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.

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-dimential 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[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[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[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[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[20] (v0.3.0p2) and FusionCatcher[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[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[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[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[25]. The genetic ancestral compositions of the test samples were quantified and then used to determine each ethnic group as described in previous reports[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[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[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[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[27]. 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[30] R package were also excluded. The Seurat[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[32] was used to annotate the cell type and B-ALL subtype for each cell.

URLs

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

The best practice workflow for calling SNVs and Indels from RNA-seq data, https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-q;

Picard, http://broadinstitute.github.io/picard;

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

ScPCA, https://www.alexslemonade.org/childhood-cancer-data-lab/single-cell-pediatric-cancer-atlas;

gnomAD, https://gnomad.broadinstitute.org/;

COSMIC database, https://cancer.sanger.ac.uk/cosmic

Reference

1. Gu Z, Churchman M, Roberts K, et al. Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. Nature communications. 2016;7(13331.

2. Gu Z, Churchman ML, Roberts KG, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nature genetics. 2019;51(2):296-307.

3. 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.

4. Montefiori LE, Bendig S, Gu Z, et al. Enhancer Hijacking Drives Oncogenic BCL11B Expression in Lineage-Ambiguous Stem Cell Leukemia. Cancer discovery. 2021;11(11):2846-2867.

5. 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.

6. Brady SW, Roberts KG, Gu Z, et al. The genomic landscape of pediatric acute lymphoblastic leukemia. Nature genetics. 2022;54(9):1376-1389.

7. 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.

8. 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 Discovery. 2021;2(4):326-337.

9. 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.

10. 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;

11. 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 Research. 2017;27(2):185-195.

12. 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.

13. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology. 2019;37(7):773-782.

14. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

15. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.

16. 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.

17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

18. 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.

19. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome biology. 2016;17(1):122.

20. 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 Biology. 2020;21(1):126.

21. 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.

22. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-26.

23. 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.

24. Bansal V, Libiger O. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. BMC bioinformatics. 2015;16(4.

25. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

26. 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;

27. 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.

28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16(321-357.

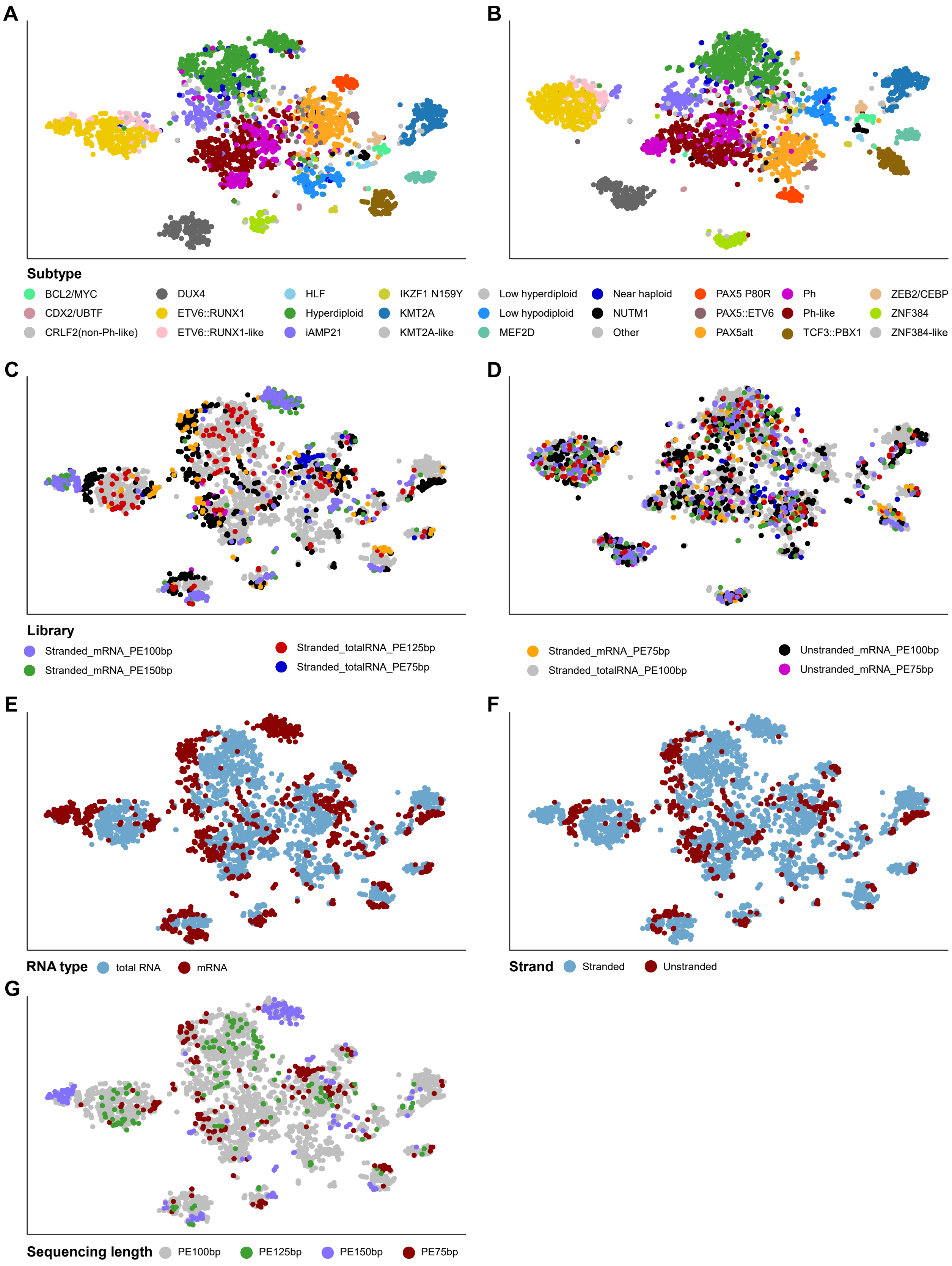
29. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(11):1-13.

30. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019;8(4):329-337 e324.

31. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015;33(5):495-502.

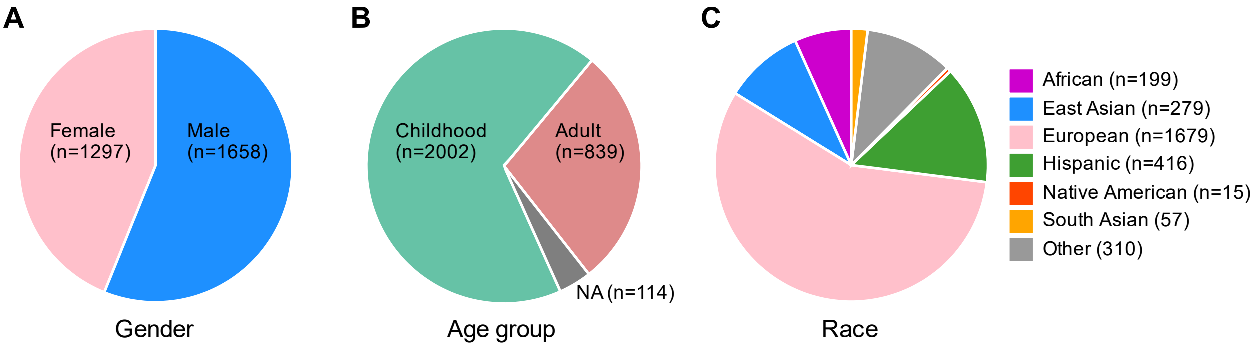
32. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163-172.

Supplementary Figures

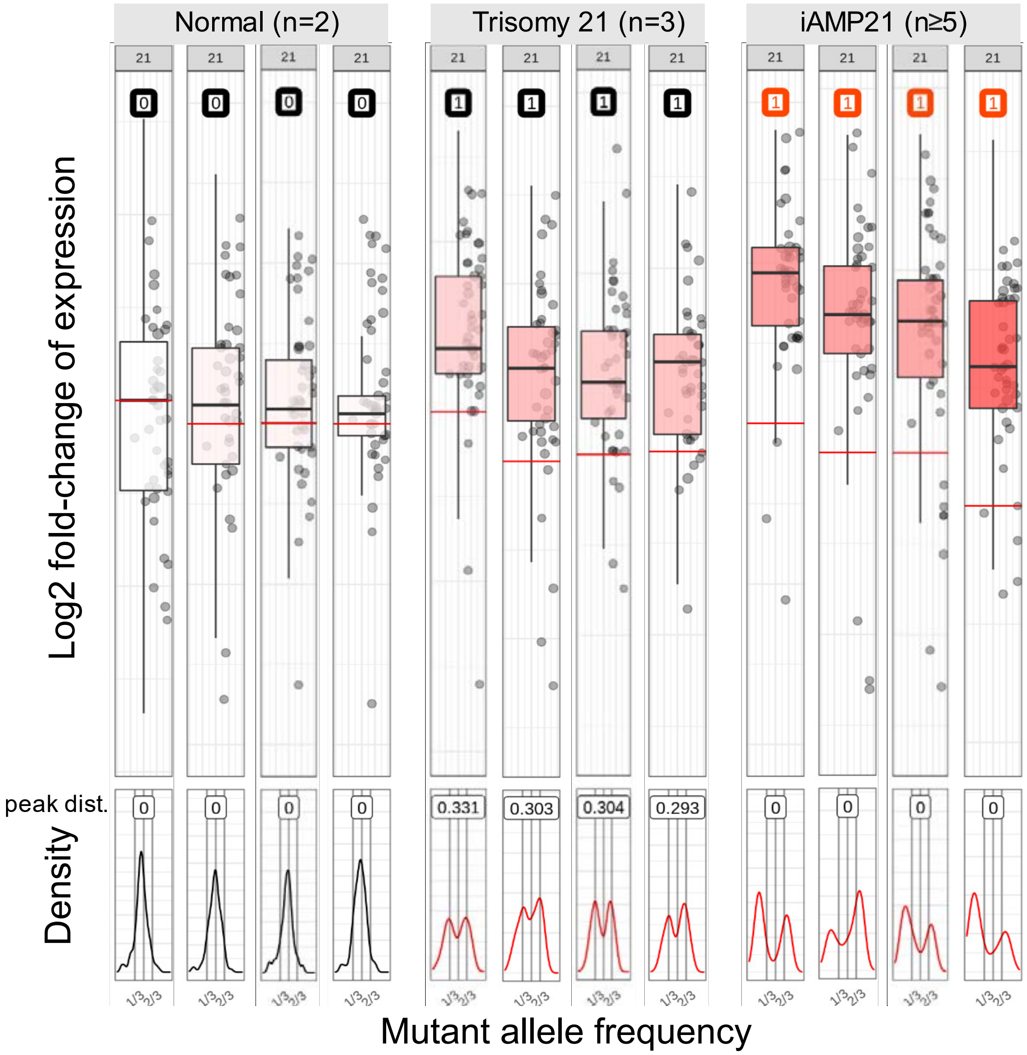


Supplementary Figure 1. Correction of batch effects in different RNA-seq datasets.

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**).

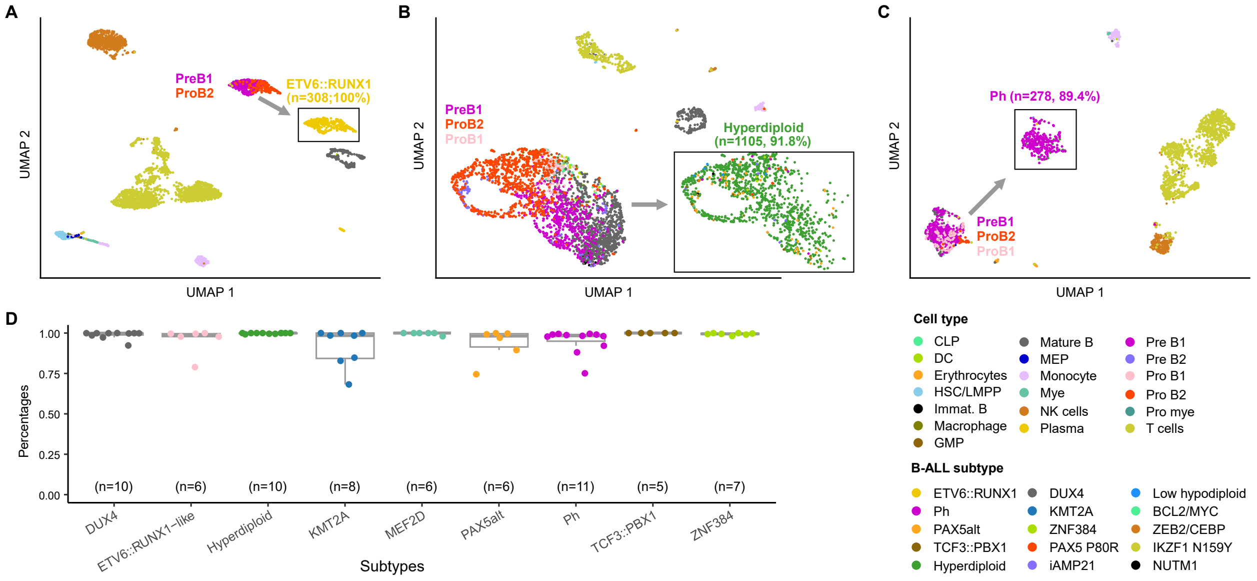


Supplementary Figure 2. Distribution of the demographic characteristics of the cohort. **A**. The study cohort has a relatively equal representation of male and female cases, with gender determined by RNAseqCNV. Out of 2,407 samples with gender information, 2,384 (99.04%) inferred genders were consistent 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.

Supplementary tables - see excel file