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

by Zunsong Hu, Zhilian Jia, Jiangyue Liu, Allen Mao, Helen Han, and Zhaohui Gu

Received: June 7, 2023.
Accepted: November 6, 2023.


Publisher’s Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors’ final approval; the final version of the manuscript will then appear in a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
MD-ALL: an integrative platform for molecular diagnosis of B-acute lymphoblastic leukemia

Zunsong Hu\textsuperscript{1,2}, Zhilian Jia\textsuperscript{1,2}, Jiangyue Liu\textsuperscript{3,4}, Allen Mao\textsuperscript{5}, Helen Han\textsuperscript{1,2}, Zhaohui Gu\textsuperscript{1,2}

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

Running Title: MD-ALL: molecular diagnosis of B-ALL

Corresponding author: Zhaohui Gu (zgu@coh.org)

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

One table, 6 figures, 4 supplementary figures, 12 supplementary tables, and one supplementary data file are associated with this manuscript.

Acknowledgements: 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 Z.G.

Author contributions: Z.H. and Z.G. conceived and designed the study, analyzed the genomics data, developed classification models, and wrote the manuscript. Z.J., J.L., and H. H. assisted the data analysis. A.M. provided HPC support for the data analysis.

Disclosures: None.
ABSTRACT

B-acute lymphoblastic leukemia (B-ALL) consists of dozens of subtypes defined by distinct gene expression profiles (GEPs) and various genetic lesions. With the application of transcriptome 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 MD-ALL (Molecular Diagnosis of ALL), an integrative platform featuring sensitive and accurate B-ALL classification based on GEPs 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 (GEPs) and constellations of genetic alterations\textsuperscript{1}. Through the application of transcriptome sequencing (RNA-seq), multiple novel B-ALL subtypes have been identified harboring recurrent genetic lesions and distinct GEPs\textsuperscript{2-4}. The current WHO Classification (5th edition) of Hematolymphoid Tumors (WHO-HAEM5)\textsuperscript{5}, along with the International Consensus Classification of Myeloid Neoplasms and Acute Leukemia (ICC)\textsuperscript{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\textsuperscript{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 GEPs (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\textsuperscript{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\textsuperscript{11-13}, they solely rely on GEP for B-ALL subtyping. Here, we present MD-ALL (Molecular Diagnosis of Acute Lymphoblastic Leukemia), 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-seq datasets

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

RNA-seq data analysis

The sequencing reads were aligned to the human genome reference (GRCh38) using STAR\textsuperscript{24}. Then Picard (see URLs) was used to mark PCR duplicates. Gene expression: Read counts were calculated by HTSeq\textsuperscript{25} and FeatureCount\textsuperscript{26}, and then normalized by DESeq\textsuperscript{27}. The ComBat function in the sva R package\textsuperscript{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 (Supplementary Figure 1). t-Distributed Stochastic Neighbor Embedding (tSNE) and Uniform Manifold Approximation and Projection (UMAP) were used for dimensionality reduction visualization. Mutations: Single nucleotide variants (SNVs) and insertions/deletions (Indels) were called by following the best practice pipeline from GATK (see URLs)\textsuperscript{29}. Deconvolution of bulk GEP: Single-cell RNA-seq (scRNA-seq) data from the 1-Million Immune Cells Project (see URLs) were reanalyzed to establish a GEP reference representing 20 primary blood cell types. Then, CIBERSORTx\textsuperscript{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\textsuperscript{31} and FusionCatcher\textsuperscript{32} were used as they can identify gene rearrangements involving highly repetitive regions such as the IgH locus. Copy number variation (CNV) calling: With read counts and SNVs called from RNA-seq, the RNAseqCNV package\textsuperscript{33} was used to detect
chromosomal level CNVs. *Ancestry inference:* The samples’ ancestral background was estimated using iAdmix\textsuperscript{34}, with the genotype of SNPs from the 1K-Genome Project used as the reference\textsuperscript{35,36}.

**GEP reference of B-ALL subtype**

Through analyzing the RNA-seq data of the 2,955 B-ALL samples, 26 subtypes were identified, with 19 having distinct GEP features. To construct a GEP reference for B-ALL classification, PhenoGraph clustering\textsuperscript{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\textsuperscript{38} was used to subsample or artificially construct additional samples, which resulted in cohorts with the same sample size per subtype. Then Boruta\textsuperscript{39} was used to identify the genes confirmed as contributing features for distinguishing different subtypes.

**GEP-based B-ALL classification**

Two GEP-based B-ALL prediction models were constructed: 1. support vector machine (SVM) classification. Among multiple tested machine learning algorithms, SVM performed the best. 2. PhenoGraph clustering\textsuperscript{37}. PhenoGraph is a clustering algorithm originally developed to identify and partition cells into subpopulations.

**Integration of genetic lesions and GEP 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.
scRNA-seq analysis and B-ALL classification

scRNA-seq reads were mapped to the GRCh38 reference. After quality control, the Seurat package\textsuperscript{40} was used for gene expression normalization and variable gene selection. With the GEP reference of blood cell types and B-ALL subtypes described above, SingleR\textsuperscript{41} was used to annotate cell types and B-ALL subtypes for each cell.

RESULTS

Characteristics of the RNA-seq cohort

In total, 2,955 B-ALL samples with high-quality RNA-seq data were included in this study (Supplementary Table 1). 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%) (Supplementary Figure 2). Through manual curation of the genetic lesions, 3,304 gene rearrangements, 2,979 sequence mutations, and 95 \textit{FLT3} internal tandem duplications (ITDs) were identified (Supplementary Tables 2-4). Subsequently, sentinel gene fusions and mutations were compiled to facilitate B-ALL classification (Supplementary Tables 5 and 6). 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, genders, 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 GEP-based B-ALL classification by MD-ALL

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 GEPs (Figure 1B). Using this GEP reference cohort, 1,058 feature
genes were consistently confirmed by the Boruta algorithm in eight SMOTE-resampled cohorts (Supplementary Table 7). Due to the substantial batch effect between mRNA-seq and total RNA-seq library preparation approaches (Supplementary Figure 1), 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 GEPs, such as Near haploid, and less recognized subtypes, such as Low hyperdiploid and CRLF2(non-Ph-like), as well as unclassified cases were excluded. 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. 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 vs. testing sample size ratio of 2:1 set for the major subtypes, fewer than 10 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 (Supplementary Table 8).

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\textsuperscript{11}, ALLSorts\textsuperscript{12}, and ALLCatchR\textsuperscript{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\textsuperscript{14}. This overlap may lead to overestimated accuracy of the alternative tools. Additionally, the \textit{PAX5::ETV6} fusion, originally reported as one of the sentinel alterations of PAX5alt subtype\textsuperscript{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 out of 938 ALLCatchR-classified), 94.8% (789 out of 832 ALLSorts-classified), and 97.1% (628 of 647 ALLSpice-classified), indicating that PhenoGraph is also a highly reliable prediction model for B-ALL subtyping (Supplementary Table 8). 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-seq analyses provide reliable and definitive B-ALL 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 strong Hyperdiploid GEP was observed in 8 cases (Supplementary Table 1). 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 (Supplementary Table 1).

In MD-ALL, users can provide raw translocations and sequence mutations for integrative B-ALL classification. Upon reanalysis of 2,955 RNA-seq samples, 96 sentinel gene rearrangements and 587 mutations were identified (Supplementary Tables 5 and 6). 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\textsuperscript{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 (Supplementary Figure 3 and Supplementary Table 9).

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 \textit{CRLF2} overexpression commonly seen in \textit{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 \textit{JAK2} p.R683 hotspot mutations, known for their high concurrence in \textit{CRLF2}-rearranged cases\textsuperscript{43}, further confirm the reliability of the \textit{IGH::CRLF2} fusion. MD-ALL then compiles all input information to assist the final subtype classification. For instance, a sample with an \textit{IGH::CRLF2} fusion and GEP-based Ph/Ph-like prediction, but lacking \textit{BCR::ABL1} fusion, can be definitively classified as Ph-like (Figure 3D). 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 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 (Supplementary Table 10). 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 (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 4 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-ALL 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 (Supplementary Table 11). 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, a malignancy transformed from more mature B lymphocytes. These conclusions agree with clinically reported immunophenotypic features of B-ALL subtypes as well as other digital deconvolution reports.

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 CIBERSORTx, and a high correlation was observed (correlation=0.85; 95% CI: 0.76 – 0.9; Figure 5C and Supplementary Table 12). 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-ALL 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.

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, we can reliably (>50% of the B-cell blasts) identify multiple B-ALL subtypes, such as KMT2A, ETV6::RUNX1/-like, Hyperdiploid, Ph, DUX4, MEF2D, PAX5alt, TCF3::PBX1, and ZNF384 (Figure 6 and Supplementary Figure 4), 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-ALL 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 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 (Supplementary Table 13). 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).

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 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\textsuperscript{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.
REFERENCES


### Table 1. Integrative criteria for B-ALL classification by MD-ALL

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

**Note:**

If genetic lesions do not agree with GEP-based prediction, genetic lesions determine the primary subtypes, while GEPs guide the decision on the secondary subtypes.

*Gene rearrangements involving ABL1, ABL2, CSF1R, PDGFR, PDGFRB, LYN, CRLF2, JAK2, EPOR, TSLP, TYK2, IL2RB, NTRK3, PTK2B, FGFR1, FLT3, DGKH, BLNK, and CBL.

iAmp, intragenic amplification.
FIGURE LEGENDS

Figure 1. Gene expression profiles (GEPs) of B-ALL subtypes

The tSNE plots display the GEP distribution using 1,058 signature coding genes identified from reference B-ALL subtypes (see Methods). GEPs are derived from bulk RNA-seq 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 GEPs. For GEP-based classification, Ph and Ph-like are combined as one Ph/Ph-like group.
Figure 2. High accuracy of B-ALL 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, SVM and PhenoGraph, were established within MD-ALL. *Phenocopy subtypes are identified by their similar GEPs 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. Gender information was inferred using the RNAseqCNV, 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.
Figure 3. Integrative summary of B-ALL classification by MD-ALL
A. Gene expression profile (GEP) -based subtype prediction by 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 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-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 MD-ALL to assist B-ALL classification. The genetic lesions, which include fusions, mutations, large-scale CNVs, will be integrated with GEP-based prediction by PhenoGraph and SVM to assist the classification of the test sample’s B-ALL subtype.
Figure 4. Integrative B-ALL classification pipeline
The integrative B-ALL classification pipeline implemented in MD-ALL consists of three steps:

Step 1. Define. In this step, MD-ALL integrates gene expression profiles (GEPs), 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, five additional rules were implemented to override the subtypes defined in Step 1, which include iAMP21, 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 will be 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.
Figure 5. Deconvolution of bulk gene expression profile of B-ALL subtypes

A. 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 URLs). B cells are classified into granular 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-erythocyte progenitor; NK cell, natural killer cell. B. Heatmap of different 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-ALL subtype classification at a single-cell level

A. scRNA-seq of a 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.
MD-ALL accepts both bulk and single-cell RNA-seq data for B-ALL classification. **a.** Bulk analysis is the main function of MD-ALL, which accepts three 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; VCF files called by GATK). Based on the input data, four aspects of information will be identified: 1) the input translocations are compared with an internal reference to identify signature fusion genes; 2) the gene expression data normalized from raw read count are analyzed by SVM and PhenoGraph to predict the subtype and shown in a UMAP plot; 3) the variants in the provided VCF files are annotated to identify the signature gene mutations; and 4) the gene expression and mutation information are integrated by RNAseqCNV to identify chromosomal CNVs, which will assist the identification of aneuploid and iAMP21 subtypes. Then, a comprehensive subtype summary from the four aspects of information will be integrated to determine the subtypes of the testing 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 standard 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 **Supplementary Methods**). With the cell type gene expression profile reference, each testing 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.

**Figure 7. Summary of integrative B-ALL classification by MD-ALL**
Figure 7

(A) Bulk RNA-seq
- mRNA or total RNA
- Gene Expression: Raw read count
- Normalization by DESeq2
- Filter
- Candidate gene rearrangements
  - ETV6::RUNX1
  - BCR::ABL1
  - GEP-subtyping: SVM, PhenoGraph, and UMAP
  - Visualization: UMAP, sig. gene, ...
- Subtype Summary

(B) scRNA-seq
- 3' or 5' library
- Gene Expression: Col.: cell; Row: gene
- Quality control
- Filter
- Mutation list
- Candidate mutations
- Aneuploid and iAMP21
- PAX5 P80R
- IKZF1 N159Y
- ZEB2 H1038R
- Cell type annotation
- Keep Pro-/Pre-B
- B-ALL subtype annotation
- Subtype Summary
SUPPLEMENTARY DATA

Supplementary Methods

RNA-seq datasets

To establish the training and validation cohorts, we collected raw RNA-seq datasets of 3,005 non-duplicate (according to sample ID) B-ALL samples from multiple published studies\(^1\)-\(^{11}\). Additionally, we inferred the genetic relationship of the enrolled samples using the KING toolkit\(^{12}\) based on the genotype of variants called from RNA-seq. We identified twenty pairs of samples as potential duplicates or related, and then removed the ones with relatively lower sequencing coverage. From the remaining 2,985 samples, we further excluded samples with low coding region coverage (<15% at 30-fold) or low B-cell ratio (<30%; estimated by the CIBERSORTx\(^{13}\); see Methods below). Eventually, 2,955 B-ALL samples with high quality RNA-seq data were kept as the primary dataset for this study (Supplementary Table 1).

RNA-seq data analysis

The raw RNA-seq data were analyzed using a uniform analysis pipeline described in our previous work\(^2\)-\(^4\). In brief, the sequencing reads were aligned to human genome reference GRCh38 using the STAR package (v2.7.6a)\(^{14}\). Gene annotation downloaded from the Ensembl database (v102; see URLs) was used for STAR mapping and the following read count evaluation. Then the Picard (v2.26.11; see URLs) was used to mark duplicates and generate the final bam files.

Gene expression level evaluation. Read count per gene was calculated by HTSeq\(^{15}\) and FeatureCount\(^{16}\), the two most popular tools for this purpose. Then gene expression level was normalized by the variance stabilizing transformation (VST) algorithm in the DESeq2 package\(^{17}\). With the VST gene expression data, R packages Rtsne and umap were used to map the samples to 2-dimentional \(t\)-Distributed Stochastic Neighbor Embedding (\(t\)-SNE) and Uniform Manifold Approximation and Projection (UMAP) plots using the top variable genes (based on median
absolute deviation). The ComBat function in the sva R package\textsuperscript{18} was used to correct the batch effects introduced by different library preparation kits and sequencing lengths (Supplementary Figure 1).

**Digital deconvolution of bulk GEP data.** To establish a GEP reference for annotating the primary blood cell types, we reanalyzed public single-cell RNA-seq (scRNA-seq) data of 166K cells obtained from eight healthy individuals used in the 1-Million Immune Cells Project (see URLs). Through stringent quality control, we established a GEP reference composed of over 10K cells representing 20 distinct cell types. To distinguish detailed differentiation stages of B cells, the annotation includes common lymphoid progenitors (CLP), pro-B1 (early pro-B), pro-B2 (late pro-B), pre-B1 (large pre-B), pre-B2 (small pre-B), immature B, mature B, and plasma cells. With the single-cell GEP reference, we used the CIBERSORTx\textsuperscript{13} to digitally deconvolute the bulk GEPs of B-ALL samples and delineate the composition of different cell types. The collective amount of B-lineage cells (pro-B1 to mature B) deconvoluted from the bulk samples were used to estimate leukemic cell ratios.

**Mutation detection from RNA-seq.** The short sequence mutation including single nucleotide variants (SNVs) and insertions/deletions (Indels) were called from RNA-seq by following the best practice workflow from the GATK forum (see URLs) as we reported before\textsuperscript{2, 4}. In brief, the bam files were processed by the SplitNCigarReads module of GATK (v4.2.2) to Splits reads that contain Ns in their cigar string. MuTect2 and HaplotypeCaller modules were used to call SNVs and Indels afterwards. The variants reported in the dbSNP (v152) and gnomAD (v3.1) databases as common single nucleotide polymorphisms (SNP; population minor allele frequency ≥ 1%) were removed. Then the remaining mutations were annotated to gene regions by VEP\textsuperscript{19} (v103). For B-ALL subtyping, the analysis was focused on a few signature mutations such as $PAX5$ P80R and other $PAX5$ mutations, $IKZF1$ N159Y, and $ZEB2$ H1038R. To further assist B-ALL subtyping, other signature mutations in gene $FLT3$, $IL7R$, $JAK1$, $JAK2$, $JAK3$, $KRAS$, $NRAS$, $PTPN11$, $NF1$,
IKZF3, and TP53 recorded in the COSMIC somatic mutation database (see URLs) were also reported.

**Fusion calling from RNA-seq.** CICERO\(^\text{20}\) (v0.3.0p2) and FusionCatcher\(^\text{21}\) (v1.33) were used as they can sensitively identify gene rearrangements involving highly repetitive regions such as the immunoglobulin heavy chain (IGH) locus. Since CICERO analysis may take a long time if the input bam files contain too many reads, we capped the bam files to 50 million reads for CICERO fusion calling. Normally, CICERO and FusionCatcher report dozens or even hundreds of fusions, but most of them are false positive. Therefore, we manually curated all the reported fusions to identify the reliable ones. Due to the complexity of DUX4 rearrangements, a few of them were rescued through manual inspection of aligned reads in the IGV browser\(^\text{22}\). Additionally, CICERO can identify the FLT3 ITD (Internal Tandem Duplication) from bulk RNA-seq both sensitively and accurately.

**Copy number variation (CNV) and iAMP21 calling from RNA-seq.** With read counts and SNVs called from RNA-seq, the RNAseqCNV package\(^\text{23}\) was used to detect chromosomal level CNVs. The gender information of the samples was also inferred by RNAseqCNV. Besides standard CNV analysis, RNAseqCNV also provides visualization results that can be used to identify intrachromosomal amplification of chromosome 21 (iAMP21) genetic lesions.

**GEP-guided detection of genetic lesions.** We detected and validated genetic lesions by using the expression level of specific genes or the overall GEPs. First, we compiled a list of candidate mutations and gene rearrangements that are signatures of different B-ALL subtypes. Then, we identified the genetic lesions that are consistent with the GEP features. For example, CRLF2 rearrangements are associated with CRLF2 overexpression, while DUX4 rearrangements are expected in DUX4 subtype defined by GEP. Similarly, GEP-defined PAX5 P80R subtype indicates both PAX5 P80R mutations and secondary PAX5 alterations.
Ancestry inference from RNA-seq

The ancestral background of enrolled samples was estimated using the iAdmix package\textsuperscript{24}, with the genotype of SNPs from the 1000 Genomes Project populations, which include European, African, Native American, East Asian, and South Asian, used as the reference\textsuperscript{25}. The genetic ancestral compositions of the test samples were quantified and then used to determine each ethnic group as described in previous reports\textsuperscript{26}.

Construct the GEP reference of B-ALL subtypes

Through integrative analysis of driver genetic lesions and GEPs, the enrolled 2,955 B-ALL samples were classified into 26 molecular subtypes, with 19 having distinct GEP features (Supplementary Table 1). To construct a GEP reference for B-ALL classification, we performed iterative sample selection using the PhenoGraph clustering\textsuperscript{27} and k-nearest neighbor (KNN) analysis of two-dimensional UMAP to identify the samples with stable and correct GEP clusters. In addition, the major subtypes with highly distinct GEPs, such as ETV6::RUNX1, KMT2A, DUX4, TCF3::PBX1, and MEF2D, were further trimmed to keep the sample size of training vs. test cohort as around 2:1.

GEP feature gene selection

Since the GEP reference cohort is not evenly distributed across different B-ALL subtypes, generic feature selection algorithms may favor the features of the major subtypes. To overcome this, cohorts with same sample size per subtype were generated by subsampling major subtypes and artificially constructing additional samples for minor ones using the SMOTE algorithm\textsuperscript{28}. Eight different samples sizes (n=10, 25, 50, 75, 100, 150, 200, and 250) per subtype were used to evaluate whether the feature genes can be stably identified. Then Boruta, a random-forest-based feature selection algorithm\textsuperscript{29}, was used to identify the genes confirmed as contributing features for
distinguishing different subtypes. Furthermore, to accommodate both mRNA and total RNA-seq libraries, only the protein-coding genes were considered for feature selection.

**GEP-based B-ALL classification model**

Using the feature genes and reference cohort described above, two GEP-based B-ALL prediction models were constructed: 1. support vector machine (SVM) classification. Among multiple machine learning algorithms, we observed that SVM performed the best. The reference samples from the 19 distinct subtypes were analyzed by SVM to train a prediction model using different numbers of feature genes (ranging from 100 to 1,058 genes in 11 rounds, with 100 as the interval). SVM algorithm with linear, polynomial, and Radial Basis Function kernels was tested in the GEP-based subtype prediction models and the accuracy for the 974 test samples was 96.1%, 95.1%, and 94.5%, respectively. Therefore, with the highest accuracy and faster training/predicting speed, the linear kernel of SMV was used for the final model. 2. PhenoGraph clustering. PhenoGraph is a clustering algorithm originally developed to identify and partition cells into subpopulations using high-dimensional single-cell mass cytometry data. Here it was applied to cluster the test samples with the reference cohort using different numbers of feature genes as described above for B-ALL classification. Ten neighbors were used in PhenoGraph analysis considering the smallest sample size for B-ALL subtypes in our training cohort is around 10. Since SVM and PhenoGraph models do not provide confidence score for classification, MD-ALL applies the 11 rounds of prediction using different numbers of genes to quantify the prediction reliability. A subtype is reported if the confidence score is above 0.5.

**Integration of genetic lesions and GEP features**

GEP-based subtype prediction and key genetic lesions identified from RNA-seq were integrated to assist definitive classification of B-ALL subtypes. For example, if the genetic lesions and GEP predictions point to the same subtypes, a highly reliable classification will be achieved. However, if GEP-based subtyping gives ambiguous prediction score or it is not consistent with the driver
genetic lesions, a knowledge-based decision-making is needed. For example, samples with both BCR::ABL1 fusion and hyperdiploid karyotype should be classified as Ph (BCR::ABL1) subtype, regardless of the GEP prediction. A detailed description of integrating GEP-based prediction and sentinel genetic lesions for B-ALL classification is summarized in Table 1.

**scRNA-seq analysis and B-ALL classification**

scRNA-seq reads were analyzed by the Cell Ranger (v6.0.1) pipeline using the human reference genome GRCh38. Genes expressed in at least 5 cells were retained, as were cells with a minimum of 200 expressed genes and less than 10% mitochondrial reads. Cells with gene counts exceeding the median plus 3 median absolute deviation of gene number were considered outliers and removed. Doublet cells identified by the DoubletFinder\(^\text{30}\) R package were also excluded. The Seurat\(^\text{31}\) (v4.0.5) was used for gene expression normalization and variable gene selection. With the GEP reference of blood cell types and B-ALL subtypes described above, the SingleR package\(^\text{32}\) was used to annotate the cell type and B-ALL subtype for each cell.

**URLs**

Ensembl, http://www.ensembl.org/;


1-Million Immune Cells Project, https://data.humancellatlas.org;


gnomAD, https://gnomad.broadinstitute.org/;
COSMIC database, https://cancer.sanger.ac.uk/cosmic
Reference


Supplementary Figures

A

Subtype
- BCL2/MYC
- CDX2/UBTF
- CRLF2(non-Ph-like)
- DUX4
- HLF
- IKZF1 N159Y
- Low hyperdiploid
- Hyperdiploid
- KMT2A
- N159Y
- Low hypodiploid
- NUP1
- MEF2D
- Other
- PAX5
- Ph-like
- ZEB2/CEBP

B

- Hyperdiploid
- iAMP21
- IKZF1
- N159Y
- KMT2A
- PAX5
- Ph
- ZEB2/CEBP

C

Library
- Stranded_mRNA_PE100bp
- Stranded_mRNA_PE150bp
- Stranded_totalRNA_PE125bp
- Stranded_totalRNA_PE150bp
- Unstranded_mRNA_PE100bp
- Unstranded_mRNA_PE75bp

D

- Stranded_mRNA_PE150bp
- Stranded_mRNA_PE75bp
- Stranded_totalRNA_PE125bp
- Stranded_totalRNA_PE150bp
- Unstranded_mRNA_PE100bp
- Unstranded_mRNA_PE75bp

E

RNA type
- total RNA
- mRNA

F

Strand
- Stranded
- Unstranded

G

Sequencing length
- PE100bp
- PE125bp
- PE150bp
- PE75bp
Supplementary Figure 1. Correction of batch effects in different RNA-seq datasets.
The t-SNE plots show the distribution of gene expression profiles (GEPs) for 2,955 B-ALL samples. This analysis is based on the top 1,000 most variably expressed genes with a perplexity parameter of 30 in t-SNE. In these plots, each point represents the GEP of one RNA-seq sample. These RNA-seq datasets, obtained from multiple sources, were generated using different library preparation kits and sequencing strategies. Therefore, substantial batch effects can be introduced, which are visible as distinct GEP clusters that appear to be driven by different RNA-seq batches (A and C). Once batch correction is applied (see Methods), the GEPs from different batches are seen to overlap evenly, indicating successful reduction of batch effects (B and D). Further investigation into different aspects of batch effects revealed that the mRNA vs. total RNA batches (E) introduces a greater batch effect compared to those from stranded vs. unstranded (F) and different sequencing lengths (G).
Supplementary Figure 2. Distribution of the demographic characteristics of the cohort.

A. The study cohort has a relatively equal representation of male and female cases, with gender determined by RNAseqCNV. Out of 2,407 samples with gender information, 2,384 (99.04%) inferred genders were consistent with the clinical report. B. The cohort includes both childhood and adult samples, with around two-thirds of the samples from pediatric cases. C. The race and ethnicity information were inferred by iAdmix based on the genotype of SNPs identified from RNA-seq. While the majority of the samples are of European descent, the cohort also includes individuals of Hispanic, East Asian, African, and other ethnic backgrounds, with a decent sample size.
Supplementary Figure 3. RNAseqCNV identifies iAMP21 genetic lesions.

The RNAseqCNV R package (Barinka et al. *Leukemia*, 2022) was initially developed to identify large scale CNVs on chromosomal or arm levels. Nonetheless, it can also identify the iAMP21 genetic lesions based on the unique gene expression and mutant allele frequency (MAF) patterns. The iAMP21 subtype is characterized by ≥ 5 RUNX1 copies per cell on a single abnormal chromosome 21 (Harrison et al., *Br J Haematol*. 2010), which exhibits elevated gene expression levels and a unique MAF density plot distribution compared with two or three copies of chromosome 21.
Supplementary Figure 4. Single-cell identification of multiple B-ALL subtypes

A, B and C, UMAP plots of B-ALL samples with ETV6::RUNX1, Hyperdiploid, or Ph subtypes, following the strategy described in Fig. 5. ETV6::RUNX1 subtype with distinct GEP (based on bulk RNA-seq) also achieves high accuracy (100%) for subtype prediction. By contrast, the subtypes with less distinct GEPs, such as Hyperdiploid and Ph, are observed with relatively lower yet still reliable subtype predictions (91.8% for Hyperdiploid and 89.4% for Ph). Raw scRNA-seq data were obtained from two published studies (Witkowski et al., Cancer Cell. 2020; Caron et al., Sci. Rep., 2020), where subtypes were all confirmed. D. The box plot displays the percentage of correct B-cell blasts classification for nine B-ALL subtypes (69 samples) at the single-cell level. Single-cell gene expression data was obtained from the Single-Cell Pediatric Cancer Atlas (ScPCA, see URLs). Each box depicts the interquartile range, spanning the 25th to the 75th percentiles. The median is represented by a horizontal line in the box. Whiskers extend from the boxes, typically encompassing up to 1.5 times the IQR. Colored dots represent the percentages of individual single-cell samples. The number of samples per subtype is shown in parentheses.

Online Supplementary tables 1-13 provided as Excel files.