MD-ALL: an integrative platform for molecular diagnosis of B-acute lymphoblastic leukemia

Zunsong Hu,[1,2] Zhilian Jia,[1,2] Jiangyue Liu,[3,4] Allen Mao,[5] Helen Han[1,2] and Zhaohui Gu[1,2]

[1]Department of Computational and Quantitative Medicine, Beckman Research Institute of City of Hope; [2]Department of Systems Biology, Beckman Research Institute of City of Hope; [3]Department of Hematology and Hematopoietic Cell Transplantation; [4]Irell and Manella Graduate School of Biological Sciences of City of Hope and [5]Research Informatics, City of Hope National Medical Center, Duarte, CA, USA

Correspondence: Z. Gu
zgu@coh.org

Received: June 7, 2023.
Accepted: November 6, 2023.
Early view: November 16, 2023.

https://doi.org/10.3324/haematol.2023.283706

Abstract

B-acute lymphoblastic leukemia (B-ALL) consists of dozens of subtypes defined by distinct gene expression profiles (GEP) and various genetic lesions. With the application of transcriptome sequencing (RNA sequencing [RNA-seq]), multiple novel subtypes have been identified, which lead to an advanced B-ALL classification and risk-stratification system. However, the complexity of analyzing RNA-seq data for B-ALL classification hinders the implementation of the new B-ALL taxonomy. Here, we introduce Molecular Diagnosis of Acute Lymphoblastic Leukemia (MD-ALL), an integrative platform featuring sensitive and accurate B-ALL classification based on GEP and sentinel genetic alterations from RNA-seq data. In this study, we systematically analyzed 2,955 B-ALL RNA-seq samples and generated a reference dataset representing all the reported B-ALL subtypes. Using multiple machine learning algorithms, we identified the feature genes and then established highly sensitive and accurate models for B-ALL classification using either bulk or single-cell RNA-seq data. Importantly, this platform integrates multiple aspects of key genetic lesions acquired from RNA-seq data, which include sequence mutations, large-scale copy number variations, and gene rearrangements, to perform comprehensive and definitive B-ALL classification. Through validation in a hold-out cohort of 974 samples, our models demonstrated superior performance for B-ALL classification compared with alternative tools. Moreover, to ensure accessibility and user-friendly navigation even for users with limited or no programming background, we developed an interactive graphical user interface for this MD-ALL platform, using the R Shiny package. In summary, MD-ALL is a user-friendly B-ALL classification platform designed to enable integrative, accurate, and comprehensive B-ALL subtype classification. MD-ALL is available from https://github.com/gu-lab20/MD-ALL.

Introduction

B-acute lymphoblastic leukemia (B-ALL) is a highly heterogeneous disease, which consists of dozens of subtypes with distinct gene expression profiles (GEP) and constellations of genetic alterations.[1] Through the application of transcriptome sequencing (RNA sequencing [RNA-seq]), multiple novel B-ALL subtypes have been identified harboring recurrent genetic lesions and distinct GEP.[2-4] The current World Health Classification (5th edition) of Hematolymphoid Tumors (WHO-HAEM5),[5] along with the International Consensus Classification of Myeloid Neoplasms and Acute Leukemia (ICC),[6] recognize a total of 11 and 26 molecular subtypes of B-ALL, respectively. Currently, clinical diagnosis and classification of B-ALL rely on a range of assays such as flow cytometry, fluorescence *in situ* hybridization (FISH), cytogenetic karyotyping, and panel-based sequencing assays.[7,8] The data generation and analysis using these platforms are time-consuming, expensive, and error-prone. Furthermore, they are inadequate to identify specific subtypes defined by cryptic genetic lesions (e.g., *DUX4* and *MEF2D* rearrangements) or the ones primarily defined by GEP (e.g., Ph-like and ETV6::RUNX1-like).

With rapid progress in discovering novel B-ALL subtypes, updating clinical test assays accordingly has become a challenging task. Alternatively, the application of RNA-seq for clinical diagnosis of B-ALL subtypes has been investigated by multiple institutions and led to encouraging outcomes.[9,10] With its easy-to-follow protocol and multiple layers of information, RNA-seq is poised to revolutionize the classification of B-ALL in both research and clinical settings. However, bioinformatics analysis of RNA-seq data to extract both the sentinel genetic lesions and the GEP signatures for classification is still challenging. Although a few bioinformatics tools have been developed for this purpose,[11-13] they solely rely on GEP for B-ALL subtyping. Here, we present Molecular Diagnosis of Acute Lymphoblastic Leukemia (MD-ALL), a user-friendly bioinformatics platform that integrates genetic and transcriptomic features from RNA-seq to provide integrative, accurate, and comprehensive B-ALL subtype classification.

Methods

RNA-sequencing datasets

Raw RNA-seq data of 3,005 B-ALL samples were collected from multiple published studies.[1,4,14-22] After removing potential duplicates (inferred by KING[23]) and samples with low coverage, 2,955 samples were kept as the primary cohort for this study (*Online Supplementary Table S1*).

RNA-sequencing data analysis

The sequencing reads were aligned to the human genome reference (GRCh38) using STAR.[24] Then Picard (see URL) was used to mark PCR duplicates.

Gene expression

Read counts were calculated by HTSeq[25] and FeatureCount,[26] and then normalized by DESeq2.[27] The ComBat function in the sva R package[28] was used to correct potential batch effects introduced by different library preparation approaches (mRNA *vs.* total RNA and stranded *vs.* unstranded) and variable sequencing lengths (*Online Supplementary Figure S1*). t-distributed stochastic neighbor embedding (tSNE) and uniform manifold approximation and projection (UMAP) were used for dimensionality reduction visualization.

Mutations

Single nucleotide variants (SNV) and insertions/deletions (Indel) were called by following the best practice pipeline from GATK (see Online *Supplementary Methods*).[29]

Deconvolution of bulk gene expression profiles

Single-cell RNA-seq (scRNA-seq) data from the 1-Million Immune Cells Project (see Online *Supplementary Methods*) were reanalyzed to establish a GEP reference representing 20 primary blood cell types. Then, CIBERSORTx[30] was used to deconvolute the bulk GEP of B-ALL to estimate their leukemic cell ratios and granular B-cell composition.

Fusion calling

CICERO[31] and FusionCatcher[32] were used as they can identify gene rearrangements involving highly repetitive regions such as the *IgH* locus.

Copy number variation calling

With read counts and SNV called from RNA-seq, the RNAseqCNV package[33] was used to detect chromosomal-level copy number variation (CNV).

Ancestry inference

The samples' ancestral background was estimated using iAdmix,[34] with the genotype of SNP from the 1K-Genome Project used as the reference.[35,36]

Gene expression profile reference of B-acute lymphoblastic leukemia subtype

Through analyzing the RNA-seq data of the 2,955 B-ALL samples, 26 subtypes were identified, with 19 having distinct GEP features. In order to construct a GEP reference for B-ALL classification, PhenoGraph clustering[37] and k-nearest neighbor analysis of two-dimensional UMAP were performed to identify the representative samples of each subtype.

Feature gene selection

Since the reference cohort is not evenly distributed across B-ALL subtypes, SMOTE algorithm[38] was used to subsample or artificially construct additional samples, which resulted in cohorts with the same sample size per subtype. Then Boruta[39] was used to identify the genes confirmed as contributing features for distinguishing different subtypes.

Gene expression profile-based B-acute lymphoblastic leukemia classification

Two GEP-based B-ALL prediction models were constructed: i) support vector machine (SVM) classification; among multiple tested machine learning algorithms, SVM performed the best; ii) PhenoGraph clustering;[37] PhenoGraph is a clustering algorithm originally developed to identify and partition cells into subpopulations.

Integration of genetic lesions and gene expression profile features

GEP-based subtype prediction and key genetic lesions identified from RNA-seq were integrated for definitive B-ALL classification. A detailed description of integrating GEP-based prediction and sentinel genetic lesions for B-ALL classification is summarized in Table 1.

Single-cell RNA-sequencing analysis and B-acute lymphoblastic leukemia classification

scRNA-seq reads were mapped to the GRCh38 reference. After quality control, the Seurat package[40] was used for gene expression normalization and variable gene selection. With the GEP reference of blood cell types and

Table 1. Integrative criteria for B-acute lymphoblastic leukemia classification by MD-ALL.

Genetic alteration	GEP subtype	GEP feature	Subtype	Note
BCL2, *MYC* or *BCL6* rearrangement	BCL2/MYC	Distinct	BCL2/MYC	The rearrangements can involve genes adjacent to *MYC*
CDX2 overexpression & *UBTF::ATXN7L3* fusion	CDX2/UBTF	Highly distinct	CDX2/UBTF	*CDX2* overexpression
CRLF2 rearrangement	Not Ph/Ph-like	Non-distinct	CRLF2(non-Ph-like)	Less recognized subtype
DUX4 rearrangement	DUX4	Highly distinct	DUX4	*DUX4* gene family overexpression
ETV6::RUNX1 fusion	ETV6::RUNX1	Highly distinct	ETV6::RUNX1	-
No *ETV6::RUNX1* fusion	ETV6::RUNX1	Highly distinct	ETV6::RUNX1-like	Commonly seen with *ETV6* or *IKZF1* rearrangements
HLF rearrangement	HLF	Distinct	HLF	*HLF* overexpression
Chromosome number ≥51	Hyperdiploid	Distinct	Hyperdiploid	-
iAMP21	iAMP21	Less distinct	iAMP21	iAMP21 can be identified by RNAseqCNV
IKZF1 N159Y mutation	IKZF1 N159Y	Highly distinct	IKZF1 N159Y	-
KMT2A rearrangement	KMT2A	Distinct	KMT2A	-
No *KMT2A* rearrangement	KMT2A	Distinct	KMT2A-like	Minor subtype; reported with *AFF1* fusion
Chromosome number 47-50	Hyperdiploid	Distinct	Low hyperdiploid	Less recognized subtype
Chromosome number 31-39	Low hypodiploid	Distinct	Low hypodiploid	Commonly seen with *TP53* mutations
MEF2D rearrangement	MEF2D	Highly distinct	MEF2D	Commonly seen with chromothripsis around MEF2D
Chromosome number 24-30	Hyperdiploid	Non-distinct	Near haploid	Less frequently with GEP of Low hypodiploid
NUTM1 rearrangement	NUTM1	Less distinct	NUTM1	*NUTM1* overexpression
PAX5 P80R mutation	PAX5 P80R	Highly distinct	PAX5 P80R	Abnormal *MEGF10* isoform overexpression
PAX5::ETV6	PAX5::ETV6	Distinct	PAX5::ETV6	Originally reported as PAX5alt
PAX5 alteration	PAX5alt	Distinct	PAX5alt	Featured with *PAX5* fusion, mutation, or iAmp, but not deletion
BCR::ABL1 fusion	Ph/Ph-like	Distinct	Ph	At least two GEP subclusters observed within Ph group
Non-Ph kinase-activating alteration*	Ph/Ph-like	Distinct	Ph-like	Commonly seen with kinase activating fusions
*TCF3::PBX*1 fusion	TCF3::PBX1	Highly distinct	TCF3::PBX1	Rare fusions with *EWSR1* have been reported
ZNF384 rearrangement	ZNF384	Highly distinct	ZNF384	Also observed in mixed phenotype acute leukemia
No *ZNF384* rearrangement	ZNF384	Highly distinct	ZNF384-like	Minor subtype; reported with *ZNF362* fusion
ZEB2 H1038R mutation and/or *CEBP* fusion	ZEB2/CEBP	Distinct	ZEB2/CEBP	Minor subtype

If genetic lesions do not agree with gene expression profile (GEP)-based prediction, genetic lesions determine the primary subtypes, while GEP guide the decision on the secondary subtypes. *Gene rearrangements involving *ABL1, ABL2, CSF1R, PDGFRA, PDGFRB, LYN, CRLF2, JAK2, EPOR, TSLP, TYK2, IL2RB, NTRK3, PTK2B, FGFR1, FLT3, DGKH, BLNK,* and *CBL*. MD-ALL: Molecular Diagnosis of Acute Lymphoblastic Leukemia; iAmp: intragenic amplification.

B-ALL subtypes described above, SingleR[41] was used to annotate cell types and B-ALL subtypes for each cell.

Results

Characteristics of the RNA-sequencing cohort

In total, 2,955 B-ALL samples with high-quality RNA-seq data were included in this study (*Online Supplementary Table S1*). This cohort comprises 67.8% pediatric and 28.4% adult cases from different racial/ethnic backgrounds, with a relative higher proportion of male patients (56.1%) (*Online Supplementary Figure S2*). Through manual curation of the genetic lesions, 3,304 gene rearrangements, 2,979 sequence mutations, and 95 *FLT3* internal tandem duplications (ITD) were identified (*Online Supplementary Tables S2-4*). Subsequently, sentinel gene fusions and mutations were compiled to facilitate B-ALL classification (*Online Supplementary Tables S5* and *S6*). Through integration of genetic lesions and GEP-based predictions, the cohort was classified into 26 molecular subtypes (Figure 1A). In summary, this well-curated large cohort encompasses all the reported B-ALL subtypes across different age groups, sex, and racial/ethnical backgrounds, making it an excellent resource for constructing and evaluating B-ALL subtype prediction models, as well as advancing our understanding of the genetic and transcriptomic features of each B-ALL subtype.

High accuracy of gene expression profile-based B-acute lymphoblastic leukemia classification by MD-ALL

In order to generate a GEP reference for subtype prediction, 1,821 samples confirmed by sentinel genetic lesions and stable GEP clusters were selected as the training cohort, representing 19 B-ALL subtypes with distinct GEP (Figure 1B). Using this GEP reference cohort, 1,058 feature genes were consistently confirmed by the Boruta algorithm in eight SMOTE-resampled cohorts (*Online Supplementary Table S7*). Due to the substantial batch effect between mRNA-seq and total RNA-seq library preparation approaches (*Online Supplementary Figure S1*), only the protein-coding genes were considered for feature selection to accommodate both library types. Each feature gene was assigned an importance score by Boruta, which was used to rank their significance for distinguishing different subtypes. Based on the reference cohort and selected feature genes, MD-ALL employs SVM and PhenoGraph algorithms to predict the subtypes of the test samples. Considering that the user-provided test RNA-seq data may use different library preparation strategies and the sample size may not be sufficient for reliable batch effect correction, our prediction models were evaluated using the test samples' GEP data without batch effect correction.

For the training cohort, 100% accuracy was achieved by both SVM and PhenoGraph algorithms as expected (Figure 2A). For the test cohort, subtypes with non-distinct GEP, such as Near haploid, and less recognized subtypes, such

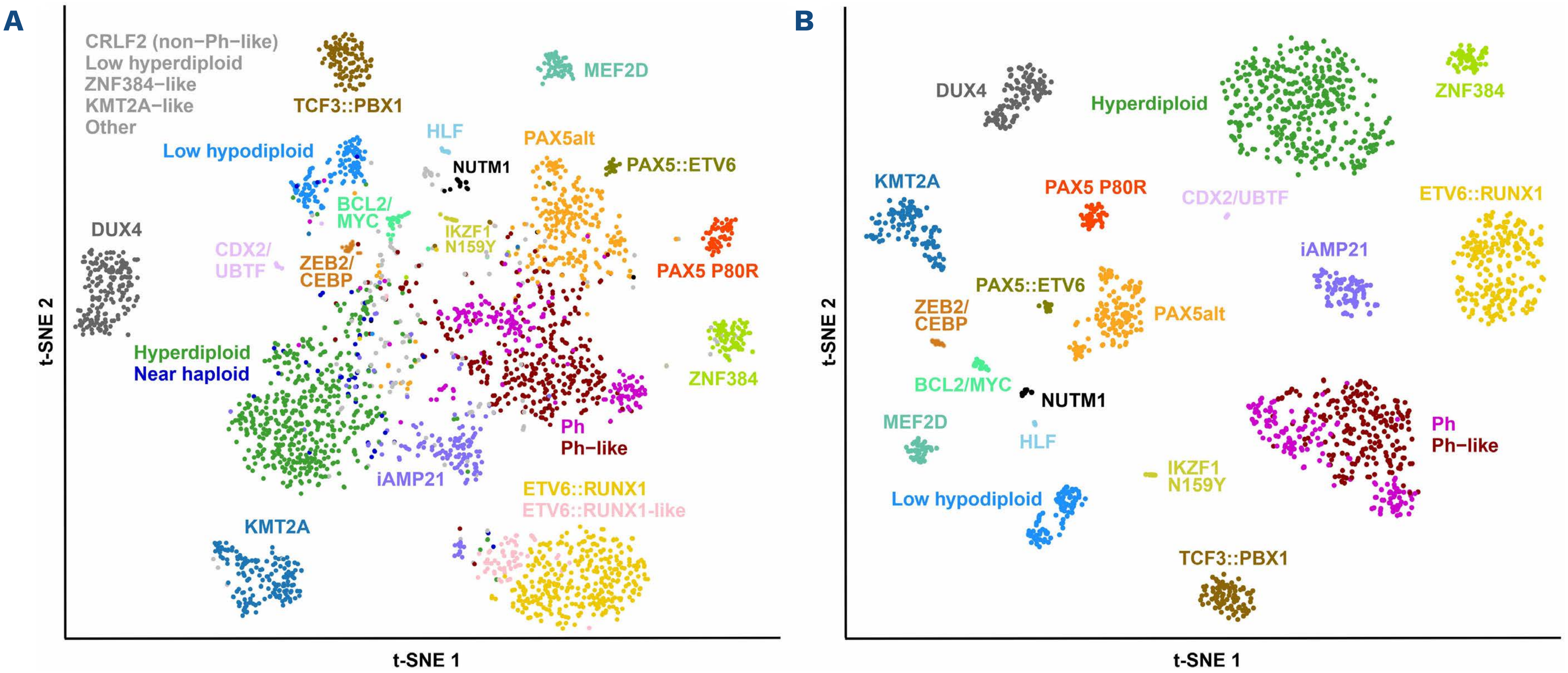


Figure 1. Gene expression profiles of B-acute lymphoblastic leukemia subtypes. The t-distributed stochastic neighbor embedding (tSNE) plots display the gene expression profiles (GEP) distribution using 1,058 signature coding genes identified from reference B-acute lymphoblastic leukemia (B-ALL) subtypes (see Methods). GEP are derived from bulk RNA-sequencing data, with each dot representing an individual sample. A perplexity parameter of 10 was used in tSNE analysis to better visualize the minor subtypes. B-ALL subtypes are color-coded and annotated, while less recognized ones such as CRLF2 (non-Ph-like), Low hyperdiploid, ZNF384-like, KMT2A-like, and unclassified are shown in grey. (A) tSNE plot of 2,955 B-ALL samples, which represents the total cohort of this study. (B) tSNE plot of reference samples (N=1,821) from 19 B-ALL subtypes with distinct GEP. For GEP-based classification, Ph and Ph-like are combined as one Ph/Ph-like group.

as Low hyperdiploid and CRLF2 (non-Ph-like), as well as unclassified cases were excluded. In order to evaluate the performance across different tools, phenocopy subtypes, including Ph-like, ETV6::RUNX1-like, KMT2A-like, and ZNF384-like, were merged with their canonical counterparts to accommodate the different strategies used by different tools for identifying them. Moreover, PAX5alt and Ph-like subtypes are primarily defined by GEP, but their GEP features are less distinct compared with others. In order to avoid potential bias of evaluating different tools for these two subtypes, only the PAX5alt and Ph-like cases confirmed by sentinel genetic lesions (i.e., *PAX5* mutation, fusion, or intragenic amplification in PAX5alt, and rearrangements involving kinase activating genes in Ph-like; see Table 1) were kept in the test cohort.

Although this study enrolled a large number of samples, seven minor subtypes have fewer than 30 qualified samples, which include *BCL2/MYC* (N=29), *PAX5::ETV6* (N=23), *ZEB2/CEBP* (N=19), *NUTM1* (N=18), *IKZF1* N159Y (N=14), *HLF* (N=11), and *CDX2/UBTF* (N=9). Following the training *versus* testing sample size ratio of 2:1 set for the major subtypes, fewer than ten samples would be left for testing. Therefore, a leave-one-out validation was used to evaluate the prediction models for these minor subtypes, eventually resulting in a test cohort of 974 samples (*Online Supplementary Table S8*).

Through GEP-based prediction, SVM and PhenoGraph successfully classified 971 and 972 samples into distinct subtypes, respectively, with high overall accuracy achieved in both models (SVM: 96.1%, N=936; PhenoGraph: 92.7%, N=903). Despite the high accuracy of both models, SVM surpassed PhenoGraph in discerning multiple subtypes such as intrachromosomal amplification of chromosome 21 (iAMP21) and Ph/Ph-like, whereas PhenoGraph demonstrated superior performance over SVM in identifying the ETV6::RUNX1/-like subtype (Figure 2B, C).

In summary, the GEP-based models in MD-ALL can achieve high classification rate as well as high accuracy for B-ALL classification.

MD-ALL classification is superior compared with alternative tools

Currently, there are three alternative tools providing the functionality of B-ALL classification, which are ALLSpice,[11] ALLSorts,[12] and ALLCatchR.[13] The subtype prediction by these tools is solely based on GEP; therefore, the comparison with them is restricted to the GEP prediction results of MD-ALL. Additionally, it should be noted that the holdout test cohort of this study partially overlaps with the training cohort of the other tools, since the majority of B-ALL RNA-seq data used in MD-ALL and these alternative tools are from our previous study, which comprises 1,988 B-ALL samples.[14] This overlap may lead to overestimated accuracy of the alternative tools. Additionally, the *PAX5::ETV6* fusion, originally reported as one of the sentinel alterations of PAX5alt subtype,[14] is still considered as PAX5alt by other tools. Therefore, the PAX5::ETV6 cases were annotated as PAX5alt when comparing the performance of different models.

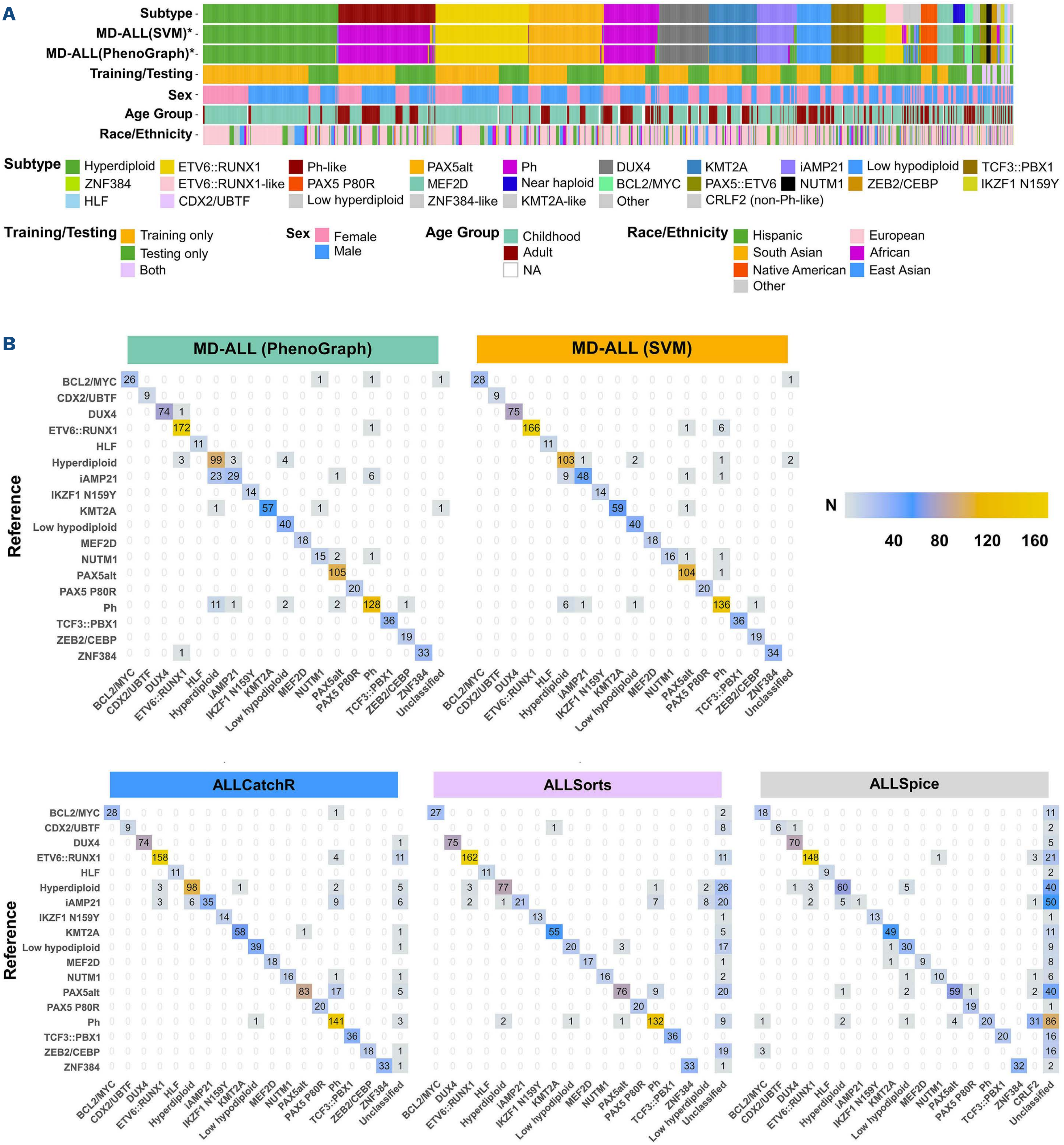
In the same test cohort of 974 samples, a much higher number of samples remained unclassified by ALLCatchR (N=36), ALLSorts (N=142), and ALLSpice (N=327) when compared to MD-ALL. The overall accuracies were 91.3% (889/974), 81.2% (791/974), and 58.8% (573/974) for each method, respectively, which were significantly lower than those achieved by both models in MD-ALL. When considering only the samples with assigned subtypes, the accuracies of ALLCatchR, ALLSorts, and ALLSpice were 94.8% (889/938), 95.1% (791/832), and 88.6% (573/647), respectively (Figure 2B). Therefore, the MD-ALL SVM prediction surpassed all other models in terms of classification rate and accuracy. For the MD-ALL PhenoGraph model, when evaluating solely the samples classified by other tools, the accuracies reached 93.7% (879/938 ALLCatchR-classified), 94.8% (789/832 ALLSorts-classified), and 97.1% (628/647 ALLSpice-classified), indicating that PhenoGraph is also a highly reliable prediction model for B-ALL subtyping (*Online Supplementary Table S8*). Among the prediction models, ALLSpice had the lowest number of correctly classified samples (N=573). Moreover, key B-ALL subtypes, such as Ph-like and ZEB2/CEBP, are not included in ALLSpice, significantly limiting its potential for clinical use. Therefore, ALLSpice will be excluded from further comparisons.

In terms of specificity, MD-ALL (SVM and PhenoGraph), ALLCatchR and ALLSorts demonstrated excellent performance for most subtypes. However, differences were observed in certain subtypes: MD-ALL algorithms outperformed ALLCatchR and ALLSorts in Ph/Ph-like subtype, while ALLCatchR and ALLSorts excelled in Hyperdiploid subtype (Figure 2C). As for sensitivity, ALLSorts consistently underperformed compared with MD-ALL and ALLCatchR in most subtypes, particularly those with less distinct GEP clusters, such as iAMP21 (35.6%), Low hypodiploid (50.0%), PAX5alt (72.4%), and Hyperdiploid (70.6%). Of note, ALLCatchR performed very well in the test cohort; especially in the Ph/Ph-like group, ALLCatchR surpassed both MD-ALL algorithms in sensitivity (97.2%) at the expense of reduced specificity (95.9%) compared to MD-ALL. As both MD-ALL SVM and ALLCatchR use the SVM algorithm, the high sensitivity levels achieved by these two models are anticipated. However, MD-ALL SVM surpassed ALLCatchR in terms of sensitivity for multiple major subtypes, such as iAMP21 (81.4% *vs.* 59.3%), PAX5alt (99.0% *vs.* 79.0%), Hyperdiploid (94.5% *vs.* 89.9%), ETV6::RUNX1/-like (96.0% *vs.* 91.3%), ZNF384 (100% *vs.* 97.1%), and Low hypodiploid (100% *vs.* 97.5%; Figure 2C)

In conclusion, the GEP-based models in MD-ALL demonstrate superior performance over alternative tools in B-ALL classification, even for the challenging subtypes.

Integrative RNA-sequencing analyses provide reliable and definitive B-acute lymphoblastic leukemia classification

Although GEP alone can provide highly accurate B-ALL classifications, sentinel genetic lesions may take precedence when GEP results are ambiguous or conflict with the genetic lesions. Additionally, genetic lesions found in the same samples may also lead to different subtypes. For example, among the 202 Ph-positive cases in this study, 22 (10.9%) carry more than 50 chromosomes, which fit the definition of Hyperdiploid subtype. Considering the associated prognosis and potential benefit of using tyrosine kinase inhibitors, Ph subtype overrides Hyperdiploid when both sentinel genetic lesions are identified, even though



Continued on following page.

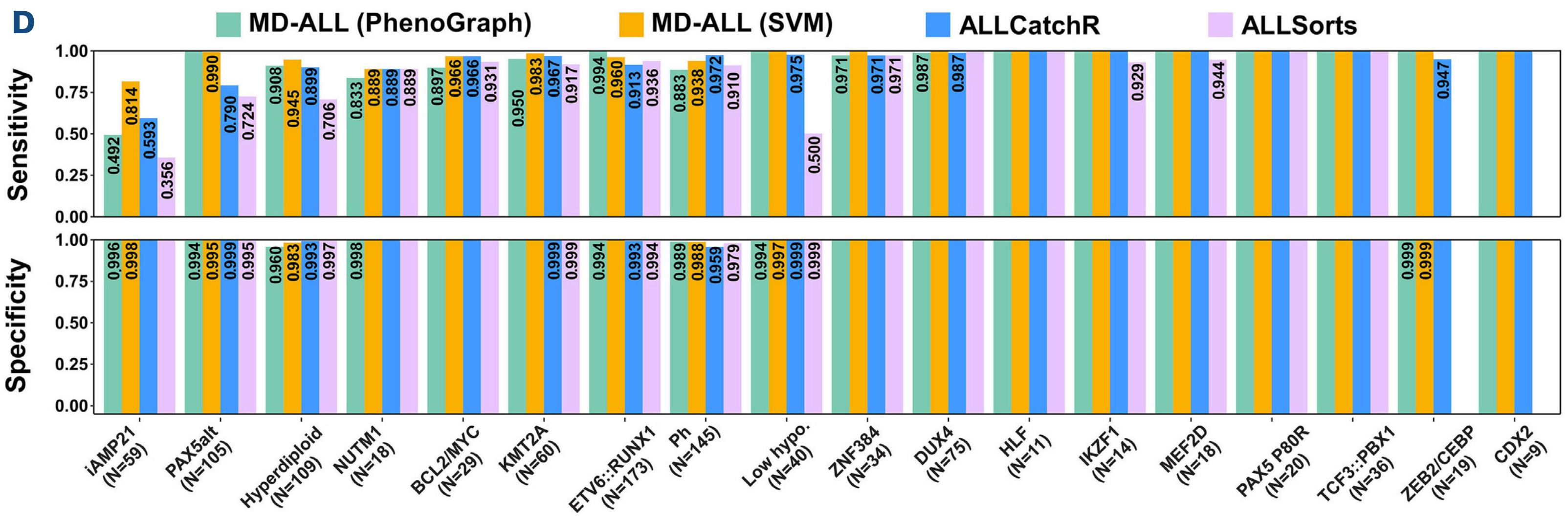
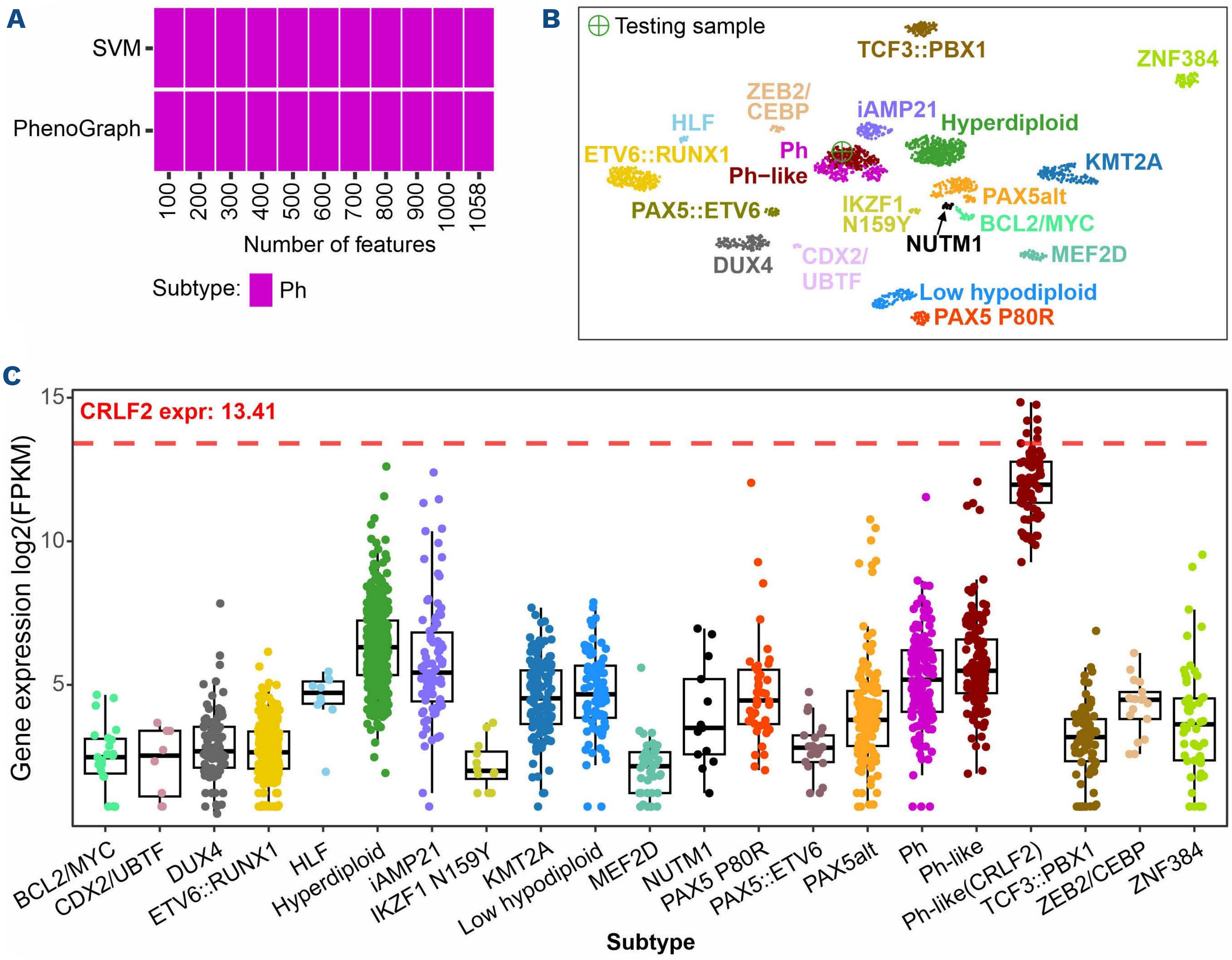


Figure 2. High accuracy of B-acute lymphoblastic leukemia subtyping with MD-ALL. (A) A heatmap showing the study cohort (N=2,955) highlights B-ALL subtypes and metadata. Each column represents a sample. Two gene expression profile (GEP)-based subtype prediction models, support vector machine (SVM) and PhenoGraph, were established within Molecular Diagnosis of Acute Lymphoblastic Leukemia (MD-ALL). *Phenocopy subtypes are identified by their similar GEP to their corresponding canonical subtypes and are thus annotated with the same colors. For the training/testing annotation, leave-one-out validation was used to evaluate the prediction for minor subtypes, which made samples in these subtypes as both training and testing data. Sex information was inferred using the package (see Methods), while race/ethnicity information was determined by the iAdmix package (see Methods). (B) A confusion matrix compares subtype predictions made by MD-ALL and alternative tools. The ground-truth subtypes of the 974-sample test cohort are displayed on the left side of each matrix, while prediction results from different models are shown at the bottom. The phenocopy subtypes and their corresponding canonical subtypes are merged for evaluation. MD-ALL, comprising SVM and PhenoGraph models, is compared with ALLCatchR, ALLSorts, and ALLSpice, with ALLSpice showing the largest number of unclassified samples. (C) Sensitivity and specificity of GEP-based B-ALL classification. The same test cohort (N=974) described above was used to evaluate all different models. The ZEB2/CEBP and CDX2 (CDX2/UBTF) subtypes are not available in the ALLSorts model. Detailed sensitivity and specificity values are labeled for conditions where they are not 100%. The evaluated sample sizes per subtype are annotated in parentheses.

strong Hyperdiploid GEP was observed in eight cases (*Online Supplementary Table S1*). By integrating multiple aspects of information, a more well-rounded subtyping result can be achieved. For example, 43 Near haploid cases were identified based on the total chromosome number (≤30). These cases were predicted as Hyperdiploid (N=40) or Low hypodiploid (N=3) by our GEP models. Of the three cases with Low hypodiploid GEP, they all carry 28 chromosomes, which are on the boundary of defining Near haploid and Low hypodiploid subtypes. Furthermore, they all carry *TP53* hotspot mutations with high mutant allele frequency (>90%), which resembles the features of Low hypodiploid.[42] Therefore, they should be categorized as Low hypodiploid subtype. This scenario highlights the importance of integrating GEP predictions with signature genetic lesions to accurately determine the subtypes (*Online Supplementary Table S1*).

In MD-ALL, users can provide raw translocations and sequence mutations for integrative B-ALL classification. Upon re-analysis of 2,955 RNA-seq samples, 96 sentinel gene rearrangements and 587 mutations were identified (*Online Supplementary Tables S5* and *S6*). By integrating GEP and mutation information, MD-ALL calls RNAseqCNV to identify aneuploid subtypes, such as Hyperdiploid, Low hypodiploid, Near haploid, and even iAMP21. Our previous work on RNAseqCNV[33] demonstrated 100% accuracy in determining aneuploid subtypes, though iAMP21 detection was not mentioned. In this study, we observed high accuracy (35/36) of detecting iAMP21 in B-ALL samples with confirmed iAMP21 status (by SNP array), further broadening the utility of RNA-seq for defining B-ALL subtypes (*Online Supplementary Figure S3*; *Online Supplementary Table S9*). In addition, MD-ALL provides visualization of subtyping results for test sample in SVM and PhenoGraph models using different numbers of genes (Figure 3A). This visualization aids in assessing the stability of the subtyping results. Furthermore, a UMAP plot of the test sample mapped to the reference cohort using all the feature genes (N=1,058) offers an insightful overview of the sample's relationship to the reference (Figure 3B). As certain gene rearrangements are strongly associated with specific gene expressions, such as *CRLF2* overexpression commonly seen in *CRLF2*-rearranged cases, MD-ALL can display a gene's expression across all B-ALL subtypes to verify the reliability of specific fusions or subtypes (Figure 3C). The *JAK2* p.R683 hotspot mutations, known for their high concurrence in *CRLF2*-rearranged cases,[43] further confirm the reliability of the *IGH::CRLF2* fusion. MD-ALL then compiles all input information to assist the final subtype classification. For instance, a sample with an *IGH::CRLF2* fusion and GEP-based Ph/Ph-like prediction, but lacking *BCR::ABL1* fusion, can be definitively classified as Ph-like (Figure 3D). In order to facilitate definitive B-ALL classification for all subtypes, MD-ALL incorporates a knowledge-based subtyping guideline that integrates both genetic lesions and GEP features (Table 1).

With the technical, biological, and clinical considerations



D

Genetic Alterations				Gene Expression Profile		Subtype Summary	Confidence
FusionCatcher	CICERO	Mutations	RNAseqCNV*	PhenoGraph	SVM		
IGH::CRLF2 CRLF2::IGH	IGH::CRLF2	JAK2 p.R683G(0.51)	male; 48,10+,X+	Ph,11(1)	Ph,11(1)	Ph-like	High

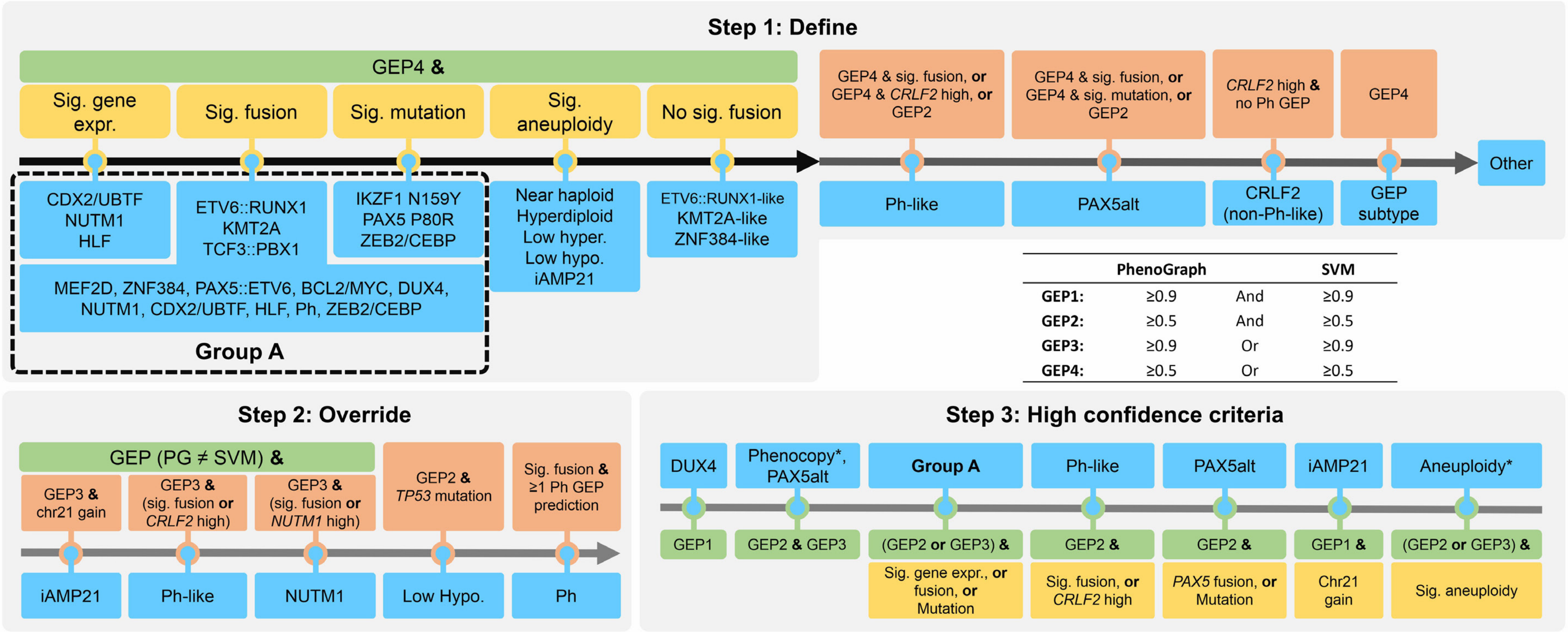
*Clinical Karyotype: 48,XY,+X,+10[17]/46,XY[3];abnormal

Figure 3. Integrative summary of B-acute lymphoblastic leukemia classification by MD-ALL. (A) Gene expression profile (GEP) -based subtype prediction by support vector machine (SVM) and PhenoGraph models. Different numbers of feature genes are used in the prediction models to evaluate classification robustness. The test sample was consistently predicted as the Ph subtype. (B) The test sample is mapped to a predefined uniform manifold approximation and projection (UMAP) space for visualizing GEP-based classification. The UMAP uses 1,058 features genes. The test sample clusters with the Ph/Ph-like group, which agrees with the SVM and PhenoGraph prediction. (C) Expression of a specific gene across different B-acute lymphoblastic leukemia (B-ALL) subtypes. Ph-like (CRLF2) is shown as a separate group here for confirming *CRLF2* rearrangements. Users can specify a gene to examine its expression for validating genetic lesions (e.g., overexpression of *CRLF2* in *CRLF2*-rearranged cases) or potential subtypes. (D) Summary of Molecular Diagnosis of Acute Lymphoblastic Leukemia (MD-ALL) to assist B-ALL classification. The genetic lesions, which include fusions, mutations, large-scale copy number variation (CNV), are integrated with GEP-based prediction by PhenoGraph and SVM to assist the classification of the test sample's B-ALL subtype.

applied in the MD-ALL platform, we developed a decision-tree-based pipeline to integrate multiple aspects of information acquired from RNA-seq to accurately determine B-ALL subtypes and the associated confidence score (Figure 4). Basically, a step-by-step process is taken for each sample to determine the subtype based on the GEP and signature genetic lesions, and then assigns the confidence score.

Of the total cohort comprising 2,955 samples, 2,689 (91.0%) were classified with high confidence. Among these, 2,682 (99.7%) were consistent with the manually curated subtypes (*Online Supplementary Table S10*). In the seven samples with discrepancies:

- Two curated B-other cases without detectable iAMP21 alteration were predicted as iAMP21, based on GEP and chr21 gain.
- In contrast, two curated iAMP21 cases were defined as Hyperdiploid (by GEP and 52 chromosomes) and PAX5alt



	PhenoGraph		SVM
GEP1:	≥0.9	And	≥0.9
GEP2:	≥0.5	And	≥0.5
GEP3:	≥0.9	Or	≥0.9
GEP4:	≥0.5	Or	≥0.5

Figure 4. Integrative B-acute lymphoblastic leukemia classification pipeline. The integrative B-ALL classification pipeline implemented in Molecular Diagnosis of Acute Lymphoblastic Leukemia (MD-ALL) consists of 3 steps: Step 1. Define. In this step, MD-ALL integrates gene expression profiles (GEP), signature (sig.) gene expression (expr.), fusions, mutations, and aneuploidies to define different B-ALL subtypes. The sequence in which subtypes are defined in this step is carefully orchestrated, primarily following the order from the most distinct subtypes, such as the ones in Group A, to less distinct ones, such as the aneuploid and phenocopy subtypes. Step 2. Override. Due to the potential overlap of some subtypes, 5 additional rules were implemented to override the subtypes defined in Step 1, which include, Ph-like, NUTM1, Low hypodiploid, and Ph. Step 3. Define high confidence score. With the subtypes defined in Step 1 and 2, a high confidence score will be assigned if they meet specific criteria, which are developed based on the GEP prediction scores and signature genetic alterations. For the less recognized subtypes such as Low hyperdiploid and CRLF2 (non-Ph-like), a low confidence is assigned. Hyper.: Hyperdiploid; hypo.: hypodiploid; GEP1 to GEP4 are defined based on the GEP prediction by PhenoGraph (PG) and SVM shown in a table. In Step 3, Phenocopy* subtypes include Ph-like, ETV6::RUNX1-like, KMT2A-like, and ZNF384-like, and Aneuploidy* subtypes include Near haploid, Hyperdiploid, and Low hypodiploid.

(by GEP) by MD-ALL, respectively.

- One curated KMT2A-like case was predicted as KMT2A subtype, based on GEP and a *KMT2A::BIRC3* rearrangement. Due to the low confidence in the *KMT2A* rearrangement, which was supported by only four reads, the sample was eventually classified as KMT2A-like after manual curation.
- One curated Ph case was classified as Low hypodiploid based on GEP and a *TP53* mutation. However, this classification was overridden and labeled as Ph subtype due to the detected *BCR::ABL1* fusion.
- One curated ETV6::RUNX1 case with a predicted Ph-like subtype, because of a strong Ph GEP signature, was eventually classified as ETV6::RUNX1 subtype based on an *ETV6::RUNX1* fusion.

Of the samples with low confidence scores (N=266), 53.0% (N=141) are concordant with the manually curated results. Approximately half of the 266 samples (N=130) are classified as aneuploid or iAMP21 subtypes, which can be easily confirmed by manually checking the RNAseqCNV or available karyotype information. In the remaining 136 samples, the subtypes can be distinguished by checking the GEP-based predictions and the signature genetic alterations provided by MD-ALL.

In summary, MD-ALL integrates multiple aspects of information derived from RNA-seq data to provide highly accurate and definitive B-ALL classification.

Distinct B-cell differentiation patterns of B-acute lymphoblastic leukemia subtypes

Using high-quality scRNA-seq data, we compiled a GEP reference consisting of over 10K cells that represent 20 major blood cell types (see Methods; Figure 5A). Subsequently, we used the single-cell GEP reference to deconvolute the bulk RNA-seq GEP of different B-ALL subtypes (*Online Supplementary Table S11*). Our analysis revealed that the PAX5 P80R and KMT2A subtypes carry a strong Pro B1 (pre-pro B stage) signature, indicating that they are at the very early stage of B-cell development. By contrast, the BCL2/MYC subtype exhibits a strong enrichment of pre B2 and even immature B-cell signatures (Figure 5B). This suggests that the leukemic B cells are more mature, which is consistent with the observation that *BCL2* and *MYC* rearrangements are more commonly seen in B-cell lymphomas,[44] a malignancy transformed from more mature B lymphocytes. These conclusions agree with clinically reported immunophenotypic features of B-ALL subtypes[18] as well as other digital deconvolution reports.[45]

In order to validate the digital deconvolution results, we compared the clinically reported B-cell blast ratio from 70 B-ALL samples and their inferred B-cell ratio by CIBER-

SORTx, and a high correlation was observed (correlation=0.85; 95% confidence interval [CI]: 0.76-0.9; Figure 5C; *Online Supplementary Table S12*). Therefore, digital deconvolution can be used to assess the potential normal cell contamination in bulk samples. In addition, we observed that samples without classified subtypes were enriched with low B-cell ratio (35.9% of 64 samples have <50% B-cell ratio) compared to those with defined subtypes (3.1% of 2,718 samples have <50% B-cell ratio). This finding indicates that contamination of normal cells can interfere with classification of B-ALL subtypes.

High sensitivity B-acute lymphoblastic leukemia subtyping at a single-cell level

In bulk RNA-seq, it is critical to obtain pure leukemic cells prior to RNA-seq assay to ensure that the GEP represents the disease. However, in clinical settings, patient samples often contain a low proportion of leukemic cells. As a result, B-cell blasts require proper enrichment prior to analysis. Even with B-cell enrichment, samples may still be contaminated by normal B-cell blasts, or contain an inadequate number of enriched cells for bulk RNA-seq.

In order to address these challenges, we explored the potential of using single-cell GEP to identify B-cell blasts (pro- to pre-B cells) using the GEP reference representing major blood cell types (Figure 5A). After identifying the blast cells, we annotated them to different B-ALL subtypes using the GEP reference compiled from bulk RNA-seq (Figure 1B). By using public scRNA-seq datasets,[46,47] we can reliably (>50% of the B-cell blasts are correctly predicted) identify multiple B-ALL subtypes, such as KMT2A, ETV6::RUNX1/-like, Hyperdiploid, Ph, DUX4, MEF2D, PAX5alt, TCF3::PBX1, and ZNF384 (Figure 6; *Online Supplementary Figure S4*), even in samples with blast percentages below 20% (Figure 6B). Furthermore, a cluster of B cells was observed with a mixture of different B-ALL subtypes in the KTM2A case (Figure 6A), indicating that they are normal B-cell blasts.

In summary, our study highlights the potential of single-cell analysis in the sensitive and accurate detection of leukemic cells and their B-ALL subtypes. With the advent of more cost-effective scRNA-seq platforms and the continual decrease in sequencing costs, single-cell analysis is expected to revolutionize clinical diagnosis of granular disease subtypes.

MD-ALL: an integrative platform for B-acute lymphoblastic leukemia classification

MD-ALL integrates both GEP and signature genetic lesions to provide a one-stop solution for B-ALL classification. This is especially important to distinguish the canonical subtypes (e.g., Ph and ETV6::RUNX1) from their phenocopy counterparts (e.g., Ph-like, and ETV6::RUNX1-like, respectively). In

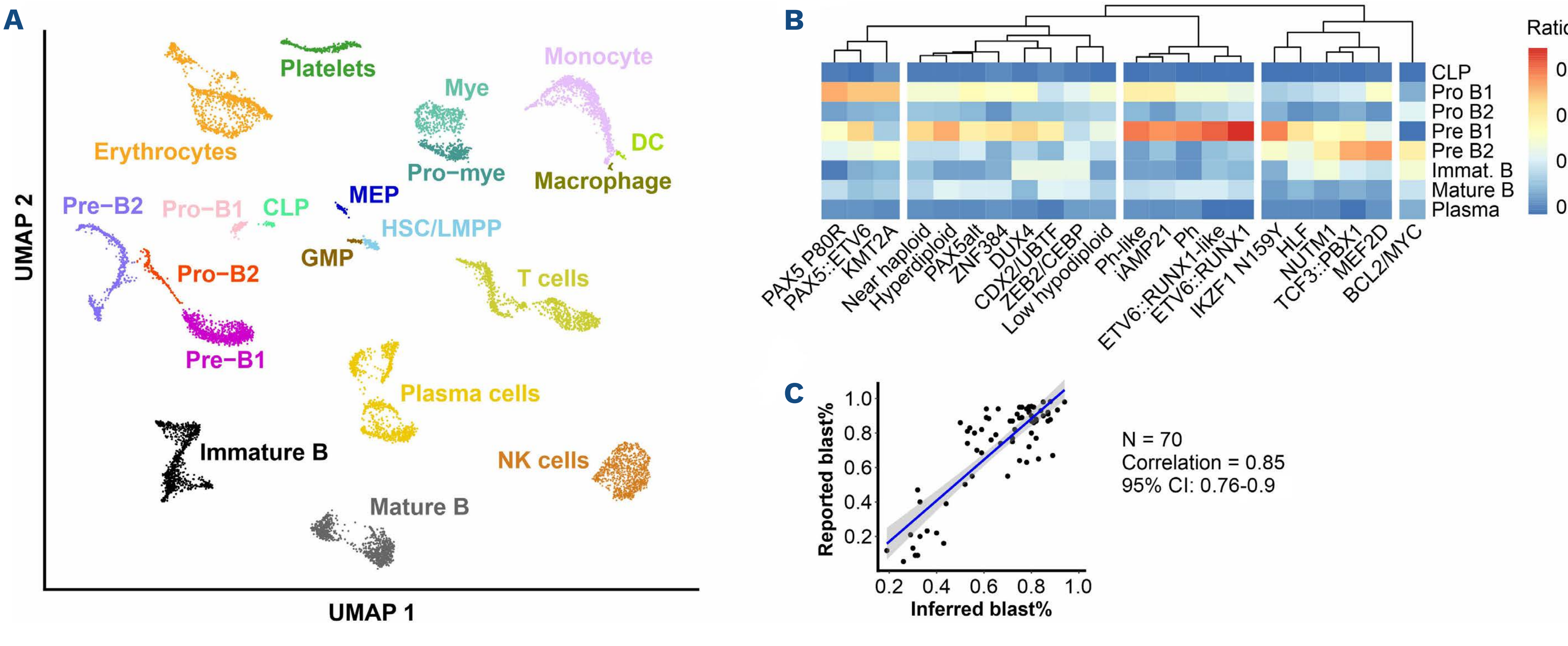


Figure 5. Deconvolution of bulk gene expression profile of B-acute lymphoblastic leukemia subtypes. (A) Uniform manifold approximation and projection (UMAP) of single-cell gene expression reference of the primary blood cell types. Over 10K cells representing 20 primary blood cell types were selected from the 1-Million Immune Cells project (see *Online Supplementary Methods*). (B) Cells are classified into granular B cell differentiation stages, including common lymphoid progenitor (CLP), pro-B1 (early pro-B), pro-B2 (late pro-B), pre-B1 (large pre-B), and pre-B2 (small pre-B). HSC: hematopoietic stem cell; LMPP: lymphoid-primed multipotential progenitor; DC: dendritic cell; Mye: myelocytes; Pro-mye: promyelocytes; GMP: granulocyte-monocyte progenitor; MEP: megakaryocyte-erythrocyte progenitor; NK cell: natural killer cell. (B) Heatmap of different B-acute lymphoblastic leukemia (B-ALL) subtypes and their inferred B-cell differentiation stages. For each subtype, the median value of each B-cell stage is calculated and presented in the heatmap. The Euclidean distance and Ward's minimum variance clustering method were used to generate the clusters. (C) Correlation of digitally inferred and clinically reported blast percentage (blast%). The inferred blast% is estimated by combining B-lineage cells from pro B1 to mature B stages (see Methods). Seventy samples from a cohort provided by the ALLSorts package were used in this analysis.

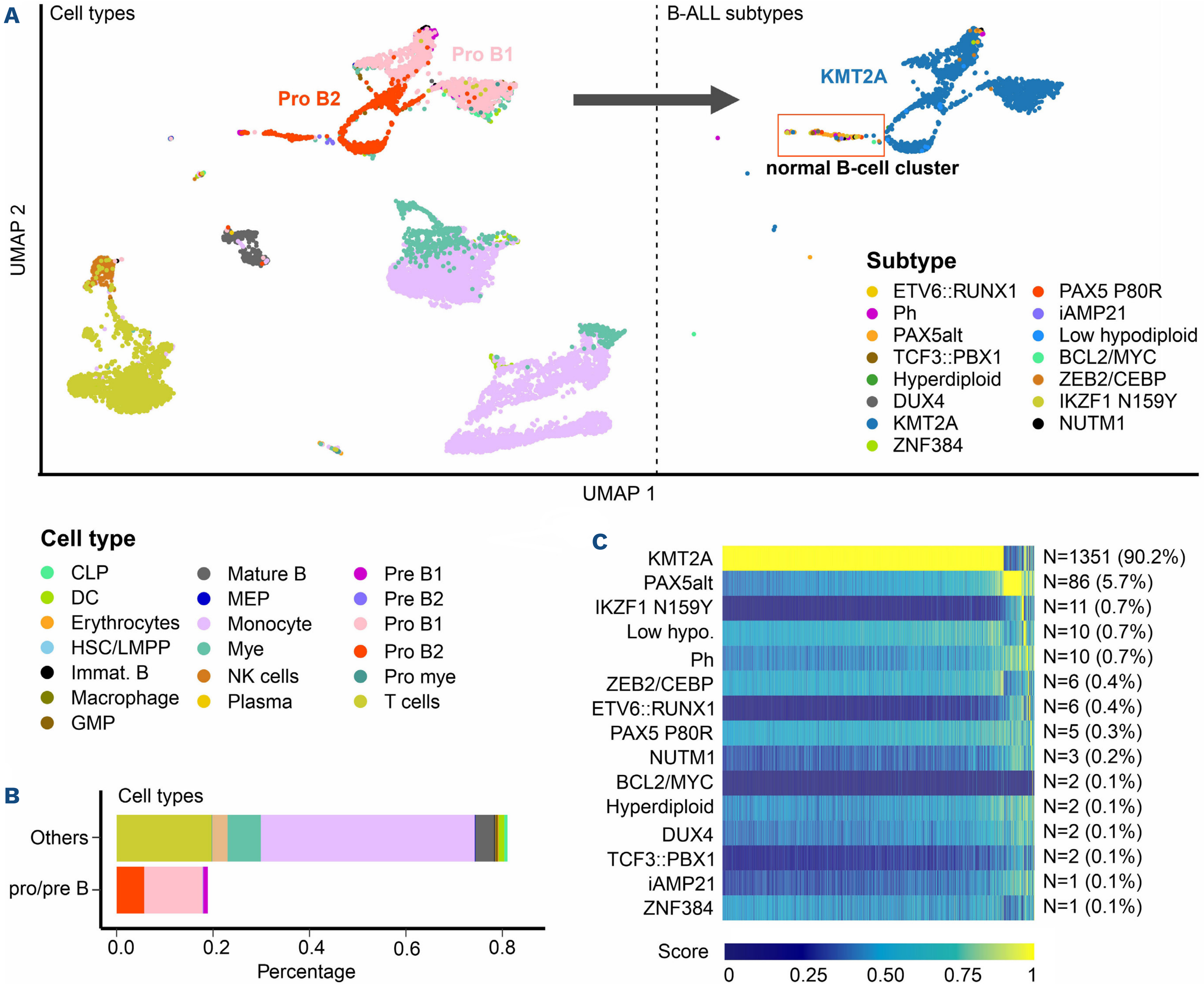


Figure 6. B-acute lymphoblastic leukemia subtype classification at a single-cell level. (A). Single-cell RNA sequencing (scRNA-seq) of a B-acute lymphoblastic leukemia (B-ALL) sample at diagnosis shown in a UMAP plot. The abnormally enriched B-cell blasts (pro- to pre-B cells) represent the leukemic cells. With the gene expression profile (GEP) reference of the B-ALL subtypes, the majority of the B-cell blasts are reliably predicted as KMT2A subtype, which is consistent with the reported subtype. A small cluster (highlighted in a red rectangle) observed with a mixture of different B-ALL subtypes indicates that they are normal B-cell blasts. (B) A bar graph shows the distribution of different cell types. Less than 20% of the test sample are B-cell blasts, which could be challenging to be accurately identified as KMT2A subtype based on bulk GEP prediction. (C) Heatmap of subtype prediction score shows that over 90% of the B-cell blasts exhibit highly reliable KMT2A GEP signature. Low hypo.: Low hypodiploid; CLP: common lymphoid progenitor; HSC: hematopoietic stem cell; LMPP: lymphoid-primed multipotential progenitor; DC: dendritic cell; Mye: myelocytes; Pro-mye: promyelocytes; GMP: granulocyte-monocyte progenitor; MEP: megakaryocyte-erythrocyte progenitor; NK cell: natural killer cell.

addition, an interactive graphical interface was provided within MD-ALL, making the tool accessible to users with limited or no computational background. The minimum required input is the raw read count from RNA-seq data. The test samples will be normalized against an internal reference cohort, which consists of 234 samples representing all reported subtypes (*Online Supplementary Table S13*). This reference cohort was sequenced using various library preparation kits, sequencing lengths, and strandness. Therefore, normalization against this reference helps minimize potential batch effects. Users may also provide raw output of gene rearrangements and mutations to MD-ALL to perform automatic filtering and genetic alteration identification based on the signature lesions identified in the large B-ALL cohort. Subsequently, MD-ALL will integrate the information of genetic alterations and GEP for robust B-ALL classification (Figure 7A). Furthermore, MD-ALL also provides the functionality for single-cell B-ALL classification, requiring only the raw read count output from standard scRNA-seq analysis (Figure 7B).

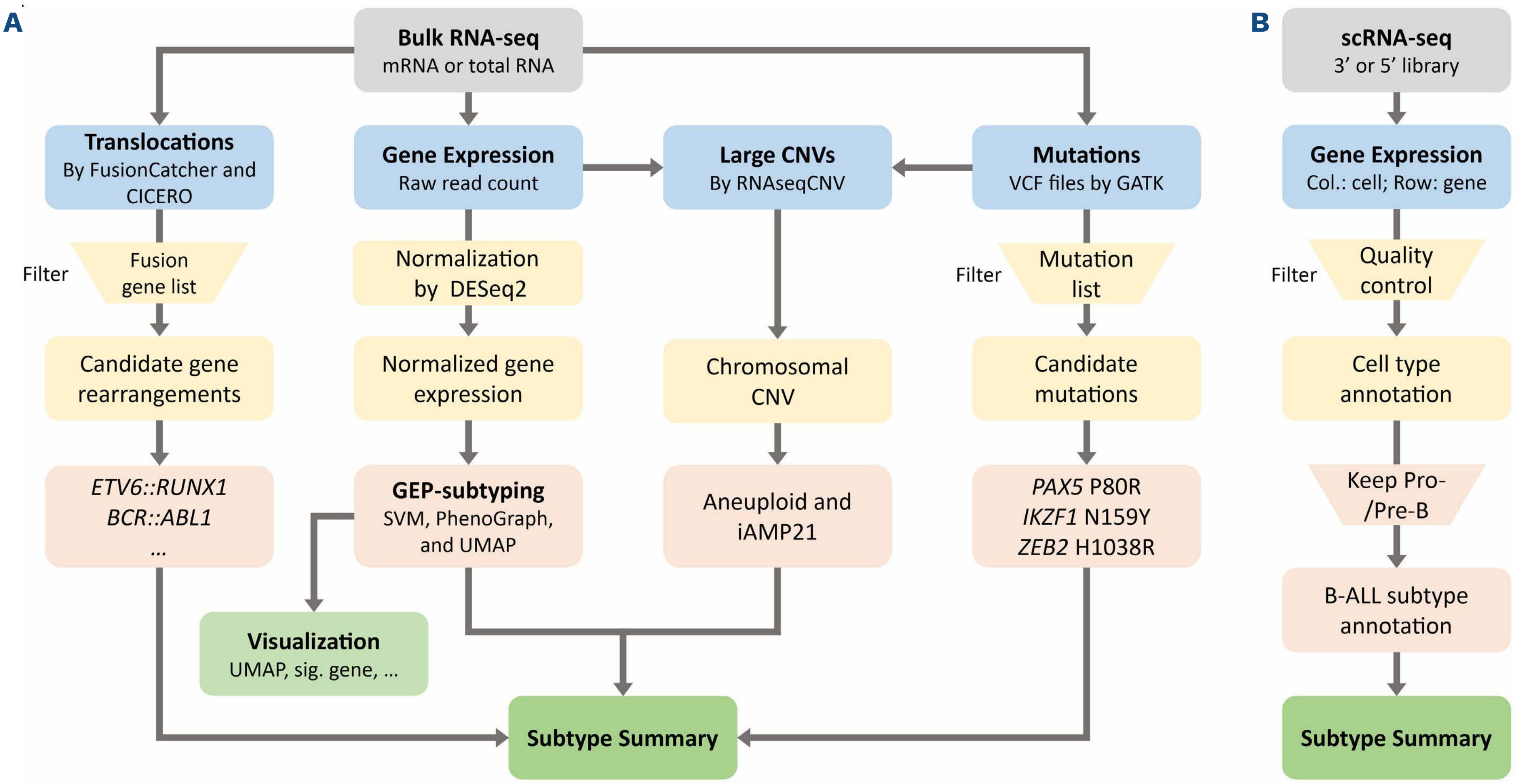


Figure 7. Summary of integrative B-acute lymphoblastic leukemia classification by MD-ALL. Molecular Diagnosis of Acute Lymphoblastic Leukemia (MD-ALL) accepts both bulk and single-cell (sc) RNA-sequencing (RNA-seq) data for B-acute lymphoblastic leukemia (B-ALL) classification. (A) Bulk analysis is the main function of MD-ALL, which accepts 3 types of standard output from bulk RNA-seq data: translocations (optional; raw output from FusionCatcher and/or CICERO), gene expression read count (required; called by HTSeq or FeatureCount), and sequence mutations (optional; Variant Call Format [VCF] files called by GATK). Based on the input data, four aspects of information will be identified: i) the input translocations are compared with an internal reference to identify signature fusion genes; ii) the gene expression data normalized from raw read count are analyzed by support vector machine (SVM) and PhenoGraph to predict the subtype and shown in a uniform manifold approximation and projection (UMAP) plot; iii) the variants in the provided VCF files are annotated to identify the signature gene mutations; and iv) the gene expression and mutation information are integrated by the RNAseqCNV package (see Methods) to identify chromosomal CNV, which will assist the identification of aneuploid and intrachromosomal amplification of chromosome 21 (iAMP21) subtypes. Then, a comprehensive subtype summary from the 4 aspects of information will be integrated to determine the subtypes of the test samples. (B) For scRNA-seq-based B-ALL classification, the input data is a count matrix with genes in rows and cells in columns. This read count matrix can be generated from either 3' or 5' scRNA-seq libraries using standard analysis pipelines. A basic quality control analysis is then performed to remove cells or genes with low sequencing coverage (see *Online Supplementary Appendix*). With the cell type gene expression profile reference, each test cell is annotated and only the B-lineage blast cells, which are pro- and pre-B cells, are retained for subsequent B-ALL subtyping, with results summarized in the report.

Thus, with minimal bioinformatics assistance to generate the raw information of GEP and genetic lesions, users can manage the subsequent analysis using MD-ALL to achieve integrative B-ALL classification.

Discussion

In this study, we present the first RNA-seq analysis platform capable of integrating both genetic lesions and GEP features for B-ALL classification. For more than 90% of the study cohort, the integrative analysis led to highly accurate B-ALL classification based on multiple layers of information. Additionally, the platform supplies detailed information for users to review and adjust the results as necessary.

This study is based on one of the largest B-ALL RNA-seq cohorts to establish a GEP reference representing all reported B-ALL subtypes, achieving high accuracy and sensitivity compared with alternative tools. By integrating genetic lesions, which other tools lack, subtypes can be determined more accurately, making this approach more feasible for future translational application in clinical settings.

Using the GEP reference compiled from bulk RNA-seq, we also explored the B-cell differentiation stages of different B-ALL subtypes. Our observations confirmed that certain B-ALL subtypes are blocked at early B-cell progenitor stages, while others progress to more mature stages. Moreover, some subtypes have been observed to have overlapping GEP features, such as iAMP21, PAX5alt, and Ph/Ph-like. Incorporating distinct B-cell differentiation patterns of different subtypes might be beneficial for better separation of these subtypes.

As genomic analysis advances towards single-cell resolution, we have demonstrated the feasibility of using GEP reference derived from bulk RNA-seq for accurate single-cell B-ALL classification in multiple subtypes. Currently, generating comparable samples size of single-cell data remains challenging due to technological and cost limitations. Moreover, scRNA-seq is unable to provide as comprehensive transcript abundance as bulk RNA-seq, and different scRNA-seq library preparation kits have been reported with larger batch effects compared with bulk RNA-seq. As a result, bulk RNA-seq remains the optimal platform for generating *bona fide* GEP signatures for each B-ALL subtype.

The classification of B-ALL subtypes using RNA-seq is revolutionizing clinical practice. Moreover, genomic data such as whole-genome sequencing can provide a more comprehensive understanding of genetic alterations. These results can further confirm the subtypes identified by RNA-seq. Importantly, genetic alterations can further differentiate patients within the same subtypes into more granular prognosis subgroups, making them critical complementary assays for B-ALL classification.[48,49]

In conclusion, we introduce MD-ALL, a highly reliable and accurate bioinformatics platform that serves the research and clinical fields for integrative B-ALL classification based on bulk or single-cell RNA-seq.

Disclosures

No conflicts of interest to disclose.

Contributions

ZH and ZG conceived and designed the study, analyzed the genomics data, developed classification models, and wrote the manuscript. ZJ, JL and HH assisted the data analysis. AM provided HPC support for the data analysis.

Acknowledgments

The authors would like to acknowledge the COH Center for Informatics, and the utilization of the POSEIDON platforms for data exploration, visualization, and analysis.

Funding

This work was supported by the Research Start-Up Budget from the Beckman Research Institute of the City of Hope, the Leukemia and Lymphoma Society Special Fellow Award, NIH/NCI Pathway to Independence Award R00 CA241297, Andrew McDonough B+ Childhood Cancer Research Grant, Leukemia Research Foundation New Investigator Grant, and The V Foundation for Cancer Research V Scholar Award to ZG.

Data-sharing statement

MD-ALL code, relevant datasets, and detailed tutorial are freely available from https://github.com/gu-lab20/MD-ALL.

References

1. Brady SW, Roberts KG, Gu Z, et al. The genomic landscape of pediatric acute lymphoblastic leukemia. Nat Genet. 2022;54(9):1376-1389.
2. Zhang J, McCastlain K, Yoshihara H, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet. 2016;48(12):1481-1489.
3. Gocho Y, Kiyokawa N, Ichikawa H, et al. A novel recurrent EP300-ZNF384 gene fusion in B-cell precursor acute lymphoblastic leukemia. Leukemia. 2015;29(12):2445-2448.
4. Gu Z, Churchman M, Roberts K, et al. Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. Nat Commun. 2016;7:13331.
5. Alaggio R, Amador C, Anagnostopoulos I, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia. 2022;36(7):1720-1748.
6. Arber DA, Orazi A, Hasserjian RP, et al. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. Blood. 2022;140(11):1200-1228.
7. Hiemenz MC, Oberley MJ, Doan A, et al. A multimodal genomics approach to diagnostic evaluation of pediatric hematologic malignancies. Cancer Genet. 2021;254-255:25-33.
8. Pui CH, Roberts KG, Yang JJ, Mullighan CG. Philadelphia chromosome-like acute lymphoblastic leukemia. Clin Lymphoma Myeloma Leuk. 2017;17(8):464-470.
9. Tran TH, Langlois S, Meloche C, et al. Whole-transcriptome analysis in acute lymphoblastic leukemia: a report from the DFCI ALL Consortium Protocol 16-001. Blood Adv. 2022;6(4):1329-1341.
10. Walter W, Shahswar R, Stengel A, et al. Clinical application of whole transcriptome sequencing for the classification of patients with acute lymphoblastic leukemia. BMC Cancer. 2021;21(1):886.
11. Makinen VP, Rehn J, Breen J, Yeung D, White DL. Multi-cohort transcriptomic subtyping of B-cell acute lymphoblastic leukemia. Int J Mol Sci. 2022;23(9):4574.
12. Schmidt BM, Brown LM, Ryland G, et al. ALLSorts: a RNA-Seq subtype classifier for B-cell acute lymphoblastic leukemia. Blood Adv. 2022;6(14):4093-4097.
13. Beder T, Hansen B-T, Hartmann AM, et al. The gene expression classifier ALLCatchR identifies B-precursor ALL subtypes and underlying developmental trajectories across age. Hemasphere. 2023;7(9):e939.
14. Gu Z, Churchman ML, Roberts KG, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet. 2019;51(2):296-307.
15. Waanders E, Gu Z, Dobson SM, et al. Mutational landscape and patterns of clonal evolution in relapsed pediatric acute lymphoblastic leukemia. Blood Cancer Discov. 2020;1(1):96-111.
16. Montefiori LE, Bendig S, Gu Z, et al. Enhancer hijacking drives oncogenic BCL11B expression in lineage-ambiguous stem cell leukemia. Cancer Discov. 2021;11(11):2846-2867.
17. Kimura S, Montefiori L, Iacobucci I, et al. Enhancer retargeting of CDX2 and UBTF::ATXN7L3 define a subtype of high-risk B-progenitor acute lymphoblastic leukemia. Blood.

2022;139(24):3519-3531.
18. Paietta E, Roberts KG, Wang V, et al. Molecular classification improves risk assessment in adult BCR-ABL1-negative B-ALL. Blood. 2021;138(11):948-958.
19. Jeha S, Choi J, Roberts KG, et al. Clinical significance of novel subtypes of acute lymphoblastic leukemia in the context of minimal residual disease–directed therapy. Blood Cancer Discov. 2021;2(4):326-337.
20. Li Z, Lee SHR, Chin WHN, et al. Distinct clinical characteristics of DUX4- and PAX5-altered childhood B-lymphoblastic leukemia. Blood Adv. 2021;5(23):5226-5238.
21. Li Z, Jiang N, Lim EH, et al. Identifying IGH disease clones for MRD monitoring in childhood B-cell acute lymphoblastic leukemia using RNA-Seq. Leukemia. 2020;34(9):2418-2429.
22. Qian M, Zhang H, Kham SK-Y, et al. Whole-transcriptome sequencing identifies a distinct subtype of acute lymphoblastic leukemia with predominant genomic abnormalities ofEP300andCREBBP. Genome Res. 2017;27(2):185-195.
23. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-2873.
24. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
25. Anders S, Pyl PT, Huber W. HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.
26. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-930.
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
28. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-883.
29. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303.
30. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773-782.
31. Tian L, Li Y, Edmonson MN, et al. CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. Genome Biol. 2020;21(1):126.
32. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv. 2014;011650. doi: https://doi.org/10.1101/011650 [preprint, not peer-reviewed].
33. Barinka J, Hu Z, Wang L, et al. RNAseqCNV: analysis of large-scale copy number variations from RNA-seq data. Leukemia. 2022;36(6):1492-1498.
34. Bansal V, Libiger O. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. BMC Bioinformatics. 2015;16:4.
35. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
36. Lee SHR, Antillon-Klussmann F, Pei D, et al. Association of genetic ancestry with the molecular subtypes and prognosis of childhood acute lymphoblastic leukemia. JAMA Oncol. 2022;8(3):354-363.
37. Levine JH, Simonds EF, Bendall SC, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184-197.
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
39. Kursa MB, Rudnicki WR. Feature selection with the Boruta Package. J Stat Softw. 2010;36(11):1-13.
40. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495-502.
41. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163-172.
42. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45(3):242-252.
43. Harvey RC, Mullighan CG, Chen IM, et al. Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia. Blood. 2010;115(26):5312-5321.
44. Rosenthal A, Younes A. High grade B-cell lymphoma with rearrangements of MYC and BCL2 and/or BCL6: double hit and triple hit lymphomas and double expressing lymphoma. Blood Rev. 2017;31(2):37-42.
45. Khabirova E, Jardine L, Coorens THH, et al. Single-cell transcriptomics reveals a distinct developmental state of KMT2A-rearranged infant B-cell acute lymphoblastic leukemia. Nat Med. 2022;28(4):743-751.
46. Witkowski MT, Dolgalev I, Evensen NA, et al. Extensive remodeling of the immune microenvironment in B cell acute lymphoblastic leukemia. Cancer Cell. 2020;37(6):867-882.
47. Caron M, St-Onge P, Sontag T, et al. Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity. Sci Rep. 2020;10(1):8079.
48. Ryan SL, Peden JF, Kingsbury Z, et al. Whole genome sequencing provides comprehensive genetic testing in childhood B-cell acute lymphoblastic leukaemia. Leukemia. 2023;37(3):518-528.
49. Leongamornlert D, Gutierrez-Abril J, Lee SW, et al. Diagnostic utility of whole genome sequencing in adults with B-other acute lymphoblastic leukemia. Blood Adv. 2023;7(15):3862-3873.