Transcriptional profiling directs the classification of acute leukemias of ambiguous lineage into AML, B-ALL, or T-ALL

Mulet-Lazaro, R.; Sijs-Szabó, A.; Hoogenboezem, R.M.; Herk, S. van; Exalto, C.; Koenders, J.E.; ... ; Sanders, M.A.

Received: 4 November 2024 | Accepted: 27 June 2025

DOI: 10.1002/hem3.70195

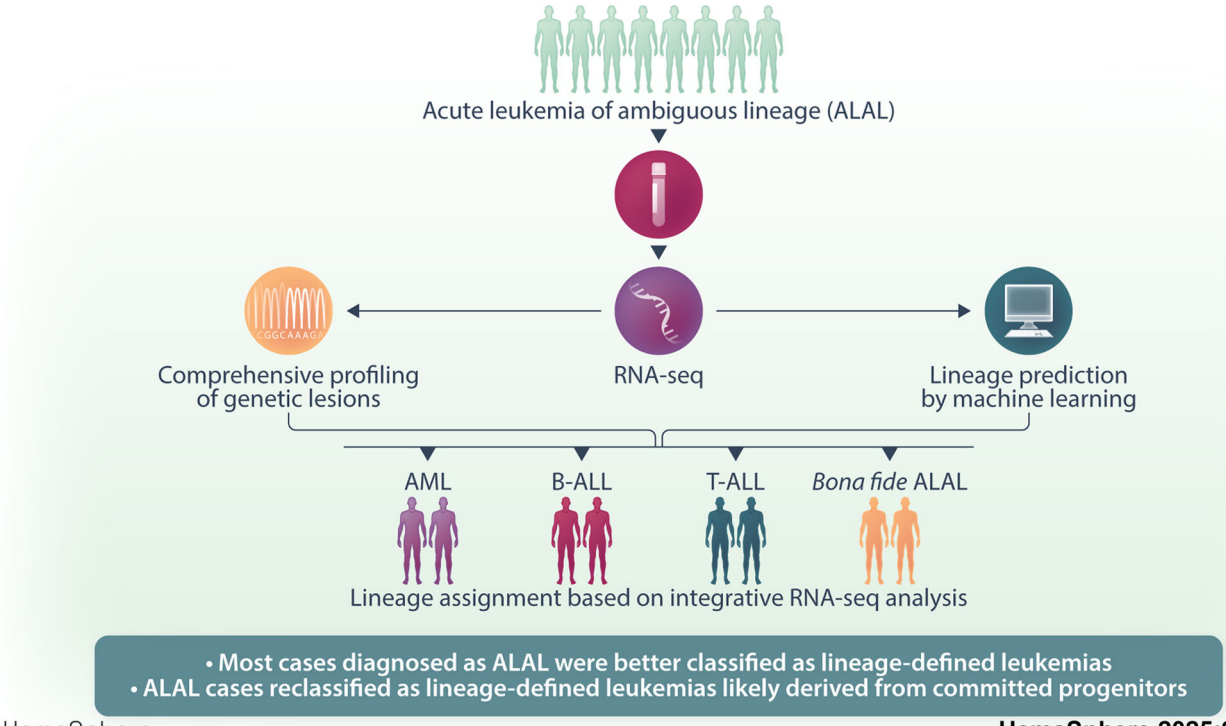
HemaSphere eha

ARTICLE

Transcriptional profiling directs the classification of acute leukemias of ambiguous lineage into AML, B-ALL, or T-ALL

Roger Mulet-Lazaro[1,^] | Anikó Sijs-Szabó[1,2,^] | Remco M. Hoogenboezem[1] | Stanley van Herk[1] | Carla Exalto[1] | Jasper E. Koenders[1] | Patricia G. Hoogeveen[3] | François G. Kavelaars[1] | Anita M. Schelen[1] | Willemijn van den Ancker[4] | Arjan A. van de Loosdrecht[4] | Charles G. Mullighan[5] | H. Berna Beverloo[6] | Vincent van der Velden[3] | Jan J. Cornelissen[1] | Peter J. M. Valk[1] | Anita W. Rijneveld[1,^] | Mathijs A. Sanders[1,^]

Graphical Abstract



Received: 4 November 2024 | Accepted: 27 June 2025

DOI: 10.1002/hem3.70195

ARTICLE

HemaSphere eha

Transcriptional profiling directs the classification of acute leukemias of ambiguous lineage into AML, B-ALL, or T-ALL

Roger Mulet-Lazaro[1,^] | Anikó Sijs-Szabó[1,2,^] | Remco M. Hoogenboezem[1] | Stanley van Herk[1] | Carla Exalto[1] | Jasper E. Koenders[1] | Patricia G. Hoogeveen[3] | François G. Kavelaars[1] | Anita M. Schelen[1] | Willemijn van den Ancker[4] | Arjan A. van de Loosdrecht[4] | Charles G. Mullighan[5] | H. Berna Beverloo[6] | Vincent van der Velden[3] | Jan J. Cornelissen[1] | Peter J. M. Valk[1] | Anita W. Rijneveld[1,^] | Mathijs A. Sanders[1,^]

Correspondence: Anita W. Rijneveld (a.rijneveld@erasmusmc.nl) and Mathijs A. Sanders (m.sanders@erasmusmc.nl)

Abstract

Acute leukemia of ambiguous lineage (ALAL) is a rare, poor-prognosis acute leukemia subtype that cannot be assigned to a single hematopoietic lineage. Although ALAL patients are typically treated with acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) regimens, optimal treatment choice is hindered by their lineage ambiguity. Therefore, we investigated the added value of transcriptomics for improving lineage assignment, currently based mainly on surface markers. First, we used an in-house pipeline to detect genetic lesions in RNA sequencing data ($n = 30$) with a sensitivity > 90% for small variants. Second, we compared ALAL gene expression profiles (GEPs) with representative AML ($n = 145$), B-ALL ($n = 223$), and T-ALL ($n = 85$) cases. In a principal component analysis (PCA), ALALs did not form a clear separate group, as most clustered with AML, B-ALL, or T-ALL. Accordingly, a machine learning classifier trained with GEPs of acute leukemias segregated 27/30 ALALs into myeloid-, B-, or T-lymphoid. These 27 cases harbored genetic abnormalities consistent with the classifier-assigned leukemia. Furthermore, deconvolution of ALAL GEPs revealed enrichment for signatures of normal hematopoietic cells corresponding to the leukemic type predicted by our algorithm. The classifier was also applied on an external ALAL cohort ($n = 24$), assigning 75% of the patients to a lineage matching their immunophenotypic and methylation profiles. In conclusion, integrative analysis of RNA sequencing data can accurately classify most ALAL cases as lineage-defined, while others show true transcriptional and epigenetic ambiguity driven by lesions like *BCL11B*. The pipeline and classifier developed here are valuable tools to improve ALAL diagnosis and guide therapeutic decisions.

INTRODUCTION

The delineation between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) can generally be established by morphological, immunophenotypic, cytogenetic, and molecular analyses. This classification is an important factor in determining the correct treatment regimen administered to the patient. However, 2%–3% of cases cannot be unequivocally assigned to any of these categories, since they express cell surface markers associated with multiple hematopoietic lineages.[1,2]

[1]Department of Hematology, Erasmus University Medical Center, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[2]Department of Hematology, Leiden University Medical Center, Leiden, The Netherlands
[3]Department of Immunology, Erasmus University Medical Center, Rotterdam, The Netherlands
[4]Department of Hematology, Amsterdam University Medical Center, Free University, Amsterdam, The Netherlands
[5]Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA
[6]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands
^These authors contributed equally to this work; Anita W. Rijneveld and Mathijs A. Sanders share senior authorship.

In 1991, Catovsky et al. designated these disorders as biphenotypic acute leukemias (BALs) and categorized them with an immunophenotype-based scoring system,[3] later refined by the European Group for the Immunological Classification of Leukemias (EGIL)[4] (Supporting Information S1: Figure 1). BALs show promiscuous expression of surface markers on the same blast, as opposed to bilineal acute leukemias, in which two separate populations of blasts from distinct lineages coexist.[1] In 2008, the World Health Organization (WHO) classification merged these two entities into mixed phenotype acute leukemias (MPALs), which are part of a broader category of acute leukemias of ambiguous lineage (ALALs) that also includes acute undifferentiated leukemias (AULs).[5]

Since the EGIL algorithm involved too many markers, some of which were not strictly lineage-specific, the WHO classification switched to a simpler scheme. Furthermore, it excluded cases harboring genetic aberrations associated with other acute leukemias, while recognizing subtypes defined by *KMT2A* rearrangements (present in 10%–15% of children[6,7] and 2%–5% of adults[8–10]) or *BCR*::*ABL1* (1%–3% of children[6,11] and 10%–30% of adults[8,10]). Following the discovery that ALALs with *ZNF384*[6] or *BCL11B* rearrangements[11,12] are transcriptionally distinct entities, these subtypes were added to the 2022 revision of the WHO classification.[13] These aberrations are frequent in children with ALAL (10%–13%[6,11] and 17%,[11] respectively), but less so in adults (4% for *ZNF384*[8,14] and 7% for *BCL11B*[8]), although precise estimates of their prevalence are limited by the scarcity of data. While the WHO classification and, more recently, the International Consensus Classification (ICC)[15] remain the gold standard in the field, their reliance on such a narrow set of immunophenotypic markers might be too restrictive for accurate differential diagnosis of ALAL.

Although certain genetic aberrations are overrepresented in ALAL, they can also be found in lineage-defined leukemias, as is the case for *ZNF384* rearrangements[16] and *BCR*::*ABL1*[17] in B-ALL, or *BCL11B* rearrangements in AML and T-ALL.[11] On the other hand, there are ALAL cases without these aberrations, indicating that other factors mediate in the development of the disease. For example, a study reported that a subtype of myeloid/lymphoid biphenotypic leukemias showed a shared epigenetic state, possibly reflecting their cell of origin, but no common ALAL-defining mutation was detected.[18] Existing evidence suggests that the lineage ambiguity characteristic is due to inherent developmental plasticity retained by the tumor cells, and for many cases, arises from the acquisition of specific genomic drivers in hematopoietic stem or progenitor cells.[6] However, a comprehensive understanding of the determinants of lineage ambiguity and aberrant epigenetic state is lacking.

Two possible explanatory hypotheses were initially put forward.[19] Ernest McCulloch proposed that their mixed phenotype could be the result of "lineage infidelity," whereby reprogramming of committed hematopoietic cells leads to the misexpression of cross-lineage markers.[20] In this scenario, it would be expected that the leukemic cells remain epigenetically and transcriptionally similar to their lineage of origin. Alternatively, Mel Greaves et al. theorized that "lineage promiscuity" arises from the multilineage potential of the cell origin, preserved upon transformation rather than acquired as a consequence of it.[21] The latter is consistent with the observations that early progenitors display multilineage gene expression before commitment to a single fate.[22] Nevertheless, more recent studies informed by genomics data have shown that genetic lesions, such as *BCL11B* or *ZNF384* rearrangements, may promote such lineage promiscuity when overexpressed in early hematopoietic progenitors.[6,11]

Despite significant progress in the understanding of ALAL biology,[6,11,14] optimal treatment remains challenging. As a result, the overall survival (OS) of ALAL patients is poor and comparable to patients with high-risk ALL or AML.[23] Most retrospective analyses suggest better outcome after ALL-like therapy compared to AML-like therapy, whereas combined AML-/ALL-therapy does not show clear benefits over ALL-like therapy alone.[9,19,24–26] Nevertheless, there is no consensus yet regarding the optimal therapeutic strategy to tackle ALAL. It stands to reason that a personalized treatment tailored to the biological features of each patient would lead to better outcomes.

Here, we investigated whether integrated mutational and transcriptional analyses of RNA sequencing (RNA-seq) data can improve lineage assignment in ALAL compared to conventional molecular genetic analyses. To this end, we profiled a cohort of 30 ALAL cases, including both bilineal and biphenotypic leukemias.

METHODS

An extended description of the methods is provided in Supporting Information S1: Supplemental Data.

Patients

Thirty adult ALAL patients classified as either BAL (EGIL,[4] $n = 28$) or MPAL (WHO 2022,[27] $n = 21$), treated between 2000 and 2025 at the Erasmus University Medical Center (EMC) or the Amsterdam University Medical Center and with available diagnostic material, were retrospectively selected. AUL patients were not included in any of the clinical trials linked to this study and therefore were not available. The study was approved by the ethics committee of the EMC (MEC-2015-155) and was conducted in accordance with the Declaration of Helsinki.

RNA sequencing

Bone marrow or peripheral blood samples were collected to perform molecular analyses. Mononuclear cells were separated using Ficoll density gradient centrifugation, ensuring a purified cell population. RNA-seq was performed on all 30 studied ALAL cases. RNA was isolated using the AllPrep DNA/RNA mini kit (Qiagen, #80204). RNA was converted into cDNA using the SuperScript II Reverse Transcriptase (Thermo Fischer Scientific) according to standard procedures. Sample libraries were prepped using 500 ng of input RNA according to the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Amplified sample libraries were paired-end sequenced (2×101 bp) on the NovaSeq 6000 (Illumina).

Identification of genetic lesions in RNA-seq data

RNA-seq data were aligned to the GRCh38 human reference genome using *STAR* (2.7.0f)[28] and processed with an in-house pipeline partially based on the GATK best practices, aimed at correcting possible sources of artifacts.[29] Point mutations were called with *HaplotypeCaller* and *Mutect2* from the GATK suite (v4.0.0),[30] and inserts/deletions with *Pindel* (v0.2.5b9).[31] Fusion genes were detected using *FusionCatcher* (v1.3),[32] *STAR-Fusion* (v1.10),[33] and *Arriba* (v2.2.21),[34] and copy number alterations (CNAs) were detected using *SuperFreq* (v1.3.2).[35] Other structural variants, such as large deletions, were identified by an in-house algorithm that extracts anomalous splicing junctions absent in the Ensembl database (v104).[36] All genetic lesions were manually inspected and validated using the *Integrative Genomics Viewer (IGV)*.[37]

Predictive signature based on gene expression profiles

To refine the lineage assignment of ALAL, we generated a machine learning classifier that can discriminate between acute leukemias based on gene expression profiles (GEPs) of AML ($n = 145$), T-ALL ($n = 85$), and B-ALL ($n = 223$). First, we log-transformed the TPM values produced by Salmon with a pseudo-count of 1 and selected the 8000 most variable genes. Then, we trained predictive models on the GEPs of lineage-restricted acute leukemias using six machine learning algorithms: multinomial logistic regression with lasso regularization (MLR), random forests (RF), gradient boosting machines (GBM), support vector machine with a radial kernel (SVM), k-nearest neighbors (KNN), and linear discriminant analysis (LDA). To optimize the hyperparameters and estimate the generalization performance of each model, we used nested cross-validation with 5 inner folds and 10 outer folds, implemented in the R packages *nestedcv* (0.7.10) and *caret* (7.0.1). The models were further validated on external data from the Munich Leukemia Laboratory (MLL). Standard performance metrics were computed using the *yardstick* (1.3.2) package.

Upon selection of the best machine learning algorithm (i.e., MLR), we performed a second round of nested cross-validation, with 10 inner and 10 outer folds, to tune the lambda hyperparameter. The final model, called "Expression-driven Classification of Acute Leukemias" (E-CAL), was then applied to the ALAL GEPs, yielding a matrix of probabilities corresponding to each leukemic class for each patient sample. Based on these predictions, a single leukemic lineage was assigned if it had a probability at least 0.25 greater than either of the other two lineages. Otherwise, the two leukemic lineages with the highest probabilities were assigned, provided that they were within a 0.25 probability range from each other. If none of these conditions were met, the leukemia was considered trilineage. Finally, to identify which genes were used in the prediction of each leukemia class, we extracted the nonzero coefficients from the model.

Bioinformatics and statistical analyses

Statistical tests were conducted with R version 4.3.2 unless otherwise specified. Most plots were generated using the *ggplot2* R package, whereas heatmaps were created using *ComplexHeatmap*.

RESULTS

ALAL patients show poor clinical outcome

We selected 30 adult leukemia patients classified as BAL ($n = 28$) and/or MPAL ($n = 21$) according to the EGIL[4] and WHO 2022[27] criteria, respectively (Table 1, Supporting Information S2: Table 3). The latter were further subcategorized into 15 B/myeloid (B/M) cases, 5 T/myeloid (T/M), and 1 B/T-lymphoid (B/T) based on their immunophenotype. No AUL cases were available in the clinical trials linked to this study. The two classification systems overlapped in 18/30 patients, with the remaining cases being classified as either ALL or AML in one system. Unless otherwise specified, in this study, we collectively refer to all cases as ALAL regardless of which set of criteria they met.

In addition to lineage-defining markers, 95% of ALAL cases were positive for the CD34 surface antigen (Supporting Information S2: Table 4). Results of cytogenetic analysis were available for 21 of the 30 cases, revealing that 90% of them showed an aberrant karyotype (Supporting Information S2: Table 5). The median age of the included patients was 55 years (range 21–75 years, Supporting Information S2: Table 3). Although complete remission (CR) was reached in 70% of cases (Supporting Information S2: Table 3), survival for the total ALAL group of 30 patients was poor, with a 2-year relapse-free survival (RFS) of 19% (95% CI: 7%–36%) (Figure 1A) and a 2-year OS of 38% (95% CI: 20%–56%) (Figure 1B). The median OS was 11.6 months (95% CI: 5.8–30.9). Five years after diagnosis, RFS and OS were 15% (95% CI: 5%–32%) and 23% (95% CI:9%–40%), respectively, comparable to high-risk AML and ALL.[38,39]

Fourteen patients received ALL-type therapy, including four who were treated in combination with a tyrosine kinase inhibitor (TKI). Another 14 patients received AML-type therapy, two of whom were switched to ALL-type therapy after the first induction due to refractory disease. Two patients did not undergo active treatment (Supporting Information S2: Table 3).

Both ALL- and AML-type therapies were administered in accordance with the applicable intensive treatment protocols established by the Dutch–Belgian Cooperative Trial Group for Hematology–Oncology (HOVON), which included HOVON37, HOVON70, HOVON71, and

TABLE 1 Patient characteristics.

Parameters	*N* (%)
Total no. of patients	30
Median follow-up, months (IQR)	71 (65–79)
Age, years	
Median (range)	55 (21–75)
EGIL classification	
MPAL	28 (93)
ALL	2 (7)
AML	0
WHO classification	
MPAL	21 (70)
ALL	3 (10)
AML	6 (20)
Cytogenetic subtype	
BCR::ABL1	6 (20)
KMT2A-rearranged	3 (10)
WBC count at diagnosis	
<30	16 (53)
30–100	9 (30)
≥100	4(13)
Unknown	1 (3)
Year of diagnosis	
Median (range)	2006 (2000–2025)
Treatment type	
ALL	10 (33)
ALL + TKI	4 (13)
AML	12 (40)
AML → ALL	2 (7)
No treatment	2 (7)
Allogeneic SCT	10 (33)

Abbreviations: ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; EGIL, European Group for the Immunological Classification of Leukemias; IQR, interquartile range; MPAL, mixed phenotype acute leukemia; SCT, stem cell transplantation; TKI, tyrosine kinase inhibitor; WBC, white blood cell; WHO, World Health Organization.

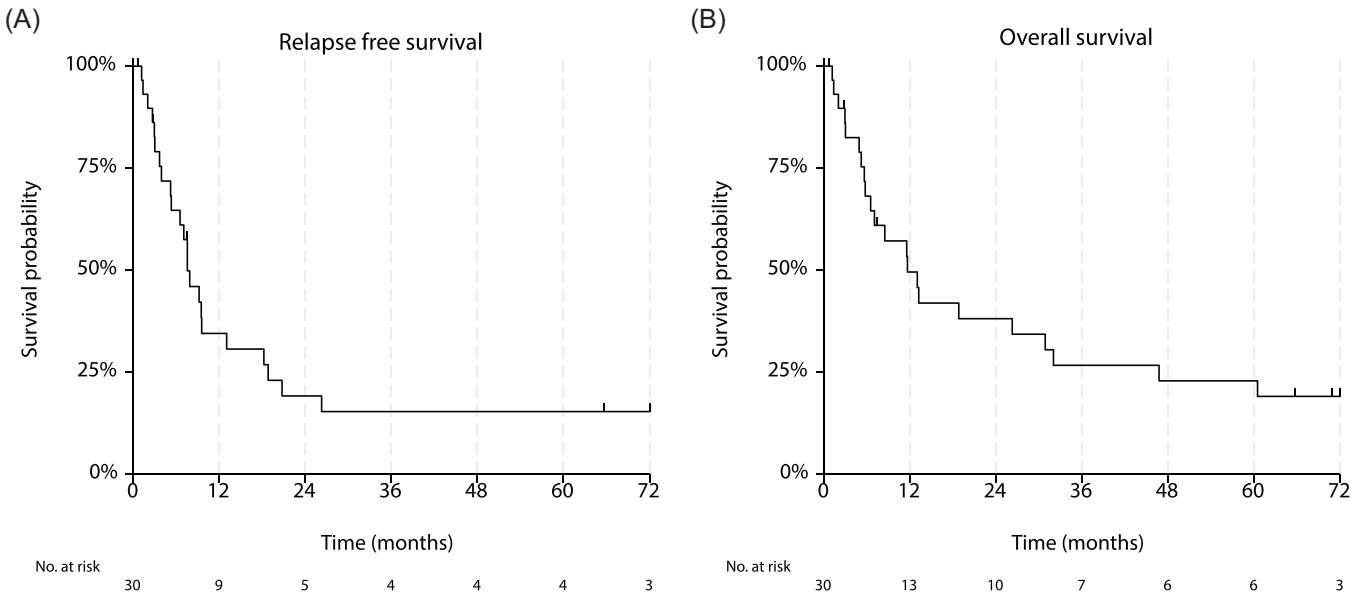


FIGURE 1 **Acute leukemia of ambiguous lineage (ALAL) patients exhibit a poor clinical outcome.** Kaplan–Meier plots of **(A)** relapse-free survival and **(B)** overall survival in ALAL patients with a maximum follow-up of 8 years.

HOVON100 for ALL and HOVON42, HOVON42A, HOVON43, and HOVON81 for AML (https://hovon.nl/en). Allogeneic stem cell transplantation was performed in 10 of the 30 patients (Supporting Information S2: Table 3), including all those who achieved long-term survival.

RNA-seq is suitable for detecting somatic mutations in ALAL

To characterize the mutational landscape of our cohort of adult ALALs, we implemented a computational pipeline comprising in-house and publicly available tools to identify genetic lesions from RNA-seq ($n = 30$) data. We validated and expanded the results of this analysis with a battery of standard diagnostics techniques, including karyotyping ($n = 21$), multiplex ligation-dependent probe amplification (MLPA, $n = 24$), and targeted DNA sequencing (DNA-seq, $n = 19$) (Figure 2A,B, Supporting Information S2: Table 6). It is noteworthy that the depth of coverage of the RNA-seq was sufficient for sensitive detection of variants in the vast majority of cases, with the exception of a few lineage-specific genes.

The RNA-seq pipeline correctly called 98% of single-nucleotide variants (SNVs) and insertions/deletions (indels) detected by DNA-seq with a variant allele frequency (VAF) greater than 5% (Supporting Information S2: Table 7). Moreover, 100% of the fusion genes identified by cytogenetics and 67% of the large copy number alterations (CNAs) were detected by RNA-seq (Figure 2B,C, Supporting Information S2: Table 8). On the other hand, the analysis revealed several genetic events missed by cytogenetics, chiefly copy number neutral losses of heterozygosity (CNN-LOHs), sub-chromosomal aberrations, and fusion genes.

Since G-banded karyotyping is prone to error, we generated single nucleotide polymorphism (SNP) array data for four ALAL patients with discrepancies between cytogenetics and RNA-seq. In three out of four, the copy number profiles inferred from RNA-seq were fully concordant with those derived from the SNP array (Supporting Information S1: Figure 2A–D), yielding a 100% recall for clonal CNAs (76% if including subclonal events). Disagreements with cytogenetics were favorably resolved to RNA-seq. For example, patient #22679 was reported to have a del(3p14) and a −7 by karyotyping, whereas both RNA-seq and SNP array identified CNN-LOH in 3p and a deletion restricted to 7p. In the remaining case (#24139), RNA-seq showed genome-wide chromosomal losses, as opposed to the high hyperdiploidy detected by cytogenetics (Supporting Information S1: Figure 2D). Considering that this patient was classified as B/M MPAL and harbored a homozygous *TP53* mutation, which is associated with hypodiploidy in B-ALL,[40,41] we surmised that the patient had undergone a duplication of an aneuploid genome, resulting in "masked hypodiploidy" that would explain the high chromosomal count.[42,43] Besides, the pattern of chromosomal losses was consistent with previous studies.[40] Taking this into account when interpreting this copy number profile, RNA-seq and SNP array showed a 78% concordance for this complex case.

To detect gene-level CNAs, we integrated the results of SuperFreq, based on depth of coverage and B-allele frequency, with those of an in-house tool that exploits anomalous splice junctions to find structural variants. This combined approach found 56% of the focal CNAs reported by MLPA. Events that exclusively affected an entire gene, too small to reliably call with SuperFreq, yet too large to affect splice junctions, were often missed. Finally, T-cell receptor (TCR) and immunoglobulin (Ig) rearrangements inferred from RNA-seq were consistent with the results from polymerase chain reaction (PCR)-heteroduplex assays in 59% of the cases (Figure 2B, Supporting Information S2: Table 3).

Overall, our pipeline achieved good accuracy across a wide range of genetic abnormalities, suggesting that RNA-seq can serve as a comprehensive diagnostic tool for ALAL, potentially replacing other molecular biology techniques currently in use.

ALAL has a heterogeneous mutational landscape

The ALAL patients had a median of four genetic lesions per patient (range: 2–8), excluding gross chromosomal aberrations (Figure 2B, Supporting Information S2: Table 6). Their mutational landscape was

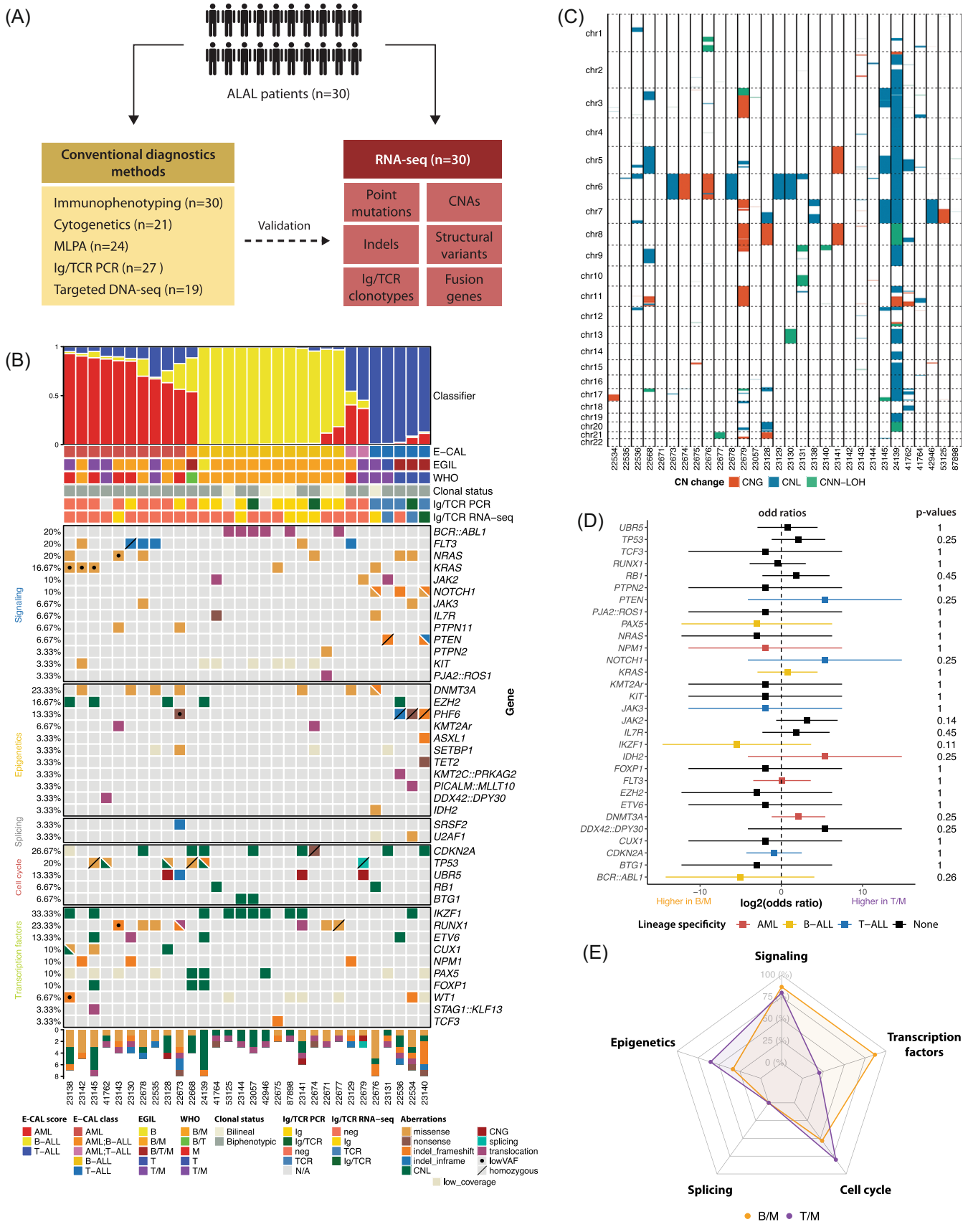


FIGURE 2 (See caption on next page).

FIGURE 2 **Comprehensive molecular diagnosis of acute leukemia of ambiguous lineage (ALAL) with RNA sequencing (RNA-seq). (A)** Overview of diagnostic modalities applied to the ALAL cohort, including conventional methods (left branch, in yellow) and RNA-seq (right, in red). Using a custom computational pipeline, we detected a wide range of genetic lesions using RNA-seq, which we validated with conventional approaches. **(B)** Oncoprint displaying single-nucleotide variants (SNVs), small insertions and deletions (indels), or copy number alterations (CNAs) affecting potential driver genes. Columns correspond to patients, and rows correspond to genes, grouped by molecular categories (indicated on the left). Different variants are represented in distinct colors, as shown in the plot legend. The annotation rows at the top indicate, in ascending order, the presence of immunoglobulin (Ig)/T-cell receptor (TCR) rearrangements detected by either RNA-seq or polymerase chain reaction (PCR), the classification of each patient according to World Health Organization (WHO) or European Group for the Immunological Classification of Leukemias (EGIL) criteria, whether the leukemia is biphenotypic or bilineal, and the results of the machine learning classifier developed in this study (Expression-driven Classification of Acute Leukemias, E-CAL). **(C)** Heatmap displaying CNAs detected by SuperFreq in RNA-seq data, where red corresponds to copy number gain (CNG), blue corresponds to copy number loss (CNL), and green corresponds to copy number neutral loss of heterozygosity (LOH). **(D)** Forest plot showing the odds ratio of recurrent mutations ($n > 2$) in T/myeloid (T/M) mixed phenotype acute leukemia (MPAL) relative to B/myeloid (B/M) MPAL, expressed in a logarithmic scale with a 0.1 offset ($n = 13$). The horizontal lines define the 95% confidence interval. On the right, the corresponding P-value from a Fisher's exact test comparing T/M MPAL and B/M MPAL is shown. The colors indicate that these genetic lesions are specific to acute myeloid leukemia (AML) (red), B-acute lymphoblastic leukemia (ALL) (yellow), or T-ALL (blue), determined as those significantly enriched in one disease with respect to the other two and present in at least 5% of those patients. **(E)** Spider plot depicting the proportion of T/M MPAL (orange) and B/M MPAL (purple) cases with mutations in five distinct molecular categories. MLPA, multiplex ligation-dependent probe amplification; VAF, variant allele frequency.

highly heterogeneous, but recurrent lesions were found in signaling pathways (86% of cases), transcription factors (77%), epigenetic modifiers (57%), and cell cycle regulators (53%). The most frequently mutated genes were *IKZF1* (33%), *CDKN2A* (27%), *RUNX1* (23%), and *DNMT3A* (23%).

In keeping with their phenotypic ambiguity, several ALAL patients harbored combinations of mutations strongly associated with different lineage-defined leukemias (Supporting Information S1: Figure 3A, Supporting Information S2: Tables 9 and 10), such as *TET2* and *NOTCH1* in #23140 or *DNMT3A* and *IKZF1* in #23141. However, as previously reported,[14,44] T/M MPAL and B/M MPAL cases were enriched for mutations associated with T-lymphoid and B-lymphoid malignancies, respectively (Figure 2D). Despite clear differences in the distribution of lineage-specific mutations in each subgroup, the small sample sizes limited the power of the analysis to detect significant associations. Thus, we extended it with published data,[8,14,45] which largely confirmed our previous findings (Supporting Information S1: Figure 3B). When grouped by pathway, B/M cases had a higher proportion of mutations involving transcription factors and cell cycle regulators, whereas splicing- and epigenetic-related genes were more often targeted in T/M (Figure 2E).

Transcriptional profiling guides the classification of adult ALAL

According to the current guidelines, ALAL classification is based on a limited series of immunophenotypic markers, which may fail to capture the true identity of the leukemic cell. Transcriptional profiling provides a better picture of cellular function than their immunophenotype, enabling a more accurate determination of the cell of arrest. Thus, in order to refine lineage assignment, we compared the GEPs of ALAL patients ($n = 30$) with those of AML ($n = 145$), B-ALL ($n = 223$), and T-ALL ($n = 85$). In a PCA, ALALs did not appear as a clear separate group, in line with previous publications showing that they are often not a distinct clinical entity at the transcriptional level[6] (Figure 3A). Instead, the majority of cases clearly clustered with one of the lineage-restricted leukemias, pointing to a well-defined cellular identity despite the misexpression of cell surface markers. It is noteworthy that early T-cell precursor (ETP)-ALL cases mainly clustered with other T-ALLs, with one exception that clustered among AML cases in the PCA (Supporting Information S1: Figure 3C).

Thus, we surmised that transcriptional profiles could be used to further refine the classification of most ALAL cases into AML, B-ALL, or T-ALL. To this end, we trained predictive models on the GEPs of lineage-restricted acute leukemias using six different machine learning algorithms, namely, MLR, RF, GBM, SVM, KNN, and LDA. All methods showed excellent performance in the 10 outer folds of nested cross-validation (Supporting Information S1: Figure 4A,B, Supporting Information S2: Table 11), with RF and KNN slightly below the others. However, evaluation of these models on an external data set from the Munich Leukemia Laboratory revealed that MLR was ahead of the rest in all performance metrics, with RF a close second (Supporting Information S1: Figure 4C,D, Supporting Information S2: Table 11).

Therefore, we selected the MLR model, which achieved nearly perfect performance in nested cross-validation (100% precision and recall) and maintained good generalizability on the external data (97.5% recall and 93.7% precision). We henceforth refer to this model as Expression-driven Classification of Acute Leukemias (E-CAL) (Figure 3B). Hierarchical clustering, PCA, and Uniform Manifold Approximation and Projection (UMAP) confirmed that clustering of lineage-defined leukemias and ALALs was largely preserved when using only the 41 genes selected by the lasso regularization (Figure 3C, Supporting Information S1: Figure 5A,B, and Supporting Information S2: Table 12). Unsurprisingly, the coefficients predictive of each class in our E-CAL model were enriched for gene sets associated with the corresponding lineage (Supporting Information S1: Figure 5C).

When applied to the ALAL GEPs, the model segregated 27/30 cases into myeloid (ALAL-M), B-lymphoid (ALAL-B), or T-lymphoid (ALAL-T) leukemias (Figure 2B, Supporting Information S2: Table 13). The remaining three patients received scores compatible with several possible lineages and were therefore labeled as "multilineage." Remarkably, patients assigned to a lineage-defined leukemia harbored genetic abnormalities known to be associated with that disease, such as *DNMT3A* and *NPM1* in ALAL-M, *IKZF1* in ALAL-B or *NOTCH1*, and *PHF6* in ALAL-T (Figure 2B, Supporting Information S2: Table 6). When grouped by pathways, ALAL-M cases were enriched for mutations in epigenetic modulators and transcription factors, whereas cell cycle lesions were more common in ALAL-T and signaling mutations in both ALAL-T and ALAL-B (Figure 3D, Supporting Information S1: Figure 5D). Furthermore, Ig/TCR rearrangements were detected (by either PCR or RNA-seq) in all patients classified as B-ALL and T-ALL, respectively (Figure 2B). Although Ig was also rearranged in three cases labeled as myeloid, such rearrangements have been previously reported in bona fide AML.[46]

Altogether, we show that most ALALs in our cohort can be transcriptionally segregated into AML, B-ALL, or T-ALL, yielding a refined classification that is consistent with their mutational status and the presence of Ig/TCR rearrangements.

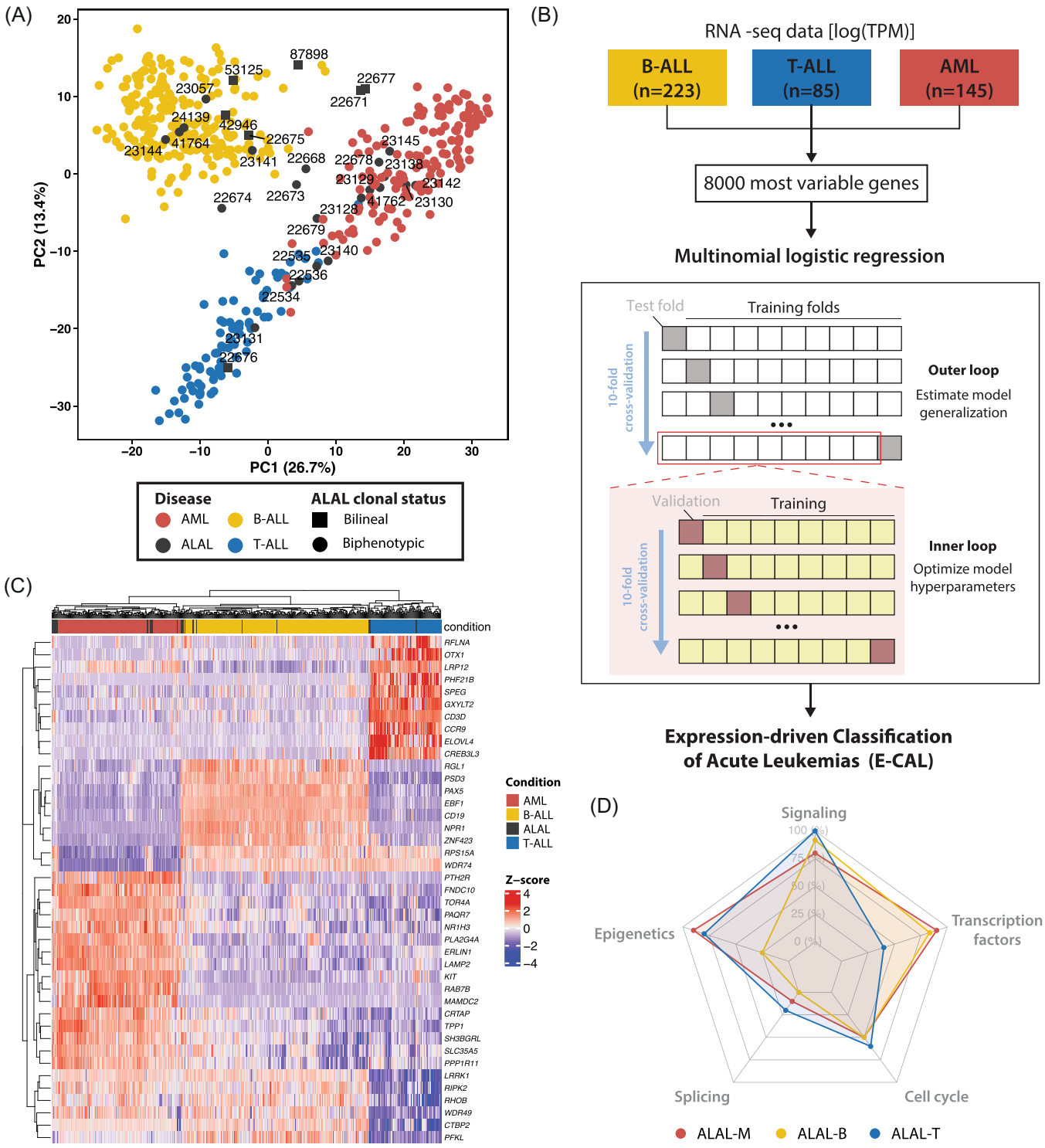


FIGURE 3 **Transcriptional profiling refines the classification of acute leukemia of ambiguous lineage (ALAL). (A)** Principal component analysis based on the 1000 most variable genes across acute myeloid leukemia (AML) (n = 145), B-acute lymphoblastic leukemia (ALL) (n = 223), T-ALL (n = 85), and ALAL (n = 30), with each point corresponding to a single patient. RNA sequencing (RNA-seq) data were normalized as transcripts per million (TPM) and log-transformed. **(B)** Diagram depicting the training of the Expression-driven Classification of Acute Leukemias (E-CAL) model. As shown here, we extracted the top 8000 most variable genes from log-transformed, TPM-normalized RNA-seq data of AML, B-ALL, and T-ALL. Then, we trained a multinomial logistic regression model with lasso regularization, using nested cross-validation for hyperparameter estimation and performance evaluation. **(C)** Heatmap displaying TPM-normalized gene expression data, as Z-scores, of the genes with nonzero coefficients in the classifier. Hierarchical clustering, indicated by the dendrogram on the edges, was performed on Euclidean distance for genes (rows) and Manhattan distance for samples (columns). **(D)** Spider plot depicting the proportion of mutations across five distinct molecular categories in ALAL cases classified as AML (ALAL-M), T-ALL (ALAL-T), or B-ALL (ALAL-B).

Lineage-defined ALALs show signatures of differentiated hematopoietic cells

The close similarity between the GEP of 27 ALAL cases and lineage-defined leukemias suggests that they derive from a committed hematopoietic cell of origin. To investigate this hypothesis, we used CIBERSORTx[47] to deconvolute the bulk ALAL RNA-seq data with publicly available single-cell transcriptional profiles of normal hematopoietic cells.[48,49] This analysis revealed strong enrichment for signatures of cell H types corresponding to the leukemic type predicted by E-CAL: myeloid for ALAL-M, B-lymphoid for ALAL-B, and lymphoid-biased multipotent progenitor (LMPP) and T-lymphoid for ALAL-T (Figure 4A, Supporting Information S1: Figure 6A, and Supporting Information S2: Table 14). However, some cases, including several classified as ALAL-T, also displayed high hematopoietic stem cell (HSC) and multipotent progenitor (MPP) scores, suggesting that they may have arisen from earlier progenitors, consistent with previous studies.[6,50] The three patients who could not be assigned to a single lineage by E-CAL showed either a mixed signature from distinct lineages (#23129, #22679) or hematopoietic stem and progenitor cell (HSPC) transcriptional programs (#22668). Patient #22673, also enriched for HSC/MPP signatures, was classified as AML, but only by a narrow margin, since it showed high probability scores for both B-ALL and T-ALL (Supporting Information S2: Table 13). Deconvolution with BayesPrism[51] largely confirmed these results, but with increased proportions of lymphoid progenitors for ALAL-T cases, and HSC/MPP for ALAL-M cases (Supporting Information S1: Figure 6B, Supporting Information S2: Table 14).

Moreover, the transcriptional levels of lineage-specific surface markers and master regulators in ALAL cases were consistent with the leukemic class assigned by the classifier (Figure 4B, Supporting Information S1: Figure 6C). Their expression patterns were also generally concordant with the immunophenotypic data (Supporting Information S1: Figure 6C), although the latter were more frequently in support of a mixed lineage, whereas transcriptional levels aligned more closely with the lineage assigned by our classifier. Interestingly, cases with intermediate E-CAL probabilities also tended to show co-expression of cross-lineage markers at both transcriptional and antigenic levels, such as TdT, CD117, and CD20 in #22668. Furthermore, gene set enrichment analysis (GSEA) detected enrichment for myeloid, B-lymphoid, and T-lymphoid gene sets in ALAL-M, ALAL-B, and ALAL-T, respectively, when compared to the remaining samples (Figure 4C).

In sum, these results provide orthogonal confirmation for the lineage predicted by E-CAL and support the notion that ALALs frequently originate from a lineage-committed cell.

E-CAL successfully classifies ALAL cases in other cohorts

We validated the applicability of E-CAL for ALAL lineage assignment using an RNA-seq data set from the MD Anderson Cancer Center (MDACC, $n = 24$),[14] processed in the same way as our in-house data set. In the original publication, the authors classified the patients into AML or ALL based on their methylation profiles. Since our classifier distinguishes between T-ALL and B-ALL, cases with ALL methylation profiles were considered T-ALL or B-ALL depending on whether they were T/M or B/M MPAL according to the WHO classification.

Analogously to our cohort, a PCA revealed that the majority of MDACC ALAL cases also clustered among lineage-defined leukemias (Figure 4D). Indeed, E-CAL unequivocally classified 75% of them into AML, T-ALL, or B-ALL, in a manner fully consistent with the labels derived from their immunophenotypic and methylation profiles (Figure 4E, Supporting Information S2: Table 15). Notably, the other 25% ALAL cases were assigned to the same two lineages inferred from their immunophenotype using WHO guidelines, namely, 5 T/M and 1 B/M. The higher percentage of multilineage cases may be due to the fact that the MDACC consisted of patients exclusively defined on the basis of the WHO classification, whereas the EMC cohort also included several leukemias that only met the EGIL criteria. In line with this hypothesis, several MDACC cases carried mutations typically associated with opposing lineages (Figure 4E). We further tested our classifier on a second cohort of 89 ALAL cases with RNA-seq and mutational data available, which the authors had previously analyzed with hierarchical clustering to assess transcriptional similarity with other acute leukemias.[8] In this cohort, our classifier assigned a lineage compatible with both the mutational build-up and immunophenotype in 84% of the cases, whereas it agreed with the hierarchical clustering in 66% of them (Supporting Information S2: Table 16). Several of the discrepancies could be explained by the fact that hierarchical clustering enforces a single cluster, whereas our classifier allows for multi-lineage labels.

These observations validate the generalizability of our classifier to external data sets and confirm that a majority of ALAL patients can be more precisely assigned to lineage-committed leukemias independently of the sequencing strategy.

Adult ALAL patients often carry mutations targetable by Food and Drug Administration-approved drugs

Since ALAL patients have a poor prognosis with the current treatment regimes, they could benefit from therapies directed at genes mutated in these patients. Comparison of the list of genes affected by genetic lesions with data of the Therapeutic Target Database[52] revealed that 19.5% of all detected mutations in this cohort are potentially targetable by Food and Drug Administration (FDA)-approved drugs (Figure 5A, Supporting Information S2: Table 17). Among these, six cases with *FLT3* mutations could be treated with *FLT3* inhibitors, which are currently the standard of care for AML with *FLT3* alterations,[53] whereas the six patients with *BCR::ABL1* should respond to specific TKIs.[54] This approach also offered insights into the potential repurposing of drugs not currently indicated for leukemia. For instance, ROS1-directed TKIs like Repotrectinib, approved for lung cancer,[55] could be a promising option for patient #22671, who showed a *PJA2::ROS1* fusion. In fact, a recent study reported successful use of such inhibitors to treat an AML patient with a *TFG::ROS1* fusion who had been refractory to other lines of therapy.[56] In addition, 19.5% of the other mutated genes may be susceptible to drugs that have been previously tested in clinical trials, but are not yet approved.

Overall, 70% of the studied ALAL cases carried mutations potentially targetable by FDA-approved drugs, whereas the remaining 30% could be targeted by drugs that have reached the clinical phase (Figure 5B). In all cases, patients harbored at least one clonal mutation amenable to treatment, with four patients showing additional targetable subclonal mutations (defined as VAF < 5%). However, in two cases, the only approved drugs available were directed against subclonal lesions. While therapeutic efficacy generally requires mutations to be present in the majority of tumor cells, targeting subclonal lesions may be clinically relevant, as therapy resistance is often mediated by these subpopulations.[57–59] Accordingly, it has been shown that the addition of FLT3 inhibitors to azacitidine and venetoclax suppresses the emergence of resistant leukemic subclones.[60]

All things considered, comprehensive profiling of the genetic lesions in ALAL patients by RNA-seq enables the identification of mutated genes potentially amenable to targeted therapies in addition to AML/ALL-based treatment backbones.

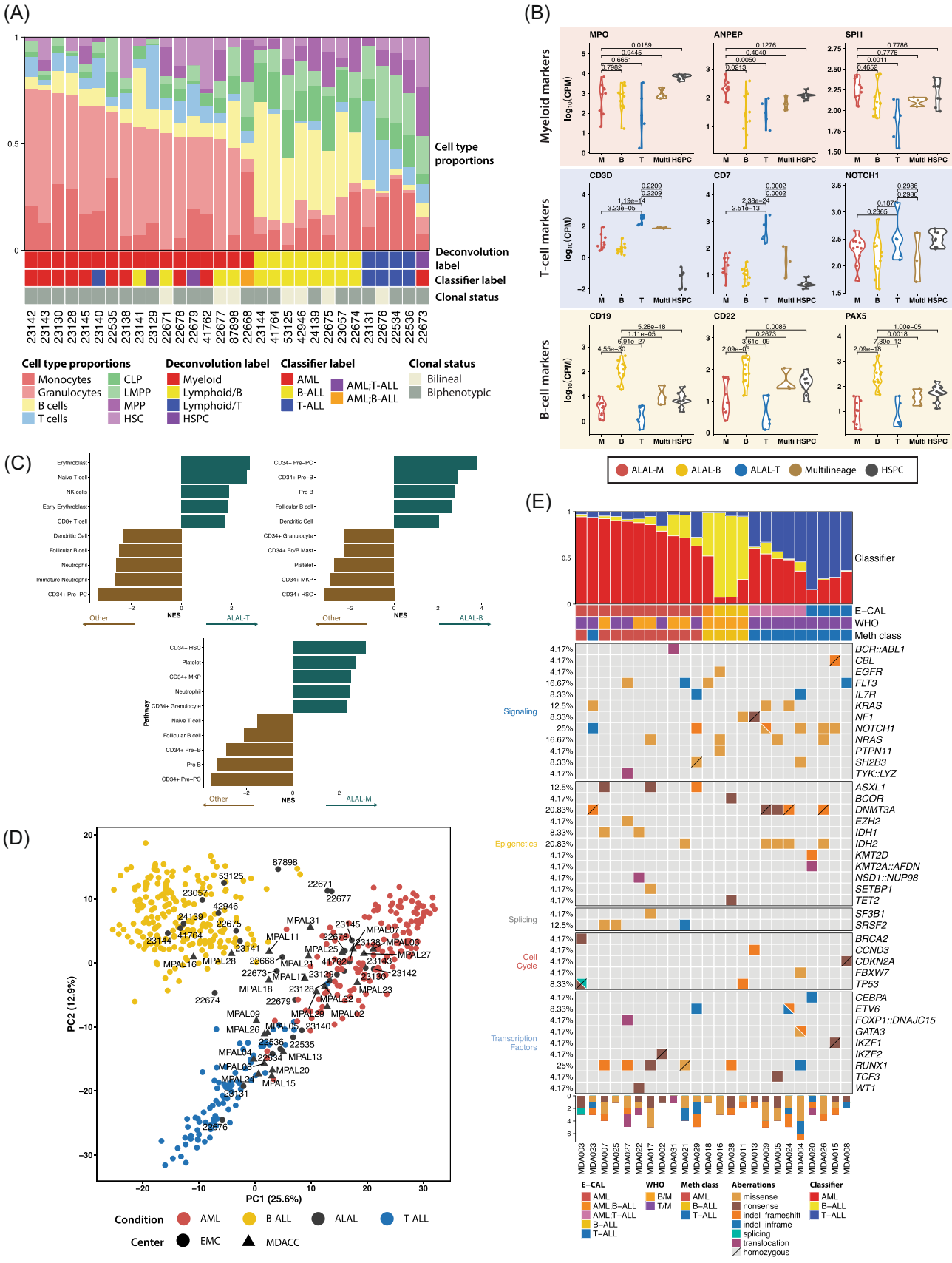


FIGURE 4 (See caption on next page).

FIGURE 4 **Orthogonal validation of Expression-driven Classification of Acute Leukemias (E-CAL) for acute leukemia of ambiguous lineage (ALAL) classification.** **(A)** Bar plot indicating the proportion of cell types estimated by CIBERSORTx in bulk RNA-sequencing data of ALAL cases, using single-cell transcriptomics data from the Human Cell Atlas[48] as a reference matrix. These scores were discretized by adding up all cell types within the same compartment (myeloid, lymphoid/B, lymphoid/T, and hematopoietic stem and progenitor cell [HSPC]), counting lymphoid-biased progenitors CLP and LMPP for both B-lymphoid and T-lymphoid populations; a unique label was assigned to the category with the highest score. These labels are depicted at the bottom, alongside the leukemic class assigned by E-CAL. **(B)** Expression of surface markers and transcription factors associated with the myeloid, T-lymphoid, or B-lymphoid lineages. Data are presented as log10-transformed counts per million (CPM). **(C)** Bar plot showing the top five results from a pre-ranked gene set enrichment analysis (GSEA) conducted on the Hay et al. data sets from the C8 MSigDB collection. The analysis was conducted on differentially expressed genes in ALAL-M (left), ALAL-T (middle), and ALAL-B (right) relative to the remaining ALAL cases in the same cohort. **(D)** Principal component analysis (PCA) of the 1000 most variable genes across acute myeloid leukemia (AML) ($n = 145$), B-acute lymphoblastic leukemia (ALL) ($n = 223$), T-ALL ($n = 85$), Erasmus University Medical Center (EMC)-ALAL ($n = 30$), and MD Anderson Cancer Center (MDACC)-ALAL ($n = 24$), with each patient corresponding to a single point. EMC patients are depicted as circles, and MDACC cases, used for external validation, are depicted as triangles. HSC, hematopoietic stem cell; MPP, multipotent progenitor; CLP, common lymphocyte progenitor; LMPP, lymphoid-biased multipotent progenitor; WHO, World Health Organization.

DISCUSSION

The existence of leukemia patients showing both myeloid and lymphoid features has puzzled researchers for more than a century.[61,62] In the last decades, various attempts have been made to establish clear-cut criteria to classify these patients, but they largely rely on a restricted set of surface markers that fail to capture the complex molecular processes that underpin cellular identity. In this study, we used RNA-seq data to simultaneously characterize the genetic landscape and assign the lineage of ALAL cases, thereby providing more accurate grounds for classification. We demonstrated that RNA-seq can serve as a comprehensive diagnostic tool, potentially replacing other molecular biology techniques currently in use. Moreover, our machine learning model classified the majority of ALAL patients into lineage-restricted leukemias on the basis of their GEP, which could have important implications for diagnosis and treatment.

The identification of biphenotypic and bilineal leukemias relies on the detection of surface markers associated with opposing hematopoietic lineages. However, it is unclear whether they are a veritable clinical entity or merely an artifact caused by the misexpression of surface antigens in an otherwise lineage-committed cell. Our findings indicate that the latter is frequently true: 90% of the ALAL patients in our cohort were classified into AML, T-ALL, or B-ALL and mostly carried genetic lesions associated with the assigned lineage. Deconvolution with data from healthy donors further corroborated the validity of the predictions. A few patients could not be reliably assigned to any single leukemic class by the classifier, suggesting that they originate from early progenitors without a well-defined cellular identity. Although no AUL cases were included in our cohort, their lack of lineage-associated surface markers may indicate that they similarly arise from primitive HSPCs. Interestingly, however, Lao and colleagues showed that many AULs closely resemble secondary AML in terms of genetic make-up and gene expression.[50]

Recently, single-cell studies have identified stem-like transcriptional signatures strongly enriched in ALAL,[63,64] supporting the notion that the cell of origin is an early progenitor.[44] We confirmed these findings in our adult cohort using the signatures from Peretz et al.[64] (Supporting Information S1: Figure 7A), while our deconvolution analyses revealed substantial enrichment for HSC/MPPs or LMPPs. Nevertheless, we and others have also shown that ALALs often have transcriptional, epigenetic, and mutational profiles that closely resemble those of lineage-defined leukemias.[8,11,14,50,65] Mapping of single-cell ALAL data onto a reference atlas of hematopoiesis also revealed frequent overlap between the differentiation landscapes of ALAL and AML.[66] These observations point to a model in which ALALs originate from primitive progenitors positioned at different levels of the hematopoietic hierarchy, yielding two broad

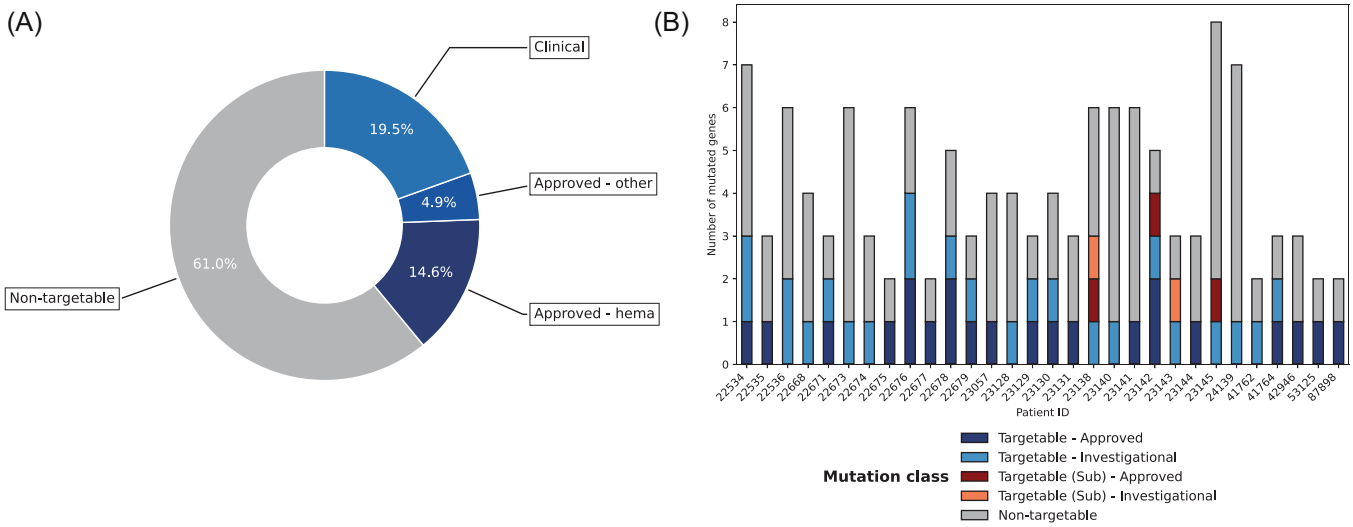


FIGURE 5 **Comprehensive profiling of acute leukemia of ambiguous lineage identifies possible targeted therapies.** **(A)** Pie chart depicting the proportion of gene mutations detected in this cohort that are targetable by Food and Drug Administration-approved or investigational drugs. **(B)** Bar plot showing the proportion of mutated genes in each patient that can be targeted by currently available targeted therapies. Clonal and subclonal mutations are shown in blue and red hues, respectively.

subgroups (Supporting Information S1: Figure 7B). In most cases, ALAL blasts originate from committed cells that misexpress surface markers typically associated with an alternative lineage, thereby conforming to the lineage infidelity hypothesis postulated by McCulloch.[20] We propose that these are, in fact, cases of ALL or AML masquerading as ALAL. In contrast, in a smaller group of patients, the leukemic cell arises from an earlier progenitor that retains lineage promiscuity upon the acquisition of driver mutations.[21] Thus, these cases may be considered "true" ALALs with a transcriptional program that straddles different branches of the hematopoietic tree. Specific aberrations, like *KMT2A*[6] or *BCL11B*[11] rearrangements, have been shown to drive lineage ambiguity when acquired in these early progenitors.

Although our conclusions are broadly in line with other studies, they are limited by the small sample size ($n = 30$) in our study, which may preclude the identification of ALAL-specific clusters. Accounting for 1%–2% of all leukemias, ALAL is a rare clinical entity, thereby hindering the acquisition of samples at a large scale. To the best of our knowledge, the largest RNA-seq ALAL data sets available are those from Montefiori et al. ($n = 126$, including mostly pediatric leukemias),[11] Wang et al. ($n = 89$, adults only),[8] and Takahashi et al.[14] Since our study focused on adults, we used the latter two to evaluate the frequency of mutations and validate our classifier. Another potential shortcoming is the use of bulk RNA-seq data instead of single-cell RNA-seq, but we chose the former because it is more suitable for the characterization of large cohorts and the detection of mutations.

Differential diagnosis of leukemia patients would benefit from a classification system based on genetic and transcriptional profiles rather than a reduced set of surface markers. Our study shows that RNA-seq, when analyzed with tailored computational pipelines, could be the technique of choice for this purpose. In our hands, it demonstrated high sensitivity for point mutations, small indels, fusion genes, and other large cytogenetic aberrations. Insufficient coverage did not seem to preclude the identification of any relevant lesions, as it was mainly restricted to a few cell-type-specific genes typically silent in the leukemic lineage assigned to the corresponding ALAL cases, such as *PAX5* in T cells and myeloid cells or *KIT* and *WT1* in B cells. Nonetheless, RNA-seq failed to detect smaller CNAs that only affect a single gene (such as *CDKN2A* or *IKZF1*), as it cannot distinguish them from local changes in expression. Likewise, RNA-seq is blind to events that do not involve any coding regions, such as various forms of enhancer dysregulation.[67] Notably, a subset of ALALs overexpresses *BCL11B* as a result of either amplification of its native enhancer or hijacking of rearranged enhancers.[11,12] To overcome these drawbacks, RNA-seq could be coupled with low-coverage whole-genome sequencing to reveal all CNAs and structural variants.[68]

Simple diagnostic algorithms based on a few immunophenotypic markers are easy to implement and understand, and therefore likely to remain in use for the foreseeable future. It is thus important to evaluate their agreement with our transcription-driven strategy. The E-CAL lineage assignment was concordant with the WHO classification for eight patients (five AML and three T-ALL) and with the EGIL scoring system for two others (one B-ALL and one T-ALL), none of which overlapped with the previous eight (Supporting Information S2: Table 4, Supporting Information S1: Figure 7C). Taken together, these findings lend greater support to the WHO scoring system over EGIL. However, both misclassified the majority of B-ALL cases as B/M MPAL, underscoring the need to reconsider the current diagnostic criteria for this clinical entity.

An important consideration in machine learning is the ability of the model to make predictions on unseen data. When dealing with RNA-seq, this can be potentially compromised by the existence of multiple experimental strategies used to enrich the sequencing library.[69] To partially address these systematic differences, we only quantified expression of protein-coding genes and log-transformed them to stabilize the variance.[70] We avoided normalization techniques that would require the original data set, such as the trimmed-mean of the *M*-values, to facilitate the application of E-CAL to new data by other researchers. Ultimately, the robustness of our classifier was proven by the successful classification of most ALAL cases from independent cohorts.[8,14] In the MDACC dataset, however, E-CAL did not assign a single label to 25% of ALAL cases, whereas methylation-based hierarchical clustering did. While this discrepancy could stem from the particularities of the RNA-seq protocol, other possible causes are the differences between methylation and transcription or between the computational strategies for classification used in each study.

The improved subclassification proposed here could have implications for diagnosis and therapeutic decisions. At the time of diagnosis, clinicians relied on expert opinion to determine whether patients should receive ALL-like or AML-like therapy. In six cases, retrospective classification using E-CAL did not align with the originally administered treatment: two patients retrospectively classified as AML by E-CAL had received ALL-type therapy, whereas four patients classified as ALL had been treated with AML-based regimens. Although the sample size of our cohort is insufficient for robust statistical analysis of survival outcomes, we observed a notable difference in CR rates between treatments that matched the E-CAL classification versus those that did not match the E-CAL classification. Patients who received therapy consistent with their E-CAL classification achieved CR in 16 of 20 cases (80%), whereas nonmatching treatments led to CR in only 1 of 6 cases (17%). In two cases where CR was not initially attained, modifying the treatment approach resulted in CR. Nevertheless, prospective research is needed to ascertain whether E-CAL-guided treatment leads to improved survival of ALAL patients.

Moreover, the detection of gene lesions by our computational pipeline opens the door to personalized treatment. Contrary to traditional molecular approaches or small DNA-seq panels, RNA-seq provides a comprehensive survey of the protein-coding mutational landscape. Using a public target-drug database, we have shown that all patients in our cohort harbor at least one mutation in a gene targetable by existing compounds, the majority of which are FDA-approved. Even though some of these drugs are not part of the standard of care in leukemia, our findings provide a mechanistic rationale for their repurposing, which is further warranted by the poor prognosis of ALAL. Given the rarity of this disease, large multi-center studies with centralized diagnostic procedures based on RNA-seq may be necessary to investigate the benefit of this approach.[71,72]

In conclusion, our work showcases the potential of RNA-seq to improve the diagnostics and, potentially, the treatment of ALAL patients, providing a comprehensive characterization of their genetic and transcriptional profiles. The E-CAL classifier developed in our study improves the lineage assignment of ALAL cases, which are often misclassified due to the misexpression of a few cross-lineage markers. These findings shed light on the biological mechanisms that underpin these poorly understood leukemias and argue for changes in the way they are currently classified and treated.

ACKNOWLEDGMENTS

We would like to thank our colleagues from the Bone Marrow Transplantation Group and the Molecular Diagnostics Laboratory of the Department of Hematology as well as collaborators from the Department of Clinical Genetics at the Erasmus University Medical Center for storage of samples, and molecular and cytogenetic analyses of the leukemia cells. Furthermore, we thank Zorica Ristic for performing experiments and Priscilla van Hilst for collecting clinical data. We also thank Petri Pölönen, from the St. Jude Children's Research Hospital, for

sharing patient sample information; Koichi Takahashi, from the MD Anderson Cancer Center, for sharing patient metadata related to the MDACC ALAL RNA-seq data set; Qian Wang and Hai-Ping Dai, from the Jiangsu Institute of Hematology, for sharing patient metadata and processed gene expression data related to the JIH ALAL RNA-seq data set; and Wencke Walter, Alessandro Baldi, and Torsten Haferlach, from the Munich Leukemia Laboratory, for sharing acute leukemia RNA-seq data used to externally validate our classifier.

AUTHOR CONTRIBUTIONS

Roger Mulet-Lazaro: Conceptualization; data curation; writing—review and editing; writing—original draft; visualization; formal analysis; validation. **Anikó Sijs-Szabó**: Conceptualization; investigation; writing—original draft; writing—review and editing; data curation; funding acquisition. **Remco M. Hoogenboezem**: Software; formal analysis; methodology. **Stanley van Herk**: Investigation. **Carla Exalto**: Investigation. **Jasper E. Koenders**: Investigation. **Patricia G. Hoogeveen**: Investigation. **François G. Kavelaars**: Investigation; formal analysis. **Anita M. Schelen**: Investigation. **Willemijn van den Ancker**: Resources. **Arjan A. van de Loosdrecht**: Resources. **Charles G. Mullighan**: Resources; writing—review and editing. **H. Berna Beverloo**: Formal analysis. **Vincent van der Velden**: Formal analysis. **Jan J. Cornelissen**: Supervision; resources. **Peter J. M. Valk**: Supervision; resources. **Anita W. Rijneveld**: Conceptualization; supervision; resources; writing—review and editing. **Mathijs A. Sanders**: Conceptualization; writing—review and editing; methodology; supervision; formal analysis.

CODE AVAILABILITY STATEMENT

Software tools used in this study are mentioned in the corresponding sections of the Methods. R code used in the analysis of the data presented here can be found on GitLab (https://gitlab.com/erasmusmc-hematology/alal-classification). This repository also contains an R object with the E-CAL model trained on acute leukemia data, as well as instructions for its use. The model is also available via a user-friendly application: https://bmbrowser.shinyapps.io/leukemia_classifier/.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The raw RNA-seq data of ALAL patients generated in this study are deposited at the European Genome-Phenome Archive (EGA, https://ega-archive.org/studies/) under accession number EGAS00001007967. These data are protected by data privacy laws and are only available under restricted access. Processed data are publicly available in ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress/) with identifier E-MTAB-15 145. The raw RNA-seq data of AML[73] and T-ALL[18] patients have been previously used in other studies and are available under accession numbers EGAD00001007581 and EGAD00001011054, respectively. The external RNA-seq data utilized for validation, provided by the Munich Leukemia Laboratory, are available upon request through the Torsten Haferlach Foundation. For access requests, please visit https://torsten-haferlach-leukaemiediagnostik-stiftung.de/en/.

ETHICS STATEMENT

This study was approved by the ethics committee of the EMC (MEC-2015-155) and was conducted in accordance with the Declaration of Helsinki.

FUNDING

This work was funded by KWF Dutch Cancer Society project nr7512 (Anita W. Rijneveld) and the Landsteiner Foundation for Blood Transfusion Research (LSBR) fellowship LSBR 2330F (Mathijs A. Sanders).

SUPPORTING INFORMATION

Additional supporting information can be found in the online version of this article.

REFERENCES

1. Weinberg OK, Arber DA. Mixed-phenotype acute leukemia: historical overview and a new definition. *Leukemia*. 2010;24(11):1844-1851.
2. Yan L, Ping N, Zhu M, et al. Clinical, immunophenotypic, cytogenetic, and molecular genetic features in 117 adult patients with mixed-phenotype acute leukemia defined by WHO-2008 classification. *Haematologica*. 2012;97(11):1708-1712.
3. Catovsky D, Matutes E, Buccheri V, et al. A classification of acute leukaemia for the 1990s. *Ann Hematol*. 1991;62(1):16-21.
4. Bene MC, Castoldi G, Knapp W, et al. Proposals for the immunological classification of acute leukemias. European Group for the Immunological Characterization of Leukemias (EGIL). *Leukemia*. 1995;9(10):1783-1786.
5. Jaffe ES, Organization WH. *Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues*. IARC Press; 2001.
6. Alexander TB, Gu Z, Iacobucci I, et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature*. 2018;562(7727):373-379.
7. Orgel E, Alexander TB, Wood BL, et al. Mixed-phenotype acute leukemia: a cohort and consensus research strategy from the Children's Oncology Group Acute Leukemia of Ambiguous Lineage Task Force. *Cancer*. 2020;126(3):593-601.
8. Wang Q, Cai W, Wang Q, et al. Integrative genomic and transcriptomic profiling reveals distinct molecular subsets in adult mixed phenotype acute leukemia. *Am J Hematol*. 2023;98(1):66-78.
9. Matutes E, Pickl WF, van't Veer M, et al. Mixed-phenotype acute leukemia: clinical and laboratory features and outcome in 100 patients defined according to the WHO 2008 classification. *Blood*. 2011;117(11):3163-3171.
10. Victoria AV, Srinivas KT. Mixed phenotype acute leukemias: real-world outcomes by WHO classifications 2010–2021. In: *Blood*. American Society of Hematology; 2024.
11. 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.
12. Di Giacomo D, La Starza R, Gorello P, et al. 14q32 rearrangements deregulating BCL11B mark a distinct subgroup of T-lymphoid and myeloid immature acute leukemia. *Blood*. 2021;138(9):773-784.
13. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*. 2022;36(7):1703-1719.
14. Takahashi K, Wang F, Morita K, et al. Integrative genomic analysis of adult mixed phenotype acute leukemia delineates lineage associated molecular subtypes. *Nat Commun*. 2018;9(1):2670.
15. 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.
16. Hirabayashi S, Ohki K, Nakabayashi K, et al. ZNF384-related fusion genes define a subgroup of childhood B-cell precursor acute lymphoblastic leukemia with a characteristic immunotype. *Haematologica*. 2017;102(1):118-129.
17. Roberts KG, Morin RD, Zhang J, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell*. 2012;22(2):153-166.
18. Mulet-Lazaro R, van Herk S, Nuetzel M, et al. Epigenetic alterations affecting hematopoietic regulatory networks as drivers of mixed myeloid/lymphoid leukemia. *Nat Commun*. 2024;15(1):5693.

19. Wolach O, Stone RM. How I treat mixed-phenotype acute leukemia. *Blood*. 2015;125(16):2477-2485.
20. McCulloch E. Stem cells in normal and leukemic hemopoiesis (Henry Stratton Lecture, 1982). *Blood*. 1983;62(1):1-13.
21. Greaves M, Chan L, Furley A, Watt S, Molgaard H. Lineage promiscuity in hemopoietic differentiation and leukemia. *Blood*. 1986;67(1):1-11.
22. Hu M, Krause D, Greaves M, et al. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev*. 1997;11(6):774-785.
23. Shi R, Munker R. Survival of patients with mixed phenotype acute leukemias: a large population-based study. *Leuk Res*. 2015;39(6):606-616.
24. Aggarwal N, Weinberg OK. Update on acute leukemias of ambiguous lineage. *Clin Lab Med*. 2021;41(3):453-466.
25. Kurzer JH, Weinberg OK. Acute leukemias of ambiguous lineage. *Surg Pathol Clin*. 2019;12(3):687-697.
26. Wolach O, Stone RM. Mixed-phenotype acute leukemia: current challenges in diagnosis and therapy. *Curr Opin Hematol*. 2017;24(2):139-145.
27. 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.
28. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
29. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43(1110):11.10.1-11.10.33.
30. 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.
31. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-2871.
32. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher—a tool f or finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014. 011650.
33. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*. 2019;20(1):213.
34. Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res*. 2021;31(3):448-460.
35. Flensburg C, Oshlack A, Majewski IJ. Detecting copy number alterations in RNA-Seq using SuperFreq. *Bioinformatics*. 2021;37(22):4023-4032.
36. Frankish A, Carbonell-Sala S, Diekhans M, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res*. 2023;51(D1):D942-D949.
37. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer. *Cancer Res*. 2017;77(21):e31-e34.
38. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.
39. Bassan R, Hoelzer D. Modern therapy of acute lymphoblastic leukemia. *J Clin Oncol*. 2011;29(5):532-543.
40. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet*. 2013;45(3):242-252.
41. Mühlbacher V, Zenger M, Schnittger S, et al. Acute lymphoblastic leukemia with low hypodiploid/near triploid karyotype is a specific clinical entity and exhibits a very high TP53 mutation frequency of 93. *Genes Chromosom Cancer*. 2014;53(6):524-536.
42. Stark B, Jeison M, Gobuzov R, et al. Near haploid childhood acute lymphoblastic leukemia masked by hyperdiploid line. *Cancer Genet Cytogenet*. 2001;128(2):108-113.
43. Carroll AJ, Shago M, Mikhail FM, et al. Masked hypodiploidy: hypodiploid acute lymphoblastic leukemia (ALL) mimicking hyperdiploid ALL in children: a report from the Children's Oncology Group. *Cancer Genet*. 2019;238:62-68.
44. Alexander TB, Orgel E. Mixed phenotype acute leukemia: current approaches to diagnosis and treatment. *Curr Oncol Rep*. 2021;23(2):22.
45. Xiao W, Bharadwaj M, Levine M, et al. PHF6 and DNMT3A mutations are enriched in distinct subgroups of mixed phenotype acute leukemia with T-lineage differentiation. *Blood Adv*. 2018;2(23):3526-3539.
46. Boeckx N, Willemse M, Szczepanski T, et al. Fusion gene transcripts and Ig/TCR gene rearrangements are complementary but infrequent targets for PCR-based detection of minimal residual disease in acute myeloid leukemia. *Leukemia*. 2002;16(3):368-375.
47. 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.
48. Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp Hematol*. 2018;68:51-61.
49. Xie X, Liu M, Zhang Y, et al. Single-cell transcriptomic landscape of human blood cells. *Natl Sci Rev*. 2021;8(3):nwaa180.
50. Lao ZT, Ding LW, An O, et al. Mutational and transcriptomic profiling of acute leukemia of ambiguous lineage reveals obscure but clinically important lineage bias. *Haematologica*. 2019;104(5):e200-e203.
51. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer*. 2022;3(4):505-517.
52. Zhou Y, Zhang Y, Zhao D, et al. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res*. 2024;52(D1):D1465-D1477.
53. Zhao JC, Agarwal S, Ahmad H, Amin K, Bewersdorf JP, Zeidan AM. A review of FLT3 inhibitors in acute myeloid leukemia. *Blood Rev*. 2022;52:100905.
54. Leak S, Horne GA, Copland M. Targeting BCR-ABL1-positive leukaemias: a review article. *Camb Prism Precis Med*. 2023;1:e21.
55. Rais T, Shakeel A, Naseem L, Nasser N, Aamir M. Repotrectinib: a promising new therapy for advanced nonsmall cell lung cancer. *Ann Med Surg*. 2024;86(12):7265-7269.
56. Sun Jie, Zhang Shaoqi, Du Ran, et al. The TFG-ROS1 fusion is an oncogenic driver of human myeloid leukemia. *Blood*. 2024;144:38.
57. McMahon CM, Ferng T, Canaani J, et al. Clonal selection with RAS pathway activation mediates secondary clinical resistance to selective FLT3 inhibition in acute myeloid leukemia. *Cancer Discov*. 2019;9(8):1050-1063.
58. Sango J, Carcamo S, Sirenko M, et al. RAS-mutant leukaemia stem cells drive clinical resistance to venetoclax. *Nature*. 2024;636(8041):241-250.
59. DiNardo CD, Tiong IS, Quaglieri A, et al. Molecular patterns of response and treatment failure after frontline venetoclax combinations in older patients with AML. *Blood*. 2020;135(11):791-803.
60. Short NJ, Nguyen D, Ravandi F. Treatment of older adults with FLT3-mutated AML: emerging paradigms and the role of frontline FLT3 inhibitors. *Blood Cancer J*. 2023;13(1):142.
61. Findlay L. Case of mixed-celled leukæmia. *Glasgow Med J*. 1906;66(4):264-271.
62. Hull MT, Griep JA. Mixed leukemia, lymphatic and myelomonocytic. *Am J Clin Path*. 1980;74(4):473-475.
63. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. 2019;37(12):1458-1465.

64. Peretz CAC, Kennedy VE, Walia A, et al. Multiomic single cell sequencing identifies stemlike nature of mixed phenotype acute leukemia. *Nat Commun*. 2024;15(1):8191.
65. Zheng R, Gagan JR, Botten GA, et al. Genomic landscape of mixed phenotype acute leukemia associated with immunophenotypic lineage predominance: impact on diagnosis and treatment. *Eur J Haematol*. 2025;114(6):1041-1051.
66. Zeng AGX, Iacobucci I, Shah S, et al Single-cell transcriptional atlas of human hematopoiesis reveals genetic and hierarchy-based determinants of aberrant AML differentiation. *Blood Cancer Discov*. 2025;6(4):307-324.
67. Mulet-Lazaro R, Delwel R. Oncogenic enhancers in leukemia. *Blood Cancer Discov*. 2024;5(5):303-317.
68. Yang L. A practical guide for structural variation detection in the human genome. *Curr Protoc Hum Genet*. 2020;107(1):e103.
69. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*. 2020;26(8):903-909.
70. Booeshaghi AS, Pachter L. Normalization of single-cell RNA-seq counts by log(x + 1) or log(1 + x). *Bioinformatics*. 2021;37(15):2223-2224.
71. Burd A, Levine RL, Ruppert AS, et al. Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial. *Nat Med*. 2020;26(12):1852-1858.
72. Döhner H, Wei AH, Löwenberg B. Towards precision medicine for AML. *Nat Rev Clin Oncol*. 2021;18(9):577-590.
73. Mulet-Lazaro R, van Herk S, Erpelinck C, et al. Allele-specific expression of GATA2 due to epigenetic dysregulation in CEBPA double-mutant AML. *Blood*. 2021;138(2):160-177. doi:10.1182/blood.2020009244