RanBALL: An Ensemble Random Projection Model for Identifying Subtypes of B-cell Acute Lymphoblastic Leukemia

Lusheng Li[1], Hanyu Xiao[1], Xinchao Wu[1], Zhenya Tang[2], Joseph D. Khoury[2], Jieqiong Wang[3], and Shibiao Wan[1]*

[1]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA

[2]Department of Pathology, Microbiology and Immunology, University of Nebraska Medical Center, Omaha, NE, USA

[3]Department of Neurological Sciences, University of Nebraska Medical Center, Omaha, NE, USA

*Correspondence: Shibiao Wan, swan@unmc.edu

ORCID: https://orcid.org/0000-0003-0661-2684

Abstract

As the most common pediatric malignancy, B-cell acute lymphoblastic leukemia (B-ALL) has multiple distinct subtypes characterized by recurrent and sporadic somatic and germline genetic alterations. Identification of B-ALL subtypes can facilitate risk stratification and enable tailored therapeutic approaches. Existing methods for B-ALL subtyping primarily depend on immunophenotypic, cytogenetic and genomic analyses, which would be costly, complicated, and laborious in clinical practice applications. To overcome these challenges, we present **RanBALL** (an Ensemble **Ran**dom Projection-Based Model for Identifying **B**-Cell **A**cute **L**ymphoblastic **L**eukemia Subtypes), an accurate and cost-effective model for B-ALL subtype identification based on transcriptomic profiling only. RanBALL leverages random projection (RP) to construct an ensemble of dimension-reduced multi-class support vector machine (SVM) classifiers for B-ALL subtyping. Results based on 100 times 5-fold cross validation tests for >1700 B-ALL patients demonstrated that the proposed model achieved an accuracy of 93.35%, indicating promising prediction capabilities of RanBALL for B-ALL subtyping. The high accuracies of RanBALL suggested that our model could effectively capture underlying patterns of transcriptomic profiling for accurate B-ALL subtype identification. We believe RanBALL will facilitate the discovery of B-ALL subtype-specific marker genes and therapeutic targets, and eventually have consequential positive impacts on downstream risk stratification and tailored treatment design.

Background

B-cell Acute Lymphoblastic Leukemia (B-ALL) is a hematological malignancy that originates from the precursor B-cells of the bone marrow. As the most common acute lymphoblastic leukemia (ALL) type, B-ALL was diagnosed among 6,000 ALL patients each year especially for children younger than 5 years of age (1,2), manifests through the abnormal proliferation of immature B-cells. The clinic diagnostic and biologic heterogeneity of B-ALL present a significant challenge in terms of subtype classification and therapy stratification (3,4) for the disease. In addition, studies have also highlighted the requirement of precise subtype identification for highly diverse therapeutic approaches according to each patient (5), since they have specific responses to treatment and prognoses (6–8). So far, multiple distinct B-ALL subtypes have been characterized through recurrent and sporadic somatic and germline genetic alterations, (e.g., BCR-ABL1 (Philadelphia (Ph) chromosome), TCF3-PBX1 (9), hypodiploid (10), etc.), and the survival rates of this malignancy in children can be dramatically increased to more than 90% (11,12) with effective identification and tailored treatment of different subtypes (13). However, the heterogeneity of B-ALL presents a significant challenge in terms of subtype classification and treatment stratification (3,4). The study comprehensively reviewed the etiologic heterogeneity of childhood acute lymphoblastic leukemia across different subtypes, highlighting the critical need for further investigations into risk factors that are specific to each subtype (14). Another research focused on BCR/ABL1-like ALL, a high-risk subtype distinguished by specific genetic alterations, emphasizing the importance of refined diagnostic algorithms and the development of targeted therapies to improve treatment outcomes (15). Based on integrated genomic analysis of 1,988 childhood and adult cases, 23 B-ALL subtypes have been identified by chromosomal rearrangements (16), sequence mutations (17,18) and heterogeneous genomic alterations (19–21).

The conventional methods for B-ALL subtype identification primarily depends on a combination of morphological, immunophenotypic, cytogenetic, and molecular characteristics (22,23). Given the advancements in next-generation sequencing (NGS) (24,25), transcriptome profiling is found to be an informative tool to unveil chromosomal rearrangements in individual tumors for genetic or clinical marker discovery (21,26). The study explored practical considerations for utilizing RNA sequencing in managing B-lymphoblastic leukemia, underscoring RNA-Seq's capability to accurately assign specific molecular subtypes in the majority of patients (27). In addition, large cohort studies for new subtype detection and rapid classification with large-scale datasets raise more interest in the progress of precision medicine (12,28,29). For similar case as B-ALL under the category of leukemia, Umeda et. al (30) have identified the genomic atlas of pediatric acute myeloid leukemia (pAML) and determined 23 distinct molecular subtypes through large-scale gene alteration analysis. Although genetic quantification presents baseline parameters needed, it is difficult and costly for systematic analysis linking existing B-ALL subtypes with expression profiles (31) or classifying rare subtypes with standard laboratory tests, cause these methods typically involve integrating different forms of NGS methodologies (32) like whole-genome sequencing (WGS) (33), whole-exome sequencing (WES) (34), cytogenetic assays (35) etc. Moreover, extensive manual curation of the results is required before being considered as standard identification.

In recent years, machine learning (ML) has emerged as a powerful tool in the field of biomedical research, enabling the analysis of complex datasets and the discovery of hidden patterns. The high volume of RNA-seq data calls for cost-effective processing algorithms like machine learning to reveal the inner relationship between genomics and clinical conditions. The application of ML models to the identification of B-ALL subtypes has the potential to revolutionize our understanding of this disease and improve patient outcomes (36). Unsupervised clustering was first applied to microarrays for prediction yet had low performance considering individual heterogeneity will result in variable group assignments under different gene set definitions among different research institutions (37). In recent years, more presented machine learning tools have started to train reliable classifiers with well-defined terms of B-ALL subtype allocation from WHO-HAEM5 (38), and ICC (39) classifications before applying the model to systematic research like new biomarker detection (40) and risk parameter recognition (41) in unknown datasets. For instance, Allspice R package was developed to predict the B-ALL subtypes and driver genes based on centroid model (26). ALLSorts introduced by Schmidt et. al (42) demonstrate high accuracy and probability of subtype classification when attributing 18 previously defined groups to more than 1200 samples with logistic regression. Beder et. al (37) then expend the possibility of multi-class and novel subtype identification with ALLCatchR while underlying development trajectories of BCP-ALL. However, the evolving landscape of B-ALL subtypes has currently encompassed 26 distinct subcategories (38,39), combining a continuously expanding, not to mention those uncharted categories that hold crucial clinical significance. Under these circumstances, fast and precise computational tools adept at subtype classifying from vast and intricate datasets are needed (43).

Here we introduce **RanBALL** (an Ensemble **Ran**dom Projection-Based Model for Identifying **B**-Cell **A**cute **L**ymphoblastic **L**eukemia Subtypes), an accurate and cost-effective model for B-ALL subtype identification based on transcriptomic profiling only. High robustness and consistency were achieved in 1743 samples with 93.35% accuracy through 100 times 5-fold cross-validation. Moreover, RanBALL has superior improvement over state-of-art classifiers, which indicates that this model will have huge potential for further clinical application. It represents a significant advancement in the precision identification of B-ALL subtypes, offering a powerful tool for clinical applications. The development of RanBALL not only improve risk stratification and optimize treatment strategies but also opens new possibilities for personalized medicine in the future.

Methods

B-ALL dataset

The RNA-seq data and clinical information of B-ALL samples were obtained from St. Jude Cloud (https://pecan.stjude.cloud/static/hg19/pan-all/BALL-1988S-HTSeq.zip). The dataset includes 1988 samples that were classified as 23 B-ALL subtypes from the study (21). In data processing, samples with two subtypes and those identified as "other" categories were filtered out. Additionally, the samples were processed by referring to the classification architecture outlined in the ALLSorts classifier (42). Due to the limited number of samples in subtypes "ZNF384-like" and "KMT2A-like" that could potentially compromise the effectiveness of the model training, they were grouped together with subtypes "ZNF384" and "KMT2A" into categories "ZNF384 Group" and "KMT2A Group", respectively. Samples classified as the “CRLF2(non-Ph-like)” subtype were excluded, as their identification is more appropriately addressed through alternate analysis. Finally, the B-ALL dataset contains a total of 1,743 samples across 20 distinct categories. The pie chart (Fig. 1A) illustrates the distribution of these samples among the various subtypes. The age distribution and numbers for different age group of B-ALL dataset is shown in the Fig. 1B. The distribution showed a higher concentration of B-ALL cases in younger age groups, with notable peaks in childhood and young adulthood.

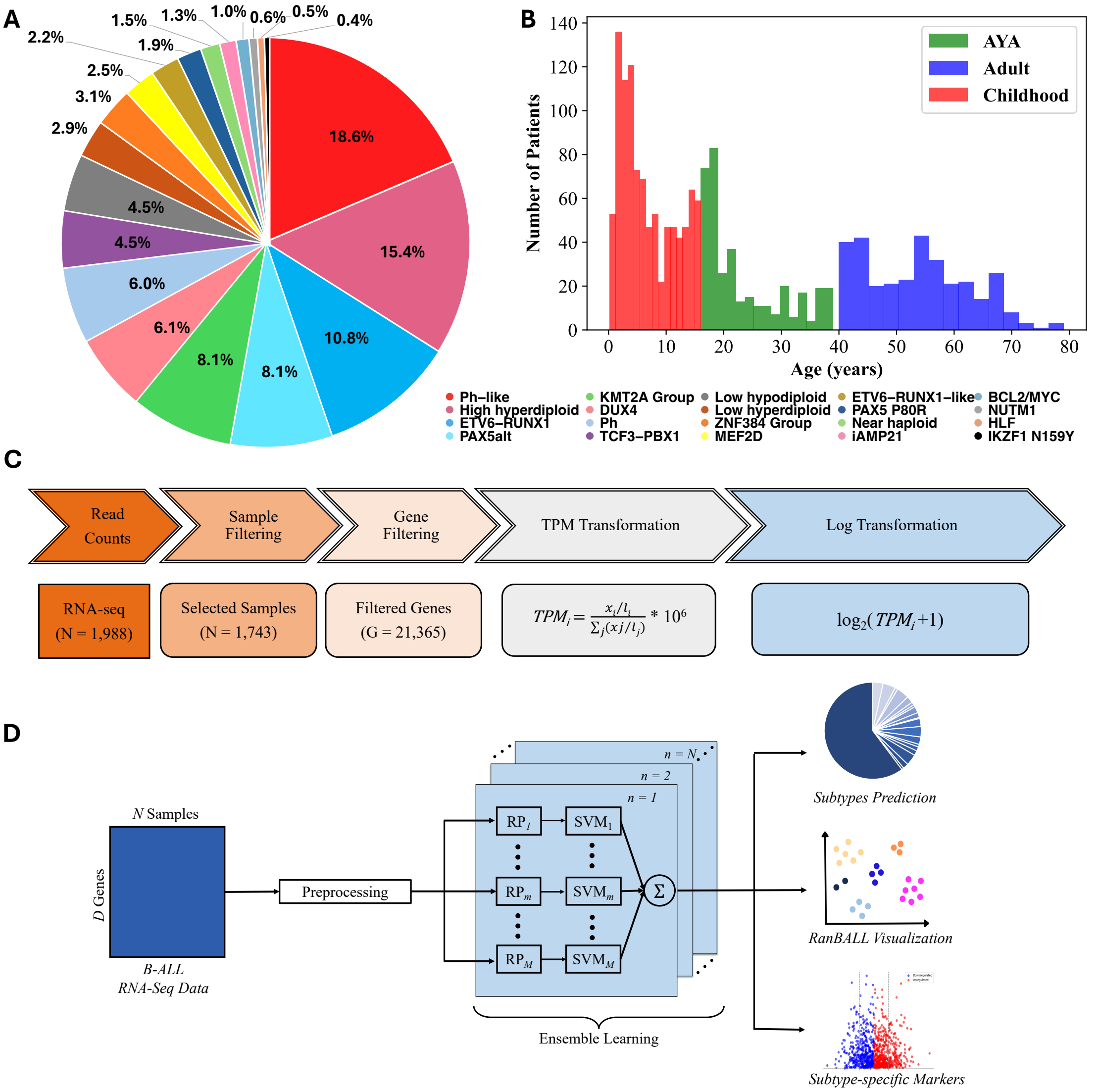


Figure 1. Overview of B-ALL subtype identification study using RanBALL framework. (A) B-ALL dataset composition. The pie chart shows the distribution of 1,743 B-ALL samples across 20 molecular subtypes, each represented by a distinct color. Percentages reflect the relative prevalence of each subtype within the dataset. **(B)** The age distribution and numbers for different age group of B-ALL dataset. Age distribution across B-ALL patients. The histogram illustrates the number of patients within each age group across three categories: childhood (red), adolescent and young adult (AYA, green), and adult (blue). **(C)** Transcriptomic data preprocessing pipeline. The flowchart outlines the multi-step preprocessing applied to the RNA-seq data, starting with raw read counts and ending with log-transformed TPM values for 21,365 genes from 1,743 selected samples. **(D)** The

framework of RanBALL. The preprocessed data is dimensionally reduced using random projection (RP), and an ensemble of multi-class Support Vector Machines (SVMs) is trained on multiple reduced matrices. The symbol m represents the m-th reduced-dimensional data matrix. We predefined dimension of 1000 in this framework. The symbol n indicates the n-th predicted subtype. The RanBALL possesses the capability to predict 20 distinct subtypes. The final prediction is an aggregated output from the ensemble. In addition to subtype prediction, RanBALL supports enhanced visualization of subtype clusters and identification of subtype-specific markers.

The RanBALL framework

RanBALL is an ensemble-based model, designed to assist healthcare professionals in accurately identifying B-ALL subtypes using RNA-seq data (Fig. 1D). Leveraging the random projection and SVM techniques, our current model enables to identify accurately and efficiently 20 distinct B-ALL subtypes, which could provide reliable diagnostic insights that can significantly aid clinical decision-making processes. The RanBALL model accepts different types of gene expression data as input data, including gene raw counts, Fragments Per Kilobase of transcript per Million mapped reads (FPKM) and Transcripts Per Million (TPM). The different data types would be uniformly transformed into $\log_2$(TPM +1) for predicting the B-ALL subtypes. Following data preprocessing and normalization, RanBALL conducts random projection to lower data dimensions. Multi-class SVM models serve as classifiers on the reduced-dimensional data in each iteration. Finally, ensemble predictions are generated by averaging probabilities across multiple runs, yielding the highest probability subtype prediction for each sample.

Data Preprocessing

The data preprocessing steps are illustrated as Fig. 1C. For the raw gene expression counts of 1988 B-ALL samples, only the gene expressed in at least 75% of the samples were retained, resulting in and final 21635 of the 52007 original genes were kept. The gene Ensembl IDs were kept in the study. Subsequently, we normalized the raw read counts to Transcripts Per Million (TPM). The total exon length of gene was calculated as effective length of the gene and the information of gene exons was extracted from the gtf file (http://ftp.ensembl.org/pub/release-109/gtf/homo_sapiens/Homo_sapiens.GRCh38.109.gtf.gz). After sequencing depth normalization, TPM values were log-transformed using the formula $\log_2$(x+1). Ultimately, the B-ALL subtype clinical information was combined with the log-transformed TPM for subsequent training and analysis.

Random Projection

Random Projection is a dimensionality reduction technique that aims to reduce the dimensionality of high-dimensional data while approximately preserving pairwise distances between data points. It is based on the Johnson–Lindenstrauss lemma (44). The Johnson–Lindenstrauss lemma provides a theoretical justification that a high-dimensional dataset can be approximately projected into a low-dimensional space while approximately preserving pairwise distances between data points. Specifically, the original D-dimensional data are projected onto a d-dimensional subspace through multiplying the original D-dimensional data matrix by the $d \times N$ random projection matrix. Namely,

$$\mathbf{A} = \frac{1}{\sqrt{d}}\mathbf{RT} \in \mathbb{R}^{d\times N}, \quad \mathbf{T} \in \mathbb{R}^{d\times N}, \quad \mathbf{R} \in \mathbb{R}^{d\times D} \tag{1}$$

The random projection matrix $\mathbf{R}$ should conform to any distributions with zero mean and unit variance, so that the random projection matrix $\mathbf{R}$ will give a mapping that satisfies the Johnson–Lindenstrauss lemma. In the study, the matrix $\mathbf{T}$ represents the original transcriptomic dataset, with D corresponding to the number of gene Ensembl ID and N denoting the number of B-ALL samples. For computational efficiency and the requirement of sparseness, we implemented a highly sparse RP method (45) This method determines the elements of $\mathbf{R}$ (i.e., $\mathbf{r}_{i,j}$) as follows:

$$r_{i,j} = \sqrt{p}\begin{cases} 1, & with\, probability\ \frac{1}{2p}, \\ 0, & with\, probability\ 1 - \frac{1}{p}, \\ -1, & with\, probability\ \frac{1}{2p}, \end{cases} \qquad \begin{matrix} \text{where } i = \{1, \dots, d\}, \\ j = \{1, \dots, D\} \end{matrix} \tag{2}$$

In accordance with the recommendation (45), we selected $p = \sqrt{D}$.

Ensemble RP Model

After data preprocessing, the transcriptomic profiling of B-ALL samples was projected to low dimensional space by random projection. To obtain reliable and robust performance, we selected 30 subspace dimensions 1000. The transformed low dimensional data matrix was used for training an ensemble of multi-class support vector machine (SVM) classifiers, each corresponding to one of the RP matrices of various dimensions. In the training process, the "linear" kernel was chosen in the SVM classifier. To develop a robust model, we ensembled the predicted probability scores of each B-ALL subtype for different low-dimensional data matrix and obtained an ensemble model. Fig. 2B shows that the ensemble method has better and stable performance than individual method. The ensemble score $\boldsymbol{S_m^{en}}$ for each subtype was calculated by averaging all the prediction probability scores from each m-th SVM model in the ensemble:

$$S_m^{en} = \frac{1}{M}\sum_{m=1}^{M}\sum_{\gamma\in S_m}\sum_{n=1}^{N}\alpha_{m,\gamma}y_{m,\gamma}\, K(\mathbf{A}, \mathbf{A}_k), \tag{3}$$

where S_m is the set of support vector indexes corresponding to the m-th SVM, $\alpha_{m,\gamma}$ are the Lagrange multipliers, N is the number of predicted subtypes, $y_{m,\gamma}$ is the class label for each subtype, $K(\cdot,\ \cdot)$ is the linear kernel function. The $\mathbf{A}$ represents the projected RNA-seq data, and the k correspond to the B-ALL sample. In addition, M is the ensemble size.

Performance Evaluation

This study applies 10 times 5-fold cross-validation (46) during the model training and testing. For model performance, we measure accuracy (Acc), F1-Score ($F1$), and Matthews correlation coefficient (MCC) (47) as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

True positives (TP) denote the count of samples predicted to possess the specific subtype, which aligns with clinical documentation. False positives (FP) represent the count of samples incorrectly classified into different categories. True negatives (TN) indicate the count of samples predicted as 'other' that genuinely do not belong to the specified subtype category, while false negatives (FN) refer to the count of samples predicted as 'other' but are indeed found within the specified subtype category. The F1-Score is a statistical measure used to evaluate the accuracy of a classification model, which is a way to balance the trade-off between precision and recall. A high precision might indicate a low tolerance for false positives, while a high recall might indicate a low tolerance for false negatives. The F1-Score helps to find a balance between these two factors, making it a useful metric for evaluating the overall quality of a classification model. It is particularly useful in situations where the class distribution is imbalanced. In addition, MCC is a balanced measure that takes into account true and false positives and negatives. This makes it particularly helpful in imbalanced datasets where the number of positive instances may be very different from the number of negative instances.

Visualization

RanBALL utilizes a weighted combination of two key matrices: a dimension-reduced feature matrix and a sample-to-subtype matrix derived from prediction results. The dimension-reduced feature matrix is obtained through Random Projection technique. This matrix is then normalized using Z-Score, centering and scaling the data along each dimension across samples. The prediction subtype for each sample is encoded by one-hot encoding to create a sample-to-subtype matrix, where each row corresponds to a sample, and each column represents a subtype. This matrix was then normalized using a Z-Score transformation across all samples to ensure that the data is centered and scaled, making the features comparable with the dimension-reduced matrix. These two matrices are then combined with different weights to formulate the final visualization matrix, combining the predicted subtype information with the dimensional features. We defined *w* as the weight ratio of the dimension-reduced feature matrix over the sample-to-subtype matrix. This weight can be adjusted to emphasize either the reduced feature space ($w > 1$) or the predicted subtype information ($0 < w < 1$) in the final visualization. This combined matrix serves as the input for t-SNE visualization, allowing for a more informative and potentially more biologically relevant representation of the data.

Differential gene expression analysis

Differential gene expression analysis was performed by edgeR package (3.40.2) (48). The voom method was applied to model differential gene expression. The raw counts were transformed to log2(CPM) for differential gene expression analysis. The cutoffs of FDR < 0.05, and absolute log2FC > 1 were applied to define significantly differentially expressed genes (DEGs). The heatmap plot was generated by Pheatmap package (1.0.12) (49).

Results

RanBALL applies ensemble random projection for multi-class prediction

RanBALL is an ensemble random projection-based multi-class classification model specifically designed for B-ALL subtyping using gene expression profiling. Employing random projection (RP) as its dimensionality reduction technique, RanBALL operates on gene expression data organized in a matrix format, where rows correspond to genes and columns represent cells. The processing pipeline encompasses four main steps: (1) data preprocessing and normalization, (2) RP-based dimension reduction, (3) multi-class classification, and (4) ensemble-based result determination, as depicted in Fig. 1D. In the step of data preprocessing and normalization, raw counts were converted to log-transformed Transcripts Per Million (TPM) values (Fig. 1C). This step is crucial for normalizing the data across different samples and reducing the impact of technical variations. RP is then applied to reduce the dimensionality of the processed data matrix. RP offers several key advantages that make it a valuable technique, particularly when working with high-dimensional data. First, RP provides significant computational efficiency (50), which is crucial for reducing the computational burden in large-scale datasets. Moreover, it approximately preserves the distances between data points (51), ensuring that the intrinsic data structure remains largely intact. This property allows RP to effectively maintain the relationships within the original data, even after dimensionality reduction. Finally, RP is theoretically grounded in the Johnson-Lindenstrauss lemma (45), which guarantees that the projection can preserve pairwise distances with high probability, making it both a practical and theoretically effective method for dimensionality reduction. In this process, a random matrix is generated to project the high-dimensional data onto a lower-dimensional space. The original data is multiplied by this random matrix, creating a lower-dimensional representation. We randomly generated 30 different low-dimensional representations, each with 1,000 dimensions. This multiple projection approach contributes to the ensemble nature of the model, increasing robustness and reducing the impact of any single projection. After dimensionality reduction, the multi-class SVM was trained on the reduced-dimension data to classify samples into different B-ALL subtypes. By aggregating the outcomes from various runs within the same dimension, the ensemble approach is applied to consolidate results, leading to the assignment of final prediction labels to samples. In addition, the predicted subtypes can provide additional information with the original gene expression profiling data for grouping data points in visualizations, aiding in the identification of clusters or patterns. In summary, RanBALL is particularly suited for the high-dimensional nature of gene expression data and the complex task of B-ALL subtyping, offering both accurate classification and improved visualization capabilities.

RanBALL preserves sample-to-sample distance

To explain the contribution of RP for dimension reduction in RanBALL, we investigated the degree of distortion caused by dimension reduction and compared the correlation of sample-to-sample distances after shrink with PCA (52), t-SNE (53) and UMAP (54), respectively, in different levels. We conducted Pearson correlation analysis to assess the similarities in sample-to-sample distances between the original and dimension-reduced data. As depicted in Fig. 2A, random projection achieves nearly perfect similarities in sample-to-sample distance, with correlation coefficients exceeding 0.93. For example, when reducing the data to 1000 dimensions (from 21,635 to 1000), the correlation remains high at 0.94, indicating the preservation of almost all embedded information post-dimension reduction. The remarkable performance of random projection (RP) can be attributed to several key factors. One critical factor is RP's ability to preserve pairwise distances (51), which plays a central role in maintaining high correlation coefficients between the original and projected data. This property is theoretically supported by the Johnson-Lindenstrauss lemma (45), which guarantees that a set of points in high-dimensional space can be projected onto a lower-dimensional space while approximately maintaining relative distances with high probability. Furthermore, RP's linear transformation ensures that the overall structure of the data (55), including relative distances between samples, is preserved without introducing complex non-linear distortions. This simplicity not only enhances computational efficiency but also minimizes the risk of overfitting to specific data patterns. In contrast, correlations observed with PCA, t-SNE, and UMAP are notably lower (overall below 0.67, with a minimum of 0.32). This disparity in performance can be explained by the inherent characteristics of these methods. While effective for linear dimensionality reduction, PCA focuses on preserving directions of maximum variance, potentially losing information crucial for maintaining sample-to-sample distances but not significantly contributing to overall variance. As non-linear techniques designed for dimension reduction and low-dimensional visualization, t-SNE and UMAP focus on preserving local structure and often distort global structure. These could be the reasons to explain their poor performance in preserving overall sample-to-sample distances in this context. RP's exceptional performance in preserving sample-to-sample distances while significantly reducing dimensionality makes it particularly well-suited for the high-dimensional, complex nature of gene expression data in B-ALL subtyping.

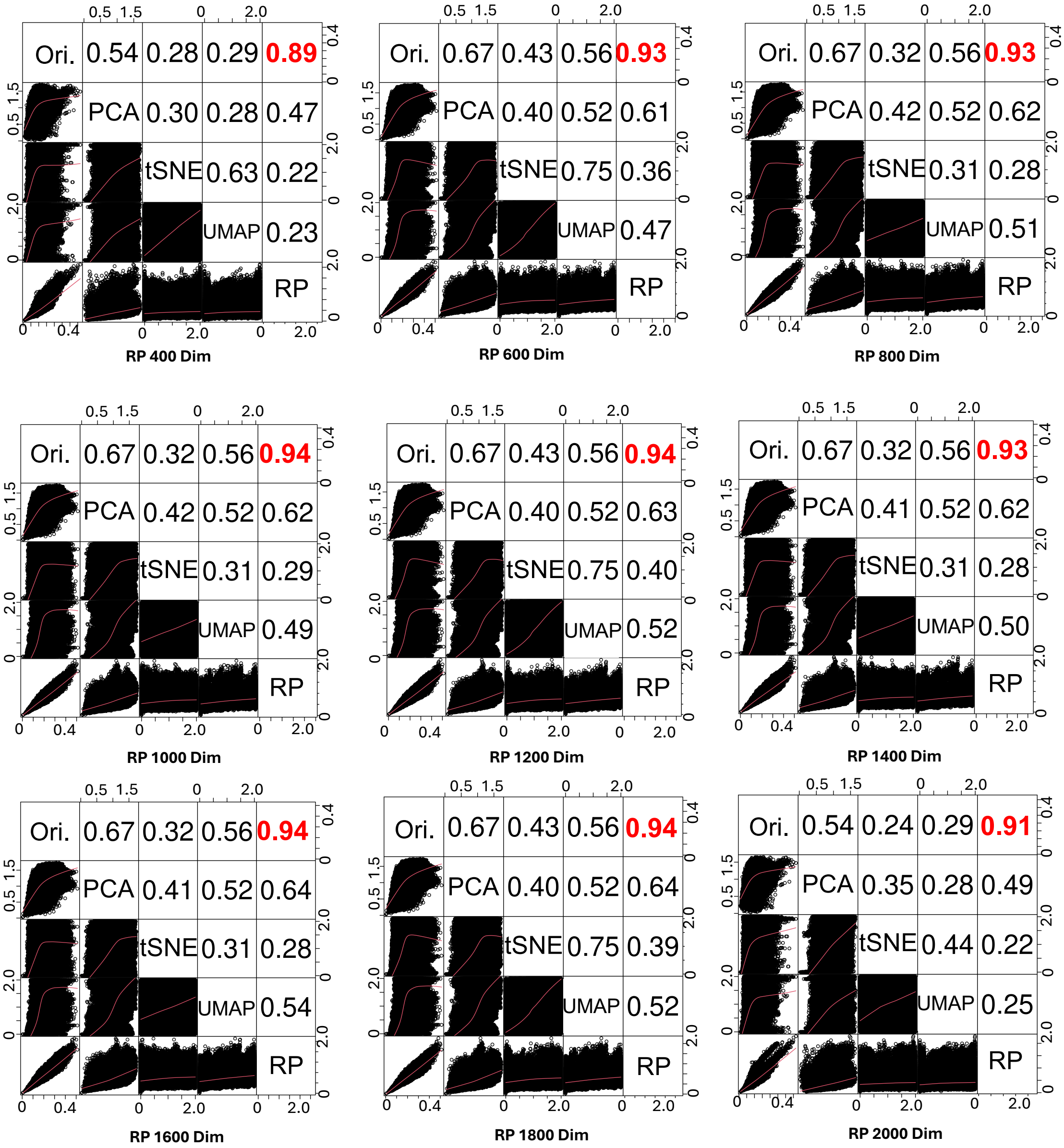


Figure 2. Comparative analysis of random projection with PCA, t-SNE, and UMAP for dimensionality reduction. This figure compares the performance of random projection (RP) with other widely used dimensionality reduction techniques across different dimensions (400 to 2000). The upper triangular section of each matrix displays the Pearson correlation coefficients (PCC) between the sample-to-sample distances in the original high-dimensional space (Ori.) and the corresponding reduced-dimensional space for each method. Higher PCC values indicate better preservation of the original data structure. RP consistently achieves higher PCCs (highlighted in red), where it outperforms PCA, t-SNE, and UMAP. The lower triangular section provides scatter plots of pairwise distances between samples before and after dimensionality reduction, illustrating how well each method preserves the relative distances between points.

Ensemble method has better performance than individual method

To ensure the robust and stable performance of RanBALL, we applied ensemble learning to the predicted results obtained after dimensionality reduction with multi-class SVM. By aggregating predictions from multiple models, ensemble methods typically lead to better performance than relying on individual models. Additionally, ensemble methods help to reduce overfitting by averaging the biases of different models, thus providing a more generalizable solution. The Fig. 3A shows the performance between ensemble and individual methods with repeated 100 times experiments. Focusing on overall accuracy metrics, the result revealed that the ensemble method's prediction exhibited greater performance and stability with statistical significance compared to individual tests across all dimensions, indicating its superiority in generating stable and trustworthy prediction outcomes. The original dimension was reduced from 400 to 2000, with an interval of 200, to test the performances of different dimensions. It also helps in finding the optimal reduced dimensionality that balances model performance and computational efficiency. The results show that there is no significant difference across conditions (Fig. 3B). Based on that the dimension of 1000 provides a substantial reduction from the original dimension while maintaining performance, 1000 was chosen for the subsequent model training. Next, we compared the performance with different ensemble sizes. Fig. 3C demonstrates that the ensemble size of 30 has better and more stable performance in term of accuracy. Based on this finding, we selected the ensemble size of 30 for the model training. The empirical approach to determining these parameters ensures that the final model configuration is well-suited to the B-ALL subtyping with complex gene expression data.

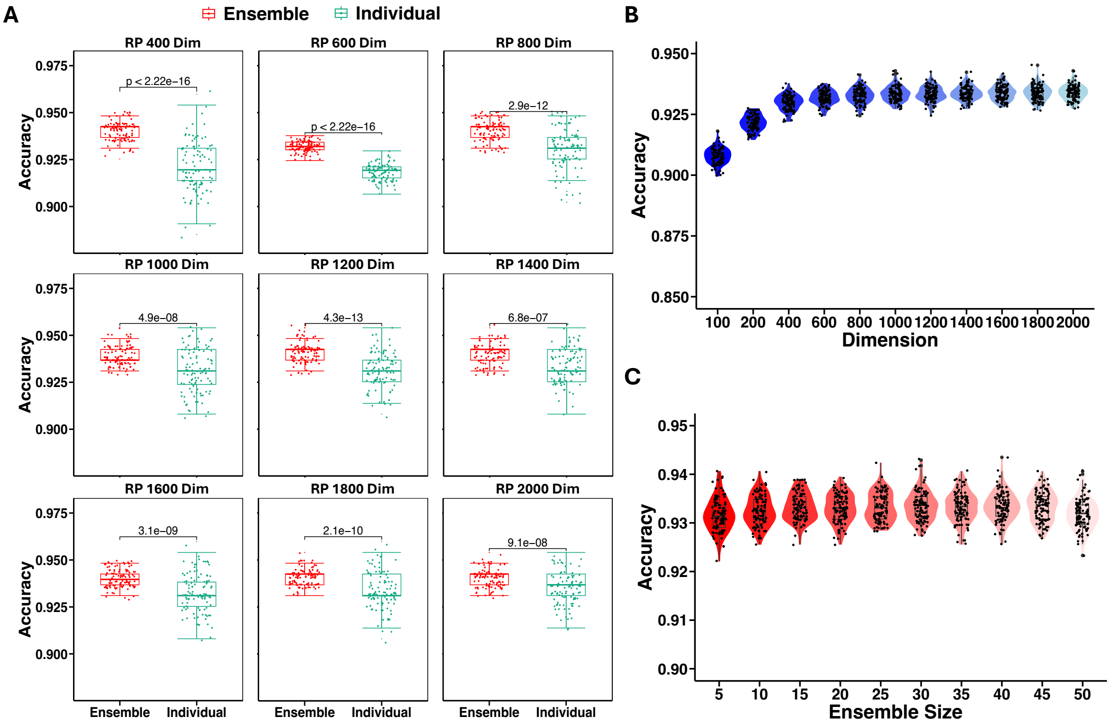


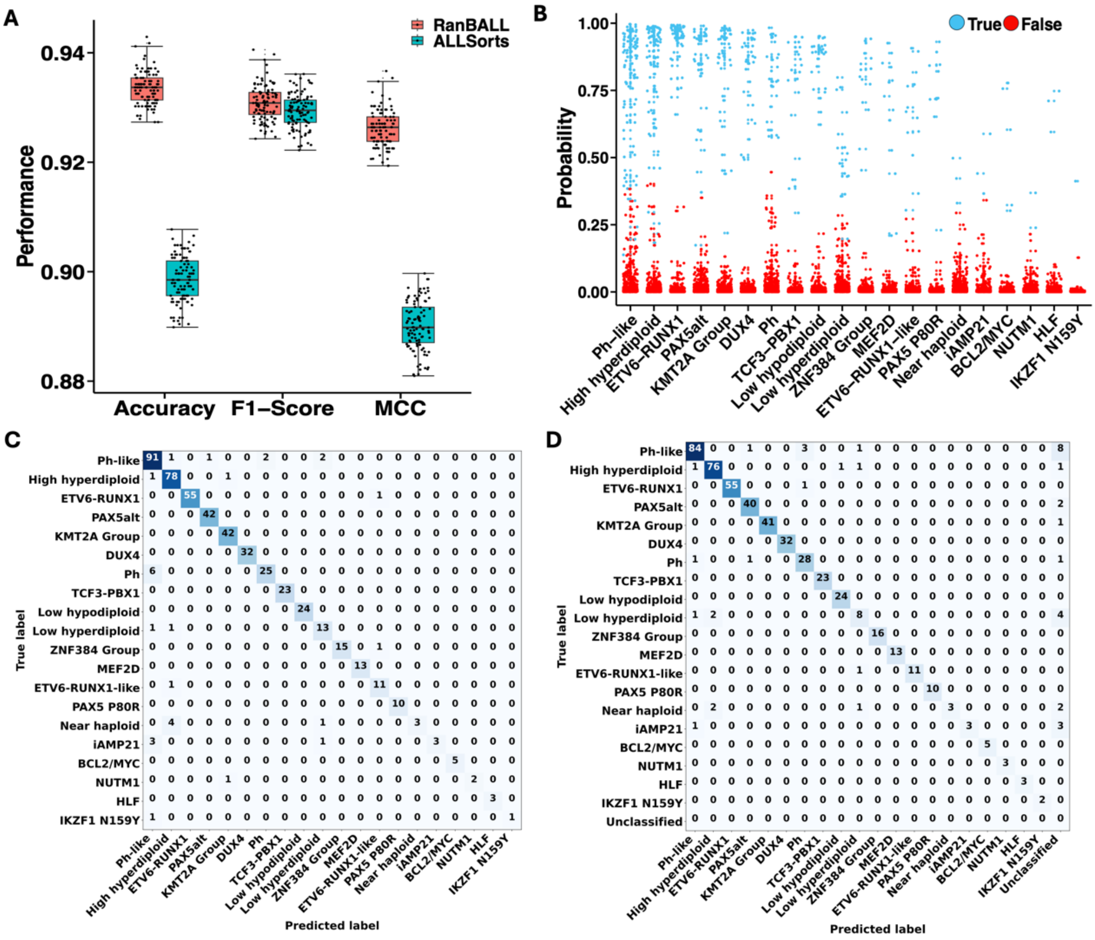
Figure 3. The performance evaluation of ensemble learning in RanBALL. (A) Comparative analysis of overall accuracy between ensemble and individual methods across different reduced dimensions. Red boxes represent the accuracy distribution of the ensemble method aggregating 30 random projections, while green boxes denote the accuracy distribution of individual classifiers on single random projections. Statistical significance was assessed using the Wilcoxon signed-rank test, with p-values displayed above each comparison. **(B)** The model performance across different reduced dimensions.

The violin plot illustrates the distribution of accuracy scores for dimensions ranging from 100 to 2000, with an interval of 200. **(C)** The model performance across different ensemble sizes. Violin plots depict the distribution of accuracy scores for ensemble sizes ranging from 5 to 50. Black dots represent individual data points, while the violin shape shows the probability density of the data.

RanBALL outperforms existing model

To assess the performance of the RanBALL model and its potential generalizability to unseen data, we employed a rigorous 10 times 5 folds cross-validation methodology on an RNA-seq dataset comprising 1743 B-ALL samples with 20 subtypes as described in Fig. 1A. Our RanBALL model yields notable average results exhibiting an accuracy of 93.35% (± 0.23%), an F1 score of 93.10% (± 0.25%) and a MCC of 92.62% (± 0.25%) (Fig. 4A). These metrics collectively offer a comprehensive evaluation of the model's efficacy. Given its exceptional performance across these metrics, the RanBALL model demonstrates significant promise for enhancing B-ALL clinical diagnosis. Additionally, we conducted a comparative analysis of the performance between RanBALL and ALLSorts (42), a well-established logistic regression classifier for B-ALL subtyping with the same data. As illustrated in Fig. 4A**,** RanBALL exhibited superior performance compared to ALLSorts in terms of Accuracy (improved by 3%), F1 Score (improved by 1%) and MCC (improved by 3%). Notably, the superior F1 score of RanBALL suggests a more balanced trade-off between precision and recall relative to ALLSorts. The MCC performance matrix offers a balanced assessment even in scenarios where classes exhibit disparate sizes, indicating that RanBALL excels particularly in multiclass classification settings with imbalanced class distributions compared to ALLSorts.

Subsequently, we applied the RanBALL model to a hold-out test set derived from the B-ALL dataset. This test set, comprised of 521 samples, generated by randomly sampling 30% of the entire B-ALL dataset. The RanBALL model demonstrated a commendable accuracy of 94.24% on this held-out test subset. The prediction probabilities of each test sample are shown in Fig. 4B**,** demonstrating the model's consistent ability to maintain high confidence levels for accurate predictions. The robust performance of the model, evidenced by high-probability predictions, underscores its proficiency in discerning intrinsic data patterns, thereby yielding confident and reliable outcomes. Notably, it exhibits the capability to deliver accurate predictions even for subtypes characterized by limited sample sizes. However, it's important to acknowledge that prediction probabilities for such subtypes may not attain exceptionally high levels. The 30% held-out test was also performed with the ALLSorts with an accuracy of 89.64% on the same test dataset. The confusion matrices are illustrated in Fig. 4C, D, provide a detailed breakdown of the model’s prediction ability for each subtype in test data. Some subtypes (9/20) have been correctly classified with no misclassifications observed for two computational models, such as PAX5alt, KMT2A, DUX4, TCF3-PBX1, Low hypodiploid, MEF2D, PAX5 P80R, BCL2/MYC and HLF Group. For some subtypes with similar characteristics and features, the model may have a certain possibility to predict the sample to be another class. For RanBALL, 2 samples were wrongly predicted as the Ph subtype in the Ph-like Group (97 samples), while 3 were wrongly classified as the Ph subtype in the Ph-like Group with ALLSorts. This situation also occurs in the subtypes related to chromosome number (Near haploid, Low hyperdiploid, and High hyperdiploid), suggesting that future research directions should improve the prediction accuracy in these subtypes with similar characteristics and features to achieve better clinical applications.



True label \ Predicted label	Ph-like	High hyperdiploid	ETV6-RUNX1	PAX5alt	KMT2A Group	DUX4	Ph	TCF3-PBX1	Low hypodiploid	Low hyperdiploid	ZNF384 Group	MEF2D	ETV6-RUNX1-like	PAX5 P80R	Near haploid	iAMP21	BCL2/MYC	NUTM1	HLF	IKZF1 N159Y
Ph-like	91	1	0	1	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0
High hyperdiploid	1	78	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ETV6-RUNX1	0	0	55	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
PAX5alt	0	0	0	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KMT2A Group	0	0	0	0	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DUX4	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ph	6	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0
TCF3-PBX1	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0
Low hypodiploid	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0
Low hyperdiploid	1	1	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0
ZNF384 Group	0	0	0	0	0	0	0	0	0	0	15	0	1	0	0	0	0	0	0	0
MEF2D	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0
ETV6-RUNX1-like	0	1	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
PAX5 P80R	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
Near haploid	0	4	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0
iAMP21	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0
BCL2/MYC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0
NUTM1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
HLF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
IKZF1 N159Y	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

True label \ Predicted label	Ph-like	High hyperdiploid	ETV6-RUNX1	PAX5alt	KMT2A Group	DUX4	Ph	TCF3-PBX1	Low hypodiploid	Low hyperdiploid	ZNF384 Group	MEF2D	ETV6-RUNX1-like	PAX5 P80R	Near haploid	iAMP21	BCL2/MYC	NUTM1	HLF	IKZF1 N159Y	Unclassified
Ph-like	84	0	0	1	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	8
High hyperdiploid	1	76	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
ETV6-RUNX1	0	0	55	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PAX5alt	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
KMT2A Group	0	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
DUX4	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ph	1	0	0	1	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	1
TCF3-PBX1	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0
Low hypodiploid	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0
Low hyperdiploid	1	2	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	4
ZNF384 Group	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0
MEF2D	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0
ETV6-RUNX1-like	0	0	0	0	0	0	0	0	0	1	0	0	11	0	0	0	0	0	0	0	0
PAX5 P80R	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
Near haploid	0	2	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	2
iAMP21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	3
BCL2/MYC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
NUTM1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
HLF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
IKZF1 N159Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4. Comprehensive performance analysis of RanBALL in comparison with ALLSorts for B-ALL subtyping. (A) Comparative performance metrics of RanBALL and ALLSorts. Accuracy, F1 Score and MCC were used for evaluating model performance. Box plots illustrate the distribution of Accuracy, F1-Score, and MCC across 100 times 5 folds cross validation. **(B)** Prediction probability distribution for the 30% held-out test set using RanBALL. Each point represents the probability of a sample (out of 521) being classified into a specific B-ALL subtype. Specifically, the blue dots indicate the specific subtype that the RanBALL model predicts to align with the categories on the horizontal axis. **(C, D)** Confusion matrices for the 30% held-out test set, comparing RanBALL (C) and ALLSorts (D) performance. Each cell shows the number of samples classified, with the diagonal representing correct classifications (True Positives). Color intensity correlates with the number of samples.

RanBALL visualizes data better than state-of-the-art methods

RanBALL demonstrates superior visualization capabilities compared to traditional methods by incorporating predicted subtype information. Specifically, the predicted subtype information for each sample was encoded using one-hot encoding and normalized by Z-score. This normalization process was applied both to the reduced dimensionality matrix and the one-hot encoded subtype information. These two matrices were then concatenated for visualization using t-SNE. We selected the t-SNE, one of the powerful and representative methods for visualizing high-dimensional data, to compare the performance of visualization.

Fig. 5A illustrates the effectiveness of RanBALL in visualizing B-ALL samples, where distinct subtypes are well-clustered, reflecting the model's capability to maintain and highlight the inherent structure in the data. This visualization allows for easy identification and interpretation of the 20 different subtypes, ranging from common subtypes like BCL2/MYC and DUX4 to rarer subtypes such as ZNF384 Group and iAMP21. Each subtype, represented by different colors and labels, forms tight, distinct clusters. In contrast, Fig. 5B presents a t-SNE visualization without the integration of predicted subtype information. This results in a less structured and more dispersed representation of the data, where subtype boundaries are less distinct and overlap more significantly. Subtypes such as High hyperdiploid, KMT2A Group, PH and Ph-like do not cluster as clearly, indicating that key relationships between subtypes may be obscured without the subtype prediction information. The difference between these two visualizations underscores the value of RanBALL's approach in enhancing the interpretability and informative visual representations of complex transcriptomic data. By leveraging predicted subtype information, RanBALL not only improves visual clarity but also potentially reveals biologically meaningful relationships between subtypes. This enhanced visualization technique could provide valuable information for researchers in identifying patterns, outliers, and potential new subgroups within B-ALL samples, ultimately leading to better understanding and classification of this complex disease.

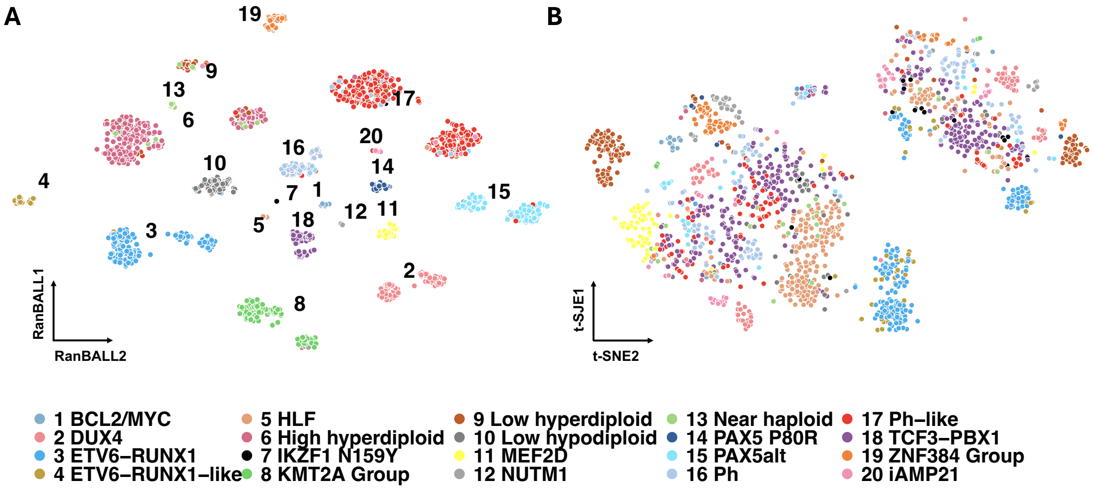


Figure 5. Comparative visualization of B-ALL subtype clustering using RanBALL-derived features and traditional t-SNE. (A) Enhanced t-SNE visualization of the reduced dimension matrix incorporating predicted subtype information. **(B)** t-SNE visualization of the reduced dimension matrix without incorporating RanBALL's predicted subtype information. The same color scheme was used in the two plots.

Differential expression analysis for B-ALL subtypes

To investigate the gene expression patterns for each B-ALL subtype, we performed differential expression analysis. Fig. 6A illustrates the differential expressed genes (DEG) between Ph-like B-ALL and the rest subtypes. The expression plots of the upregulated DEG ENAM across all B-ALL samples are shown in Fig. 6C, highlighting its specific overexpression in the Ph-like subtype. The ENAM gene was specifically expressed at the samples with Ph-like subtype. The heatmap displays the expression profiles of top 20 DEG (Fig. 6B). It indicates the potential differences among subtypes within the biological functions and processes. Among the most upregulated genes, CRLF2, one of the most important genes in Ph-like ALL, is consistent with its known role in activating JAK-STAT signaling in a subset of Ph-like cases (56,57). Other significantly overexpressed genes, including GPR110, ENAM, LDB3, and IGJ, suggesting alterations in cell adhesion, signaling, and immunoglobulin production (56,58–60). Notably, SPATS2L overexpression has been associated with poor prognosis (61,62). We also conducted differential expression analysis on the PAX5alt subtype (Fig. 6D~F). These upregulated genes may play crucial roles in promoting cell proliferation, survival, and signaling pathways in PAX5alt B-ALL. For instance, TPBG is upregulated in high-risk cytogenetic subgroups and overexpressed on the plasma membrane of lymphoblasts collected at relapse in patients with B-cell precursor ALL (63). Similarly, KSR2, a kinase suppressor of Ras 2, has been implicated in dysregulation of multiple signaling (64), suggesting a similar altered signaling pathway in PAX5alt B-ALL. Additionally, TIFAB has been shown to regulate USP15-mediated p53 signaling in stressed and malignant hematopoiesis (65). Interestingly, NFATC4 significant upregulation in PAX5alt B-ALL contrasts with its significant downregulation in Ph-like B-ALL, highlighting distinct transcriptional programs between these subtypes. For differential expression analysis between High hyperdiploid and other subtypes (Fig. 6G~I), the upregulated gene DDIT4L has been identified as therapeutic targets in PDX ALL carrying the recently described DUX4-IGH translocation (66). Notably, the upregulated gene OVCH2 was observed that it was downregulated in ALL (67,68). Additionally, S100A16 has been implicated in suppressing the growth and survival of leukemia cells in adults with Ph-negative B-ALL (69).

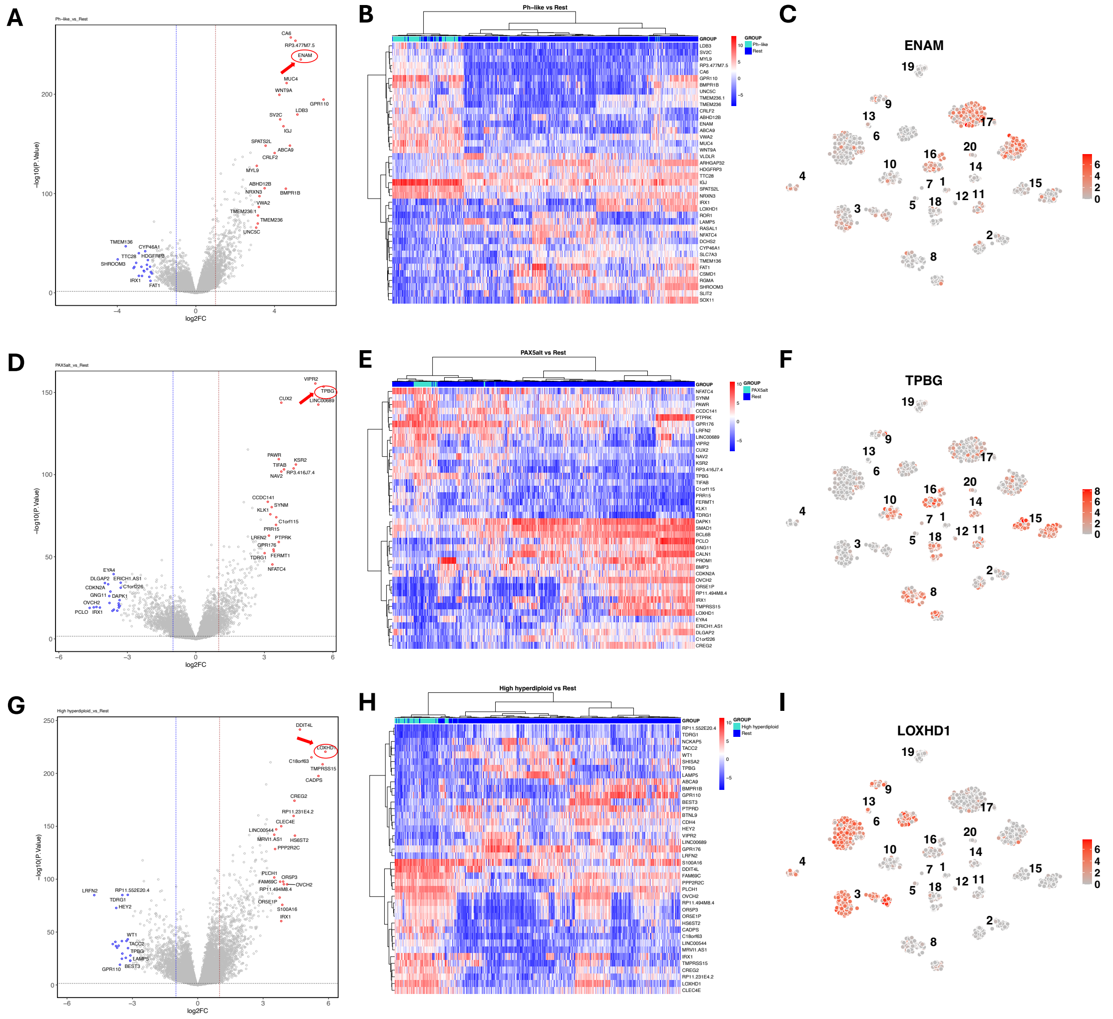


Figure 6. Differential expression analysis within B-ALL subtypes. (A, D, G) Volcano plots illustrating differential gene expression between specific B-ALL subtypes and all other subtypes. The x-axis represents log2 fold change, while the y-axis shows -log10(p-value). Red dots indicate 20 significantly up-regulated genes, blue dots represent 20 significantly down-regulated genes. Top 20 DEGs are labeled, with the most significant gene circled in red. (A) Ph-like vs. rest; (D) PAX5alt vs. rest; (G) High hyperdiploid vs. rest. **(B, E, H)** Heatmaps displaying expression patterns of the top 20 DEGs for each subtype comparison. Rows represent genes, columns represent samples. Color scale ranges from blue (low expression) to red (high expression). Hierarchical clustering dendrograms are shown for both genes and samples. Sidebar annotations indicate sample subtypes and relative level of gene expression. (B) Ph-like vs. rest; (E) PAX5alt vs. rest; (H) High hyperdiploid vs. rest. **(C, F, I)** The expression plot of the up-regulated DEG for Ph-like subtype. RanBALL plots visualizing the expression levels of the significantly up-regulated gene for each subtype across all B-ALL samples. Each point represents a sample, colored by expression intensity (red: high, grey: low). Numbers indicate different B-ALL subtypes. (C) DEG for Ph-like (ENAM); (F) DEG for PAX5alt (TPBG); (I) DEG for High hyperdiploid (LOXHD1).

Discussion

In this study, we introduced an ensemble-based model, RanBALL, which integrates Random Projection and Support Vector Machine (SVM) techniques to accurately identify B-cell Acute Lymphoblastic Leukemia (B-ALL) subtypes using solely RNA-seq data. Random Projection demonstrates efficacy in reducing the dimensionality of high-dimensional data while retaining informative features present in RNA-seq data. The experiments indicated that the ensemble method achieve superior stability and better performance than individual method. The RanBALL model runs independent from prior genomic knowledge for B-ALL subtype identification. Our results underscored the robustness of the proposed model, attaining high levels of accuracy, F1 score, and MCC value, indicating promising prediction capabilities of RanBALL for B-ALL subtyping. The application of ML models in B-ALL subtype identification demonstrates the feasibility of leveraging complex datasets to discover subtle differences among patients. This approach overcomes the limitations of traditional subtyping methods, which often rely on a limited set of markers and may not capture the full spectrum of disease heterogeneity.

Compared to existing methods for B-ALL subtyping, RanBALL consistently exhibited superior performance metrics over ALLSorts, particularly in terms of Accuracy, F1 Score and MCC value. However, there is still room for improvement in certain B-ALL subtypes, necessitating further enhancement of prediction capabilities. First, the generalizability of our findings may be limited by the composition of the training datasets, which were derived from specific patient populations. Future studies should aim to validate our models in diverse and independent cohorts to ensure their broad applicability. Second, the predictive performance of our models could be influenced by technical and biological confounders (70), such as batch effects and sample quality. Rigorous data preprocessing and quality control measures will be essential to mitigate these factors in future work. Advanced computational methods can be applied to remove the batch effects to improve the performance of model. Finally, the observed imbalance among B-ALL subtypes within the dataset may also potentially impede model performance. To address this issue, data augmentation techniques (71) can be applied to augment the representation of minority subtypes.

Additionally, future research efforts may focus on mitigating batch effects between different B-ALL clinical cohorts to better address real-world challenges and facilitate clinical applications (72). Furthermore, the integration of additional data types, such as genetic (73,74), epigenetic (75,76) and imaging data (43,77), may further enhance the accuracy and reliability of ML models in B-ALL subtype identification. The advent of single cell sequencing technologies has revolutionized our ability to dissect heterogeneity of B-ALL, enabling the characterization of cellular subpopulations and their functional states at an unprecedented resolution (78–82). The integration of multi-scale multi-omics and multi-modality can provide valuable insights into the molecular landscape of B-ALL subtypes and inform personalized therapeutic approaches.

We anticipate that the deployment of RanBALL will yield significant positive impacts on clinical diagnosis, personalized treatment strategies, and risk stratification within the realm of biomedical research and practical clinical settings. This is particularly critical as distinct B-ALL subtypes may respond differentially to various treatments, and precise subtype identification can aid clinicians in selecting the most efficacious treatment regimen for individual patients. Moreover, the diverse outcomes and survival rates associated with different B-ALL subtypes underscore the importance of accurate subtype classification. To facilitate further extending and accessibility of RanBALL, we have developed an open-source Python package, available at https://github.com/wan-mlab/RanBALL.

Acknowledgements

The authors would like to express our gratitude to St. Jude Cloud platform (https://www.stjude.cloud), which provided publicly accessible genomic data. Special thanks to all members of Dr. Wan's lab for insightful discussions. The abstract of this work was published at AACR Annual Meeting 2024 (83).

Authors' contributions

L.L.: data preprocessing, machine learning model development, data analysis and interpretation, manuscript preparation, editing, and review. H.X.: data analysis and interpretation, manuscript preparation, editing, and review. X.W.: manuscript preparation, editing, and review. Z.T.: biological and clinical expertise, manuscript editing and review. J.D.K.: biological and clinical expertise, manuscript editing and review. J.W.: manuscript editing and review. S.W.: study concept and design, manuscript editing and review.

Data availability

The RNA-seq data of B-ALL samples can be publicly accessed from St. Jude Cloud (https://pecan.stjude.cloud/static/hg19/pan-all/BALL-1988S-HTSeq.zip). The RanBALL package can be accessed at https://github.com/wan-mlab/RanBALL.

Competing Interests

The authors declare no conflict of interest.

Funding information

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA036727, and by the Office Of The Director, National Institutes Of Health of the National Institutes of Health under Award Number R03OD038391. This work was supported by the American Cancer Society under award number IRG-22-146-07-IRG, and by the Buffett Cancer Center, which is supported by the National Cancer Institute under award number CA036727. This work was supported by the Buffet Cancer Center, which is supported by the National Cancer Institute under award number CA036727, in collaboration with the UNMC/Children's Hospital & Medical Center Child Health Research Institute Pediatric Cancer Research Group. This study was supported, in part, by the National Institute on Alcohol Abuse and Alcoholism (P50AA030407-5126, Pilot Core grant). This study was also supported by the Nebraska EPSCoR FIRST Award (OIA-2044049). This work was also partially supported by the National Institute of General Medical Sciences under Award Numbers P20GM103427 and P20GM130447. This study was in part financially supported by the Child Health Research Institute at UNMC/Children's Nebraska. This work was also partially supported by the University of Nebraska Collaboration Initiative Grant from the Nebraska Research Initiative (NRI). The content is solely the responsibility of the authors and does not necessarily represent the official views from the funding organizations.

Reference:

1. Hunger Stephen P., Mullighan Charles G. Acute Lymphoblastic Leukemia in Children. N Engl J Med. 373(16):1541–52.

2. Chouvarine P, Antić Ž, Lentes J, Schröder C, Alten J, Brüggemann M, et al. Transcriptional and Mutational Profiling of B-Other Acute Lymphoblastic Leukemia for Improved Diagnostics. Cancers. 2021 Nov 12;13(22):5653.

3. Avraham Frisch, Yishai Ofran. How I diagnose and manage Philadelphia chromosome-like acute lymphoblastic leukemia. Haematologica. 2019 Oct 30;104(11):2135–43.

4. Meyers S, Alberti-Servera L, Gielen O, Erard M, Swings T, De Bie J, et al. Monitoring of Leukemia Clones in B-cell Acute Lymphoblastic Leukemia at Diagnosis and During Treatment by Single-cell DNA Amplicon Sequencing. HemaSphere [Internet]. 2022;6(4). Available from: https://journals.lww.com/hemasphere/fulltext/2022/04000/monitoring_of_leukemia_clones_in_b_cell_acute.2.aspx

5. Bassan R, Hoelzer D. Modern Therapy of Acute Lymphoblastic Leukemia. J Clin Oncol. 2011 Feb 10;29(5):532–43.

6. Mullighan CG. How advanced are we in targeting novel subtypes of ALL? Acute Leuk Myelodysplasia Adv Controv. 2019 Dec 1;32(4):101095.

7. Jeha S, Choi J, Roberts KG, Pei D, Coustan-Smith E, Inaba H, et al. Clinical Significance of Novel Subtypes of Acute Lymphoblastic Leukemia in the Context of Minimal Residual Disease–Directed Therapy. Blood Cancer Discov. 2021 Jul 1;2(4):326–37.

8. Lee SHR, Yang W, Gocho Y, John A, Rowland L, Smart B, et al. Pharmacotypes across the

genomic landscape of pediatric acute lymphoblastic leukemia and impact on treatment response. Nat Med. 2023 Jan 1;29(1):170–9.

9. Shirai R, Osumi T, Sato-Otsubo A, Nakabayashi K, Mori T, Yoshida M, et al. Genetic features of B-cell lymphoblastic lymphoma with *TCF3-PBX1* . Cancer Rep. 2022 Sep;5(9):e1559.

10. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013 Mar;45(3):242–52.

11. Roberts KG, Mullighan CG. Genomics in acute lymphoblastic leukaemia: insights and treatment implications. Nat Rev Clin Oncol. 2015 Jun;12(6):344–57.

12. Brown LM, Lonsdale A, Zhu A, Davidson NM, Schmidt B, Hawkins A, et al. The application of RNA sequencing for the diagnosis and genomic classification of pediatric acute lymphoblastic leukemia. Blood Adv. 2020 Mar 10;4(5):930–42.

13. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. The Lancet. 2008 Mar 22;371(9617):1030–43.

14. Williams LA, Yang JJ, Hirsch BA, Marcotte EL, Spector LG. Is There Etiologic Heterogeneity between Subtypes of Childhood Acute Lymphoblastic Leukemia? A Review of Variation in Risk by Subtype. Cancer Epidemiol Biomarkers Prev. 2019 May 3;28(5):846–56.

15. Płotka A, Lewandowski K. BCR/ABL1-Like Acute Lymphoblastic Leukemia: From Diagnostic Approaches to Molecularly Targeted Therapy. Acta Haematol. 2021 Nov 24;145(2):122–31.

16. Kimura S, Montefiori L, Iacobucci I, Zhao Y, Gao Q, Paietta EM, et al. Enhancer retargeting of CDX2 and UBTF::ATXN7L3 define a subtype of high-risk B-progenitor acute lymphoblastic leukemia. Blood. 2022 Jun 16;139(24):3519–31.

17. Roll JD, Reuther GW. CRLF2 and JAK2 in B-Progenitor Acute Lymphoblastic Leukemia: A Novel Association in Oncogenesis. Cancer Res. 2010 Sep 29;70(19):7347–52.

18. Gough SM, Goldberg L, Pineda M, Walker RL, Zhu YJ, Bilke S, et al. Progenitor B-1 B-cell acute lymphoblastic leukemia is associated with collaborative mutations in 3 critical pathways. Blood Adv. 2017 Sep 8;1(20):1749–59.

19. Cox CV, Evely RS, Oakhill A, Pamphilon DH, Goulden NJ, Blair A. Characterization of acute lymphoblastic leukemia progenitor cells. Blood. 2004 Nov 1;104(9):2919–25.

20. Roberts KG, Mullighan CG. The Biology of B-Progenitor Acute Lymphoblastic Leukemia. Cold Spring Harb Perspect Med [Internet]. 2020 Jul 1;10(7). Available from: http://perspectivesinmedicine.cshlp.org/content/10/7/a034835.abstract

21. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven

subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet. 2019 Feb;51(2):296–307.

22. Chiaretti S, Zini G, Bassan R. DIAGNOSIS AND SUBCLASSIFICATION OF ACUTE LYMPHOBLASTIC LEUKEMIA. Mediterr J Hematol Infect Dis. 2014 Oct 24;6(1):e2014073.

23. Mrózek K, Harper DP, Aplan PD. Cytogenetics and Molecular Genetics of Acute Lymphoblastic Leukemia. Recent Prog Treat Acute Lymphoblastic Leuk. 2009 Oct 1;23(5):991–1010.

24. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child-Educ Pract. 2013;

25. Dahui Qin. Next-generation sequencing and its clinical application. Cancer Biol Med. 2019 Feb 1;16(1):4.

26. Mäkinen VP, Rehn J, Breen J, Yeung D, White DL. Multi-Cohort Transcriptomic Subtyping of B-Cell Acute Lymphoblastic Leukemia. Int J Mol Sci. 2022 Apr 20;23(9):4574.

27. Ni Chin WH, Li Z, Jiang N, Lim EH, Suang Lim JY, Lu Y, et al. Practical Considerations for Using RNA Sequencing in Management of B-Lymphoblastic Leukemia: Malaysia-Singapore Acute Lymphoblastic Leukemia 2020 Implementation Strategy. J Mol Diagn. 2021 Oct 1;23(10):1359–72.

28. Li JF, Dai YT, Lilljebjörn H, Shen SH, Cui BW, Bai L, et al. Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. Proc Natl Acad Sci. 2018 Dec 11;115(50):E11711–20.

29. Coccaro N, Anelli L, Zagaria A, Specchia G, Albano F. Next-Generation Sequencing in Acute Lymphoblastic Leukemia. Int J Mol Sci. 2019;20(12).

30. Umeda M, Ma J, Westover T, Ni Y, Song G, Maciaszek JL, et al. A new genomic framework to categorize pediatric acute myeloid leukemia. Nat Genet. 2024 Feb;56(2):281–93.

31. Mullighan CG. Genomic profiling of B-progenitor acute lymphoblastic leukemia. Adv Controv Biol Ther Acute Leuk Myelodysplasia. 2011 Dec 1;24(4):489–503.

32. Grada A, Weinbrecht K. Next-Generation Sequencing: Methodology and Application. J Invest Dermatol. 2013 Aug 1;133(8):1–4.

33. Ng PC, Kirkness EF. Whole Genome Sequencing. In: Barnes MR, Breen G, editors. Genetic Variation: Methods and Protocols [Internet]. Totowa, NJ: Humana Press; 2010. p. 215–26. Available from: https://doi.org/10.1007/978-1-60327-367-1_12

34. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59(1):5–15.

35. Antonio Agraz-Doblas, Clara Bueno, Rachael Bashford-Rogers, Anindita Roy, Pauline

Schneider, Michela Bardini, et al. Unraveling the cellular origin and clinical prognostic markers of infant B-cell acute lymphoblastic leukemia using genome-wide analysis. Haematologica. 2019 May 31;104(6):1176–88.

36. Gu Z, Hu Z, Jia Z, Liu J, Mao A, Han H. MD-ALL: an Integrative Platform for Molecular Diagnosis of B-cell Acute Lymphoblastic Leukemia [Internet]. 2023 [cited 2024 May 6]. Available from: https://www.researchsquare.com/article/rs-2798895/v1

37. Beder T, Hansen BT, Hartmann AM, Zimmermann J, Amelunxen E, Wolgast N, et al. The Gene Expression Classifier ALLCatchR Identifies B-cell Precursor ALL Subtypes and Underlying Developmental Trajectories Across Age. HemaSphere. 2023 Sep;7(9):e939.

38. Alaggio R, Amador C, Anagnostopoulos I, Attygalle AD, Araujo IBDO, Berti E, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia. 2022 Jul;36(7):1720–48.

39. Arber DA, Orazi A, Hasserjian RP, Borowitz MJ, Calvo KR, Kvasnicka HM, et al. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. Blood. 2022 Sep 15;140(11):1200–28.

40. Lin C, Xu JQ, Zhong GC, Chen H, Xue HM, Yang M, et al. Integrating RNA-seq and scRNA-seq to explore the biological significance of NAD + metabolism-related genes in the initial diagnosis and relapse of childhood B-cell acute lymphoblastic leukemia. Front Immunol. 2022 Nov 11;13:1043111.

41. Nishiwaki S, Sugiura I, Koyama D, Ozawa Y, Osaki M, Ishikawa Y, et al. Machine learning-aided risk stratification in Philadelphia chromosome-positive acute lymphoblastic leukemia. Biomark Res. 2021 Dec;9(1):13.

42. Schmidt B, Brown LM, Ryland GL, Lonsdale A, Kosasih HJ, Ludlow LE, et al. ALLSorts: an RNA-Seq subtype classifier for B-cell acute lymphoblastic leukemia. Blood Adv. 2022 Jul 26;6(14):4093–7.

43. Anilkumar KK, Manoj VJ, Sagi TM. A review on computer aided detection and classification of leukemia. Multimed Tools Appl. 2024 Feb 1;83(6):17961–81.

44. Johnson WB, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. In: Beals R, Beck A, Bellow A, Hajian A, editors. Contemporary Mathematics [Internet]. Providence, Rhode Island: American Mathematical Society; 1984 [cited 2023 Dec 1]. p. 189–206. Available from: http://www.ams.org/conm/026/

45. Li P, Hastie TJ, Church KW. Very sparse random projections. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. Philadelphia PA USA: ACM; 2006 [cited 2023 Dec 1]. p. 287–96. Available from: https://dl.acm.org/doi/10.1145/1150402.1150436

46. Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: LIU L, ÖZSU MT, editors. Encyclopedia

of Database Systems [Internet]. Boston, MA: Springer US; 2009. p. 532–8. Available from: https://doi.org/10.1007/978-0-387-39940-9_565

47. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000 May 1;16(5):412–24.

48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

49. Kolde R. Pheatmap: pretty heatmaps. R Package Version. 2019;1(2):726.

50. Lin J, Gunopulos D. Dimensionality reduction by random projection and latent semantic indexing. In 2003.

51. Bingham E, Mannila H. Random projection in dimensionality reduction: applications to image and text data. In 2001. p. 245–50.

52. Reisfeld B, Mayeno AN, editors. Computational Toxicology: Volume II [Internet]. Totowa, NJ: Humana Press; 2013 [cited 2024 Mar 19]. (Methods in Molecular Biology; vol. 930). Available from: https://link.springer.com/10.1007/978-1-62703-059-5

53. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008 11/1/2008;9(11):2579–605.

54. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2020 [cited 2024 Mar 19]. Available from: http://arxiv.org/abs/1802.03426

55. Vu K, Poirion PL, Liberti L. Random projections for linear programming. Math Oper Res. 2018;43(4):1051–71.

56. Sadras T, Heatley SL, Kok CH, Dang P, Galbraith KM, McClure BJ, et al. Differential expression of MUC4, GPR110 and IL2RA defines two groups of CRLF2-rearranged acute lymphoblastic leukemia patients with distinct secondary lesions. Cancer Lett. 2017 Nov 1;408:92–101.

57. Pui CH, Roberts KG, Yang JJ, Mullighan CG. Philadelphia Chromosome–like Acute Lymphoblastic Leukemia. Clin Lymphoma Myeloma Leuk. 2017 Aug 1;17(8):464–70.

58. Sánchez R, Ribera J, Morgades M, Ayala R, Onecha E, Ruiz-Heredia Y, et al. A novel targeted RNA-Seq panel identifies a subset of adult patients with acute lymphoblastic leukemia with BCR-ABL1-like characteristics. Blood Cancer J. 2020 Apr 24;10(4):43.

59. Blunck CB, Poubel CP, Lopes BA, Barbosa TC, Maciel ALT, da Costa ES, et al. Characterisation of cells markers associated with IKZF1plus in BCP-ALL. Transl Oncol.

2024 Dec 1;50:102127.

60. Gestrich CK, Oduro KA. Restricted Immunoglobulin Joining Chain (IgJ) Protein Expression in B Lymphoblastic Leukemia (B-ALL) Based on B-ALL Subtype. Blood. 2020 Nov 5;136:7.

61. Li F, Ye W, Yao Y, Wei W, Lin X, Zhuang H, et al. Spermatogenesis associated serine rich 2 like plays a prognostic factor and therapeutic target in acute myeloid leukemia by regulating the JAK2/STAT3/STAT5 axis. J Transl Med. 2023 Feb 11;21(1):115.

62. Lin J, Yan J, Deng X ling, Wang C shan, Wang H sheng. SPATS2 is correlated with cell cycle progression and immune cells infiltration in hepatocellular carcinoma. BMC Gastroenterol. 2023 Jan 11;23(1):8.

63. McGinn OJ, Krishnan S, Bourquin JP, Sapra P, Dempsey C, Saha V, et al. Targeting the 5T4 oncofetal glycoprotein with an antibody drug conjugate (A1mcMMAF) improves survival in patient-derived xenograft models of acute lymphoblastic leukemia. Haematologica. 2017 Jun;102(6):1075–84.

64. Pearce LR, Atanassova N, Banton MC, Bottomley B, van der Klaauw AA, Revelli JP, et al. KSR2 Mutations Are Associated with Obesity, Insulin Resistance, and Impaired Cellular Fuel Oxidation. Cell. 2013 Nov 7;155(4):765–77.

65. Niederkorn M, Hueneman K, Choi K, Varney ME, Romano L, Pujato MA, et al. TIFAB Regulates USP15-Mediated p53 Signaling during Stressed and Malignant Hematopoiesis. Cell Rep. 2020 Feb 25;30(8):2776-2790.e6.

66. Carlet M, Völse K, Vergalli J, Becker M, Herold T, Arner A, et al. In vivo inducible reverse genetics in patients’ tumors to identify individual therapeutic targets. Nat Commun. 2021 Sep 27;12(1):5655.

67. McClure BJ, Heatley SL, Kok CH, Sadras T, An J, Hughes TP, et al. Pre-B acute lymphoblastic leukaemia recurrent fusion, EP300-ZNF384, is associated with a distinct gene expression. Br J Cancer. 2018 Apr 1;118(7):1000–4.

68. Zhu L, Bai W, Cheng Q, Fang J. ZNF384-Related Fusion Genes in Acute Lymphoblastic Leukemia. Cancer Control. 2023 Jan 1;30:10732748231182787.

69. Zhang J, Lu WY, Zhang JM, Lu RQ, Wu LX, Qin YZ, et al. S100A16 suppresses the growth and survival of leukaemia cells and correlates with relapse and relapse free survival in adults with Philadelphia chromosome-negative B-cell acute lymphoblastic leukaemia. Br J Haematol. 2019 Jun 1;185(5):836–51.

70. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, et al. RNA-seq: technical variability and sampling. BMC Genomics. 2011 Jun 6;12(1):293.

71. Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. Array. 2022 Dec 1;16:100258.

72. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct 1;11(10):733–9.

73. Marketa Zaliova, Jan Stuchly, Lucie Winkowska, Alena Musilova, Karel Fiser, Martina Slamova, et al. Genomic landscape of pediatric B-other acute lymphoblastic leukemia in a consecutive European cohort. Haematologica. 2019 Jun 30;104(7):1396–406.

74. Ryan SL, Peden JF, Kingsbury Z, Schwab CJ, James T, Polonen P, et al. Whole genome sequencing provides comprehensive genetic testing in childhood B-cell acute lymphoblastic leukaemia. Leukemia. 2023 Mar 1;37(3):518–28.

75. Diedrich JD, Dong Q, Ferguson DC, Bergeron BP, Autry RJ, Qian M, et al. Profiling chromatin accessibility in pediatric acute lymphoblastic leukemia identifies subtype-specific chromatin landscapes and gene regulatory networks. Leukemia. 2021 Nov 1;35(11):3078–91.

76. Wang H, Sun H, Liang B, Zhang F, Yang F, Cui B, et al. Chromatin accessibility landscape of relapsed pediatric B-lineage acute lymphoblastic leukemia. Nat Commun. 2023 Oct 25;14(1):6792.

77. Leszczenko P, Nowakowska AM, Jakubowska J, Pastorczak A, Zabczynska M, Mlynarski W, et al. Raman spectroscopy can recognize the KMT2A rearrangement as a distinct subtype of leukemia. Spectrochim Acta A Mol Biomol Spectrosc. 2024 Jun 5;314:124173.

78. Good Z, Sarno J, Jager A, Samusik N, Aghaeepour N, Simonds EF, et al. Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. Nat Med. 2018 Apr 1;24(4):474–83.

79. Khabirova E, Jardine L, Coorens THH, Webb S, Treger TD, Engelbert J, et al. Single-cell transcriptomics reveals a distinct developmental state of KMT2A-rearranged infant B-cell acute lymphoblastic leukemia. Nat Med. 2022 Apr 1;28(4):743–51.

80. Iacobucci I, Witkowski MT, Mullighan CG. Single-cell analysis of acute lymphoblastic and lineage-ambiguous leukemia: approaches and molecular insights. Blood. 2023 Jan 26;141(4):356–68.

81. Zhang Y, Wang S, Zhang J, Liu C, Li X, Guo W, et al. Elucidating minimal residual disease of paediatric B-cell acute lymphoblastic leukaemia by single-cell analysis. Nat Cell Biol. 2022 Feb 1;24(2):242–52.

82. Ilaria Iacobucci, Andy G.X. Zeng, Qingsong Gao, Laura Garcia-Prat, Pradyumna Baviskar, Sayyam Shah, et al. SINGLE CELL DISSECTION OF DEVELOPMENTAL ORIGINS AND TRANSCRIPTIONAL HETEROGENEITY IN B-CELL ACUTE LYMPHOBLASTIC LEUKEMIA. bioRxiv. 2023 Jan 1;2023.12.04.569954.

83. Li L, Xiao H, Khoury JD, Wang J, Wan S. Abstract 4907: RanBALL: Identifying B-cell acute

lymphoblastic leukemia subtypes based on an ensemble random projection model. Cancer Res. 2024 Mar 22;84(6_Supplement):4907–4907.