. (2020) Bonn Dataset 1 of meta-analysis on AML classification. Gene Expression Omnibus (GEO), GSE122505. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122505

Warnat-Herresthal, S., Ulas, T., Schultze, J.L., et al. (2020) Bonn Dataset 2 of meta-analysis on AML classification. Gene Expression Omnibus (GEO), GSE122511. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122511

Warnat-Herresthal, S., Ulas, T., Schultze, J.L., et al. (2020) Bonn Dataset 3 of meta-analysis on AML classification (RNA-seq). Gene Expression Omnibus (GEO), GSE122515. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122515

Yoo, M.S., et al. (2003) Oxidative stress-regulated genes in nigral dopaminergic neuronal cells: correlation with known pathology in Parkinson's disease. Brain Research Molecular Brain Research, 110, 76–84.