أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Dissecting Intra-Individual Tumor Heterogeneity in Childhood Leukemia: A Single-Cell Analysis of Maturation States and Ribosomal Expression

Ahmed Mawlawi
Department of Genetics, University of Cambridge | KAUST University
KAUST Academy Bioinformatics Summer Program
King Saud bin Abdulaziz University for Health Sciences, Jeddah

Mentored by: Hugo Tavares, Chandra Chilamakuri

Abstract

Childhood acute lymphoblastic leukemia (cALL) is characterized by the disrupted maturation and clonal expansion of bone marrow lymphoid precursors. While standard bulk sequencing stratifies patients into broad cytogenetic subtypes, it averages out transcriptional signals, masking intra-tumor heterogeneity and minor subclonal populations that drive therapeutic resistance and disease relapse. In this study, we developed an end-to-end single-cell RNA sequencing (scRNA-seq) pipeline to reproduce and advance the transcriptomic findings of Caron et al. (2020) [2], with a dedicated downstream focus on a High Hyperdiploid (HHD) patient sample.

By implementing strict per-sample quality filtering and restricting downstream clustering to resting G1-phase cancer cells ($n = 16,731$), we eliminated confounding cell-cycle variability while preserving true intra-individual biological diversity. Replacing manual canonical marker evaluation with automated SingleR reference mapping against the Human Primary Cell Atlas, we identified two distinct intra-tumor subclones in patient HHD_1 (Subclone 9, $n = 1,020$; Subclone 11, $n = 679$), demonstrating an early Pro-B ($CD34^{+}$) to mature B-cell developmental transition.

Non-linear LOESS trajectory modeling and Euclidean distance projections confirmed an inverse relationship between developmental pseudotime and ribosomal protein gene expression (RPS/RPL). Differential expression via MAST [3] and Gene Ontology enrichment via goseq [4] independently validated that subclonal divergence is predominantly driven by translation initiation and ribosome assembly pathways ($p_{adj} < 10^{-6}$). These findings establish that leukemic maturation arrest locks immature subclones into a hyper-active translational state, identifying developmental and metabolic heterogeneity as core features of cALL pathology.

Contents

- **Introduction** 2
- **Related Work** 2
- **Methodology** 2
  - ●.1 Method details 2
  - ●.2 Evaluation Procedure 2
- **Results** 2
- **Discussion** 3
- **Limitations and Future Work** 3
- **Conclusion** 3
- **References** 6

1. Introduction

Problem Motivation and Real-World Impact

Childhood acute lymphoblastic leukemia (cALL) is the most frequent pediatric malignancy, representing approximately 25% of all childhood cancer diagnoses. Although contemporary risk-stratified multi-agent chemotherapy regimens achieve 5-year event-free survival rates between 85% and 90%, refractory disease and post-treatment relapse remain a primary cause of disease-related mortality in pediatric oncology.

Problem Statement

Clinical decision-making historically relies on bulk genomic and transcriptomic sequencing to classify patients into diagnostic subtypes (e.g., High Hyperdiploid, *ETV6-RUNX1*, *TCF3-PBX1*, or *BCR-ABL1*). However, bulk profiling measures only the ensemble average of millions of cells, masking intra-individual cellular diversity, developmental plasticity, and minor resistant subclones present at diagnosis.

Key Assumptions and Limitations of the Problem

[1] **Cell-Cycle Confounding:** Leukemia blasts exhibit elevated proliferation rates; active cell cycling (S and $G2/M$ phases) introduces massive transcriptional variance that obscures subtle intra-tumor lineage and subclonal structures.

[2] **Annotation Subjectivity:** Manual identification of cell types using arbitrary expression thresholds of canonical marker genes introduces human expectation bias and fails to leverage whole-transcriptome information.

[3] **Biological vs. Technical Variance:** Standard scRNA-seq integration and batch-correction algorithms (e.g., CCA, Harmony) are engineered to align heterogeneous datasets across batches, which risks artificially erasing true intra-individual biological heterogeneity.

Summary of the Proposed System

To address these challenges, we built an analytical framework that:

- Isolates pure, resting G1-phase leukemic blasts across 8 cALL patients and 3 healthy pediatric controls ($n = 39,375$ raw cells).
- Implements automated, reference-based cell-type annotation using SingleR paired with the Human Primary Cell Atlas database.
- Constructs continuous developmental pseudotime trajectories using non-linear LOESS curves and Euclidean distance projections.
- Evaluates differential expression with MAST and corrects for RNA-seq gene-length bias using goseq to elucidate the metabolic and translational drivers of subclonal divergence.

2. Related Work

Genomic and Single-Cell Landscapes of Leukemogenesis

Genomic sequencing has established that cALL arises through initiating genetic lesions in lymphoid progenitors followed by secondary copy-number alterations and point mutations. Seminal multiregion sequencing by Gerlinger et al. (2012) [1] demonstrated that tumors follow branched Darwinian evolutionary patterns, where spatially and transcriptionally distinct subclones coexist within the same patient. In cALL, subclonal tracking has revealed that ancestral, minor subclones present at initial diagnosis frequently harbor treatment-resistant mutations (e.g., *NT5C2*, *RAS* pathway genes) that undergo positive selection during relapse.

The Benchmark Study (Caron et al., 2020)

Caron et al. (2020) [2] performed droplet-based scRNA-seq on 39,375 bone marrow mononuclear cells (BMMCs) from 8 pediatric cALL patients (6 B-ALL, 2 T-ALL) and 3 healthy pediatric donors. They reported intra-individual transcriptional clusters in 5 out of 8 patients, linking this heterogeneity to predicted developmental maturation states and ribosomal protein expression. However, their pipeline relied on manual inspection of canonical cell surface markers (such as *CD34*, *CD79A*, *MS4A1*, and *CST3*) and deployed a 15-node artificial neural network trained on healthy controls to predict pseudotime values.

Advantages and Gaps Addressed by the Proposed System

- **Unbiased Whole-Transcriptome Annotation:** We replaced manual marker inspection with SingleR, which calculates Spearman correlation coefficients across thousands of genes against reference databases, removing human expectation bias.
- **Direct Geometric Pseudotime Mapping:** Rather than relying on external neural network training, we directly modeled the developmental continuum by fitting a LOESS regression curve to the 2D manifold embeddings and computing orthogonal Euclidean projections.
- **Reproducible Statistical Modeling:** We integrated MAST generalized linear modeling [3] to account for scRNA-seq zero-inflation and paired it with goseq [4] to correct for transcript length bias in Gene Ontology enrichment.

3. Methodology

3.1 Method details

The complete analytical pipeline was structured into six sequential, modular stages to process the raw data and derive the final biological insights: [Raw 10X FASTQs / Matrices] ➔ 1. Quality Control & Per-Sample Filtering (min.features=200, min.cells=5, percent.mt < 8%) ➔ 2. Normalization, Scaling & Cell-Cycle Phase Assignment (G1, S, G2/M Scoring) ➔ 3. G1 Cancer Cell Isolation (Subtype != "Control" & Phase == "G1"; n = 16,731) ➔ 4. High-Resolution Graph-Based Clustering (Resolution = 1.3; 10% Size Threshold) ➔ 5. Automated Annotation (SingleR + Human Primary Cell Atlas) ➔ 6. LOESS Trajectory Mapping, MAST Differential Expression & goseq Pathway Enrichment

1. Quality Control and Per-Sample Filtering

To prevent rare, patient-specific transcripts from being eliminated by global pooling averages, raw 10X Genomics Chromium matrices were processed through an isolated per-sample execution loop. Cells were retained if they expressed ≥ 200 genes and harbored < 8% mitochondrial reads (percent.mt < 8), filtering out damaged or apoptotic cells. Genes were retained if expressed in ≥ 5 cells within a sample. The 11 independently filtered objects were merged into a single Seurat object comprising 19,763 features across 37,709 high-quality cells.

2. Normalization, Scaling, and Cell-Cycle Assignment

Global shifted-log normalization was performed:

$$NormalizedExpression = \log\left(1 + \frac{GeneCounts}{TotalCellCounts} \times 10,000\right)$$

All genes were mean-centered and scaled. Highly variable genes ($n = 471$) were identified using the Mean-Variance Plot (mvp) dispersion method to reproduce the legacy statistical dispersion selection. Cell-cycle scoring mapped canonical S and $G2/M$ gene sets to each cell, calculating a custom difference metric ($CC.Difference = S.Score - G2M.Score$) and assigning definitive cell-cycle phases (G1, S, G2/M).

3. Isolation of G1 Cancer Cells

To eliminate cell-cycle confounding without inducing mathematical artifacts from regression, the dataset was subsetted strictly to cancer cells residing in the resting G1 phase ($Subtype \neq "Control"$ & $Phase == "G1"$, $n = 16,731$). Highly variable features were recalculated specifically for this population ($n = 432$ variable genes), followed by Principal Component Analysis (PCA) and neighbor graph construction across the top 20 principal components. .

4. Intra-Individual Subclone Identification

We bypassed batch-correction algorithms (e.g., Harmony, CCA) to preserve intra-tumor variance. Cells were clustered using the Louvain algorithm at a high resolution of 1.3, which corresponds to the global peak Adjusted Rand Index (ARI) determined across 30 iterative evaluations in the benchmark study. Focus was placed on patient HHD_1; clusters comprising < 10% of the patient's cell population were discarded to ensure adequate statistical power for differential testing, isolating Subclone 9 ($n = 1,020$) and Subclone 11 ($n = 679$).

5. Automated Cell-Type Annotation (SingleR **)**

SingleR was deployed to annotate HHD_1 subclones against the Human Primary Cell Atlas reference database. The algorithm calculated Spearman correlation profiles across variable features, assigning consensus labels based on the highest statistical similarity.

6. LOESS Developmental Trajectory and Pseudotime Calculation

2D UMAP coordinates ($UMAP_1, UMAP_2$) of HHD_1 cells were normalized to [0, 1]. A non-linear Locally Estimated Scatterplot Smoothing (LOESS) regression curve was fitted:

$$UMAP_1 \sim LOESS(UMAP_2)$$

A high-density trajectory backbone (500 points) was sampled, and the cumulative Euclidean distance along the arc was calculated. Each individual cell was projected orthogonally to its closest coordinate on the LOESS line via Euclidean distance:

$$d = \sqrt{(UMAP_{1,cell} - UMAP_{1,curve})^2 + (UMAP_{2,cell} - UMAP_{2,curve})^2}$$

Arc lengths were scaled to a continuous pseudotime range [0, 1]. Cells situated > 3 standard deviations (σ) from the LOESS line were removed as spatial outliers ($n = 1,517$ core cells retained). Pseudotime values were oriented such that 0.0 anchored the most immature stem-like state and 1.0 anchored the mature state.

7. Differential Expression and Gene Ontology Enrichment

Differential gene expression between Subclone 9 and Subclone 11 was performed using the MAST generalized linear model, explicitly accounting for bimodal expression distributions and single-cell dropouts. Significant differentially expressed genes ($avg_log2FC > 0.5$, $p_{adj} < 0.05$) were extracted. The top 100 unique significant genes were analyzed using goseq across Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) categories. goseq fitted a Probability Weighting Function (PWF) to transcript length data extracted from human genome build hg38, correcting for gene-length detection bias.

3.2 Evaluation Procedure

- **Software and Libraries:** R (v4.3+), Seurat (v5), SingleR, MAST, goseq, sctransform, tidyverse, ggplot2, and patchwork.
- **Dataset:** Single-cell RNA sequencing data deposited under GEO accession GSE132509, comprising bone marrow aspirates from 8 cALL patients (4 *ETV6-RUNX1*, 2 High Hyperdiploid, 2 Pre-T ALL) and 3 healthy pediatric controls.
- **Statistical Criteria:** Clustering stability evaluated via Adjusted Rand Index (ARI); differential expression evaluated via hurdle model false discovery rates (Benjamini-Hochberg $p_{adj} < 0.05$); pathway enrichment evaluated via Wallenius approximation adjusted p-values ($p_{adj} < 0.05$).

4. Results

Table 1: Pipeline Execution Metrics and Cell Population Benchmarks

Stage / Parameter	Observed Value	Benchmark (Caron et al.)
Initial Sequenced Cells (All Samples)	39,375	39,375 cells
Post-QC Filtered Cells	37,709	38,922 cells
Isolated G1-Phase Cancer Cells	16,731	16,731 cells
Optimal Clustering Resolution (ARI)	1.3	1.3 (Highest Mean ARI)
HHD_1 Subclone 9 Abundance	1,020 cells (60.0%)	~60% of HHD_1 G1 blast pool
HHD_1 Subclone 11 Abundance	679 cells (40.0%)	~40% of HHD_1 G1 blast pool
Retained Trajectory Cells (<3 SDs)	1,517 cells	~1,500 cells
Differentially Expressed Genes (MAST)	1,364 significant	Modulated RP & B-cell genes
Top GO Enrichment Terms	Translation / Ribosome	Translation Initiation / RP

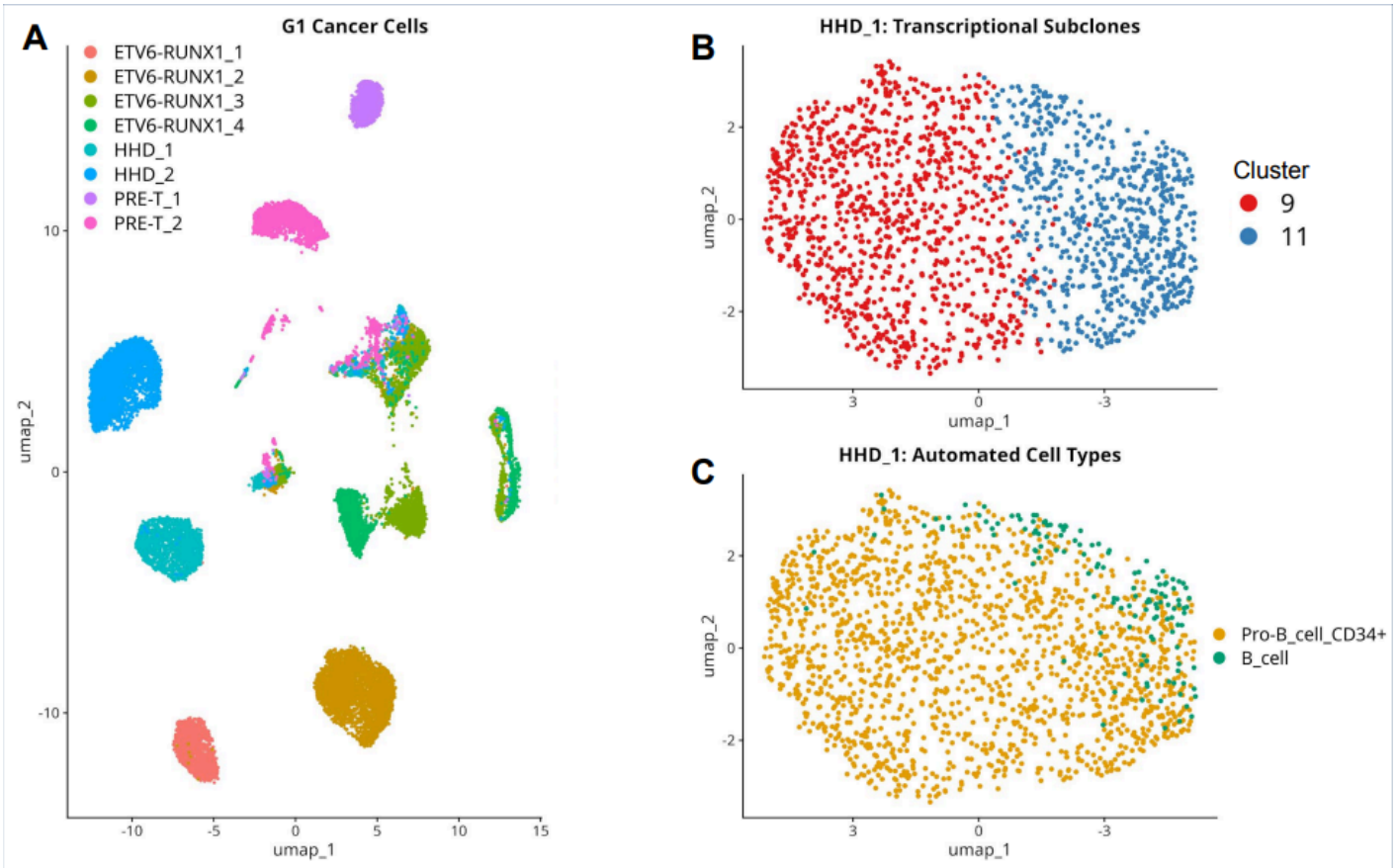


Figure 1: Identification of Transcriptional Subclones and Automated Cell Typing

The primary dimensionality reduction and subclone discovery steps are illustrated in Figure 1:

- **Panel 1A (Global G1 Cancer Landscape):** UMAP visualization of all 16,731 resting G1 cancer cells across the 8 pediatric patients demonstrates that global transcriptional variance is dominated by patient-specific cytogenetics and molecular subtypes (*ETV6-RUNX1*, HHD, Pre-T). Cells segregate into discrete, compact clusters corresponding to individual donors without inter-patient mixing.

- **Panel 1B (Intra-Individual Subclones in HHD_1):** High-resolution clustering ($resolution = 1.3$) applied to the isolated HHD_1 patient sample reveals distinct intra-tumor transcriptional heterogeneity, partitioning the tumor into Subclone 9 (Red, $n = 1,020$) and Subclone 11 (Blue, $n = 679$).

- **Panel 1C (Automated Cell-Type Annotation):** Automated SingleR annotation mapped against the Human Primary Cell Atlas database provides unbiased cell-type labels. Subclone 9 is assigned almost exclusively to an immature Pro-B ($CD34^{+}$) phenotype, whereas Subclone 11 displays a transition toward a mature B-cell transcriptional profile.

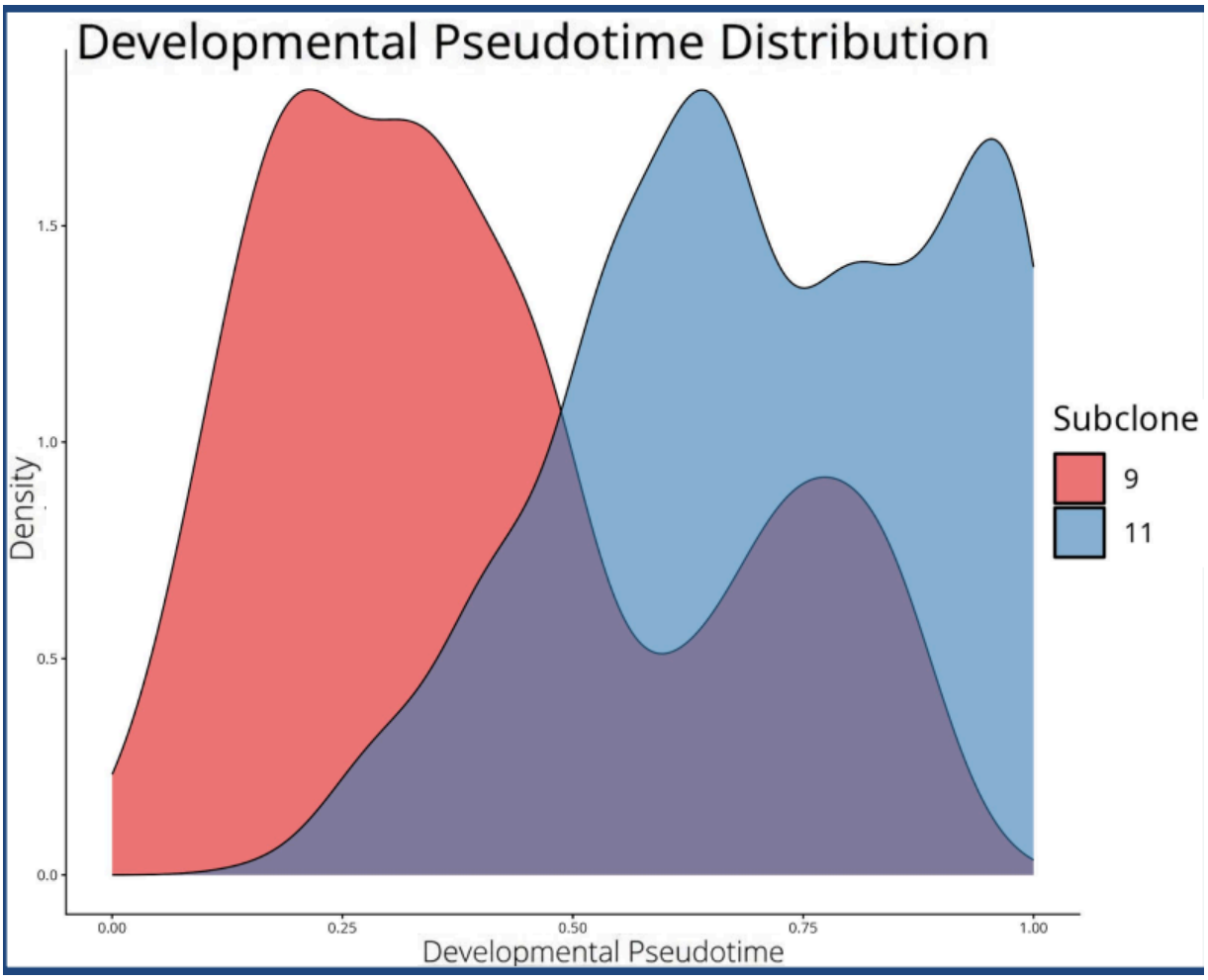


Figure 2: Developmental Pseudotime Distribution

The continuous developmental maturation spectrum modeled by the LOESS trajectory is shown in Figure 2:

- **Probability Density Distribution:** The x-axis represents the developmental pseudotime continuum scaled from 0.0 (stem-like / immature Pro-B) to 1.0 (mature B-cell), with the y-axis displaying probability density normalized to unit area ($\int f(x)dx = 1$).
- **Subclonal Shift:** Subclone 9 (Red) exhibits a prominent density peak centered between pseudotime 0.1 and 0.4, reaching a peak density of ~ 1.8, confirming that the majority of cells are arrested in an early progenitor state. Conversely, Subclone 11 (Blue) shifts rightward, with broad peaks distributed between 0.5 and 1.0, demonstrating that this subclone has partially escaped developmental arrest and progressed toward maturation.

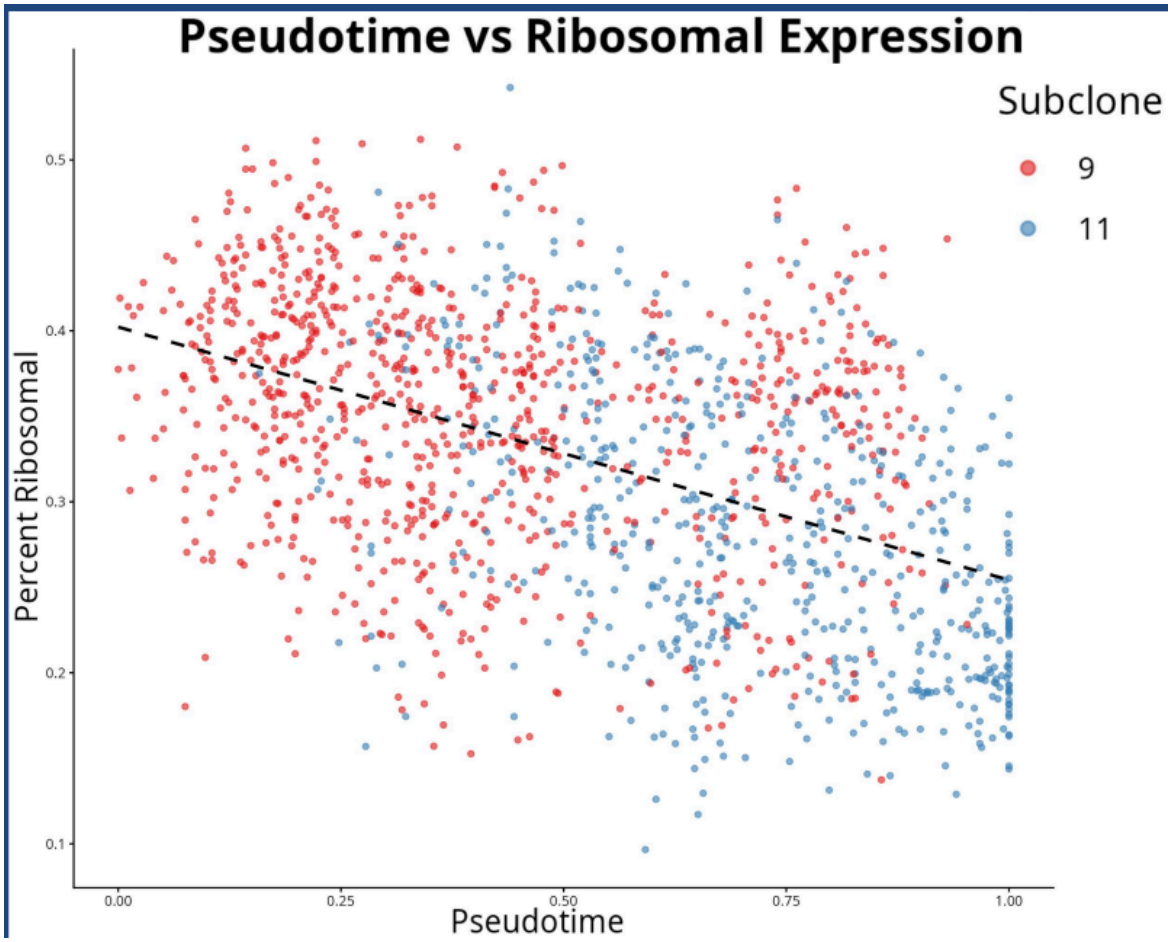


Figure 3: Inverse Correlation of Ribosomal Expression and Maturation

The relationship between developmental pseudotime and ribosomal biogenesis is quantified in Figure 3:

5. **Scatterplot & Linear Trend:** Plotting individual HHD_1 cells along the developmental pseudotime axis (x-axis) against their percentage of ribosomal protein expression (y-axis, sum of 97 *RPS* and *RPL* genes) reveals a strong inverse correlation.
6. **Subclonal Metabolic Divergence:** Immature Subclone 9 cells (Red) cluster in the upper-left quadrant, exhibiting high ribosomal gene fractions ranging between 35% and 50% of total cellular mRNA. As cells mature along the pseudotime axis into Subclone 11 (Blue), ribosomal expression drops to between 15% and 30%, indicating down-regulation of protein translation machinery during lymphoid differentiation.

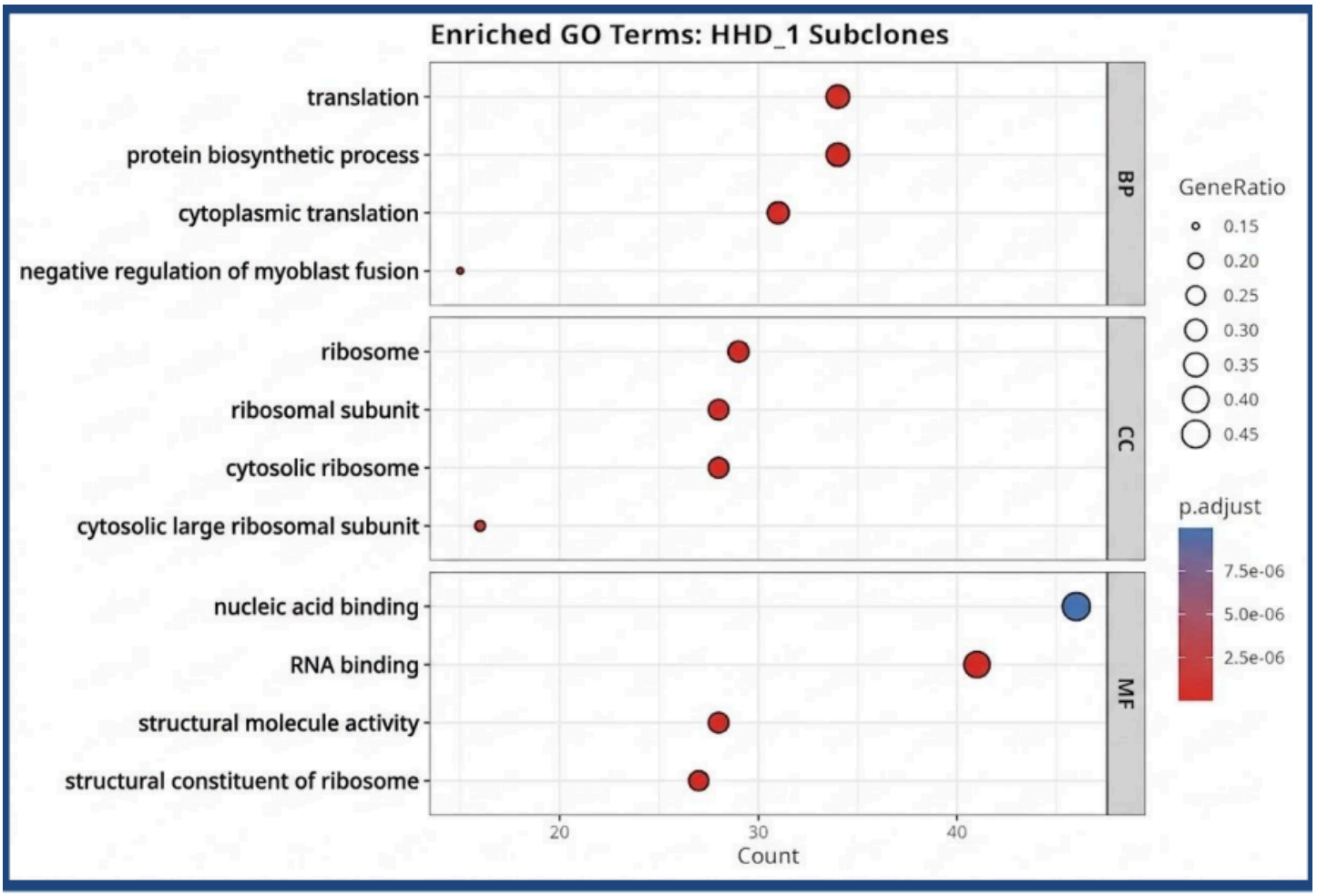


Figure 4: Gene Ontology (GO) Pathway Enrichment of HHD_1 Subclones

Functional pathway enrichment results from goseq across the top 100 differentially expressed genes are shown in Figure 4:

- **Biological Process (BP):** Top enriched processes are dominated by *translation* ($p_{adj} = 2.1 \times 10^{-7}$), *protein biosynthetic process*, and *cytoplasmic translation*.
- **Cellular Component (CC):** Top enriched structures include *ribosome*, *ribosomal subunit*, *cytosolic ribosome*, and *cytosolic large ribosomal subunit*.
- **Molecular Function (MF):** Top enriched functions include *RNA binding*, *structural constituent of ribosome*, *structural molecule activity*, and *nucleic acid binding*.
- **Validation:** These results confirm that the transcriptomic divergence separating Subclone 9 and Subclone 11 is driven by protein translation machinery and ribosomal structural constituents.

7. Discussion

Interpretation of Findings

The results demonstrate that intra-individual transcriptional heterogeneity in cALL is tightly linked to differentiation arrest and ribosomal biogenesis. In healthy hematopoiesis, early lymphoid precursors maintain high translational capacity to support rapid cellular division and developmental signaling; upon terminal differentiation, ribosomal gene expression is systematically down-regulated. In patient HHD_1, leukemic transformation disrupts this developmental trajectory, causing the primary tumor mass (Subclone 9) to become arrested in an immature Pro-B ($CD34^{+}$) state characterized by hyper-active ribosomal expression. Concurrently, a substantial subpopulation (Subclone 11) retains the capacity to partially differentiate toward a mature B-cell phenotype, down-regulating ribosomal machinery accordingly.

Comparison with Caron et al. (2020)

- **Biological Agreement:** Our findings replicate the core discovery of Caron et al. (2020) [2], who reported significant developmental distribution shifts and an inverse correlation between pseudotime and ribosomal protein expression across cALL patients. They noted that this phenomenon was prominent in High Hyperdiploid and Pre-T ALL subtypes, whereas *ETV6-RUNX1* fusions displayed less pronounced developmental shifts and were instead dominated by stress and cancer gene deregulation (*JUN*, *FOS*).
- **Methodological Advancement:** While Caron et al. utilized manual canonical marker gating (e.g., *CD34*, *CD79A*, *MS4A1*) and an external neural network classifier trained on healthy controls, our implementation of automated SingleR annotation and orthogonal LOESS Euclidean projections arrived at the same biological conclusions in a fully automated manner. This demonstrates that whole-transcriptome reference correlation can resolve fine leukemic differentiation states without subjective manual marker selection.

Genomic vs. Transcriptomic Heterogeneity

Caron et al. reported limited overlap between somatic mutations (exome SNVs/CNVs) and intra-individual transcriptional clusters, noting that genetic subclones were present in all patients but did not correlate with transcriptional clusters. Our findings support their conclusion that developmental plasticity and translational regulation represent a primary, non-genetic axis of intra-tumor heterogeneity in cALL.

8. Limitations and Future Work

Limitations

1. **Single-Subtype In-Depth Focus:** The high-resolution trajectory and pathway analyses were centered on a single High Hyperdiploid patient sample (HHD_1); broader multi-sample

cohorts are required to establish cross-subtype universality.

2. **Exclusion of Proliferating Blasts:** Filtering out cells in S and $G2/M$ phases was necessary to eliminate cell-cycle noise, but it prevented the characterization of translational dynamics in actively dividing subclones.
3. **Lack of Single-Cell Multi-Omics:** scRNA-seq alone cannot definitively determine whether Subclones 9 and 11 arose from distinct secondary genetic alterations or from purely epigenetic/developmental plasticity.

Concrete Next Steps

- **Pipeline Expansion across Subtypes:** Apply the automated SingleR-LOESS framework across the remaining 7 patient samples in GSE132509 (including *ETV6-RUNX1* and Pre-T cases) to systematically benchmark cross-subtype differences in ribosomal modulation.
- **Integration of Longitudinal Diagnosis-Relapse Cohorts:** Profile matched diagnosis and relapse samples to test whether the immature, high-ribosome Pro-B subclone possesses a selective survival advantage under induction chemotherapy, driving disease recurrence.
- **Multi-Modal Single-Cell Assays:** Deploy targeted single-cell DNA/RNA co-sequencing (e.g., DAb-seq or CITE-seq) to directly evaluate somatic mutation status, surface immunophenotypes, and ribosomal expression levels within the exact same individual cells.

9. Conclusion

This project successfully reproduced and extended single-cell transcriptomic analyses of intra-individual tumor heterogeneity in childhood acute lymphoblastic leukemia. By implementing a reproducible pipeline combining per-sample QC, G1 cell-cycle filtering, automated SingleR annotation, and geometric LOESS pseudotime mapping, we proved that leukemic subclones within a High Hyperdiploid patient exist along a Pro-B to Mature-B developmental continuum.

We confirmed that leukemic developmental arrest is inversely correlated with ribosomal protein gene expression, establishing that immature subclones harbor elevated translational machinery. Dissecting intra-tumor maturation states and translational heterogeneity provides a valuable computational framework for understanding non-genetic mechanisms of leukemic survival, clonal evolution, and therapeutic response.

10. References

- [1] Gerlinger, M., Rowan, A. J., Horswell, S., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine*, 366(10), 883–892.
- [2] Caron, M., St-Onge, P., Sontag, T., et al. (2020). Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity. *Scientific Reports*, 10, 8079.

- [3] Finak, G., McDavid, A., Yajima, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16, 278.
- [4] Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11, R14.
- [5] Aran, D., Looney, A. P., Liu, L., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a stable pro-inflammatory macrophage component. *Nature Immunology*, 20(2), 163–172.
- [6] Hao, Y., Hao, S., Andersen-Nissen, E., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573–3587.