Received: 7 June 2022 | Revised: 21 June 2022 | Accepted: 21 June 2022

DOI: 10.1002/ajh.26647

RESEARCH ARTICLE

HERVs characterize normal and leukemia stem cells and represent a source of shared epitopes for cancer immunotherapy

Vincent Alcazer[1,2] | Paola Bonaventura[2,3] | Laurie Tonon[4] | Emilie Michel[5] | Virginie Mutez[5] | Clémentine Fabres[2,3] | Nicolas Chuvin[5] | Rasha Boulos[5] | Yann Estornes[5] | Véronique Maguer-Satta[2] | Kevin Geistlich[2] | Alain Viari[4] | Klaus H. Metzeler[6,7] | Wolfgang Hiddemann[5] | Aarif M. N. Batch[8,9] | Tobias Herold[6] | Christophe Caux[2,3] | Stéphane Depil[2,3,5,10]

[1]Department of Hematology, Hospices Civils de Lyon, Lyon Sud Hospital, Pierre-Bénite, France

[2]Cancer Research Center of Lyon, INSERM U1052 and CNRS UMR5286, Lyon, France

[3]Centre Léon Bérard, Lyon, France

[4]Synergie Lyon Cancer Foundation, Gilles Thomas Bioinformatics Center, Centre Léon Bérard, Lyon, France

[5]Ervaccine Technologies, Centre Leon Bérard, Lyon, France

[6]Laboratory for Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Munich, Germany

[7]Department of Hematology and Cell Therapy, University of Leipzig, Leipzig, Germany

[8]Institute of Medical Data Processing, Biometrics and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Munich, Germany

[9]Data Integration for Future Medicine (DiFuture, www.difuture.de), LMU Munich, Munich, Germany

[10]University Claude Bernard Lyon 1, Lyon, France

Correspondence
Vincent Alcazer, Department of Hematology, Hospices Civils de Lyon, Lyon Sud Hospital, Pierre-Bénite, France.
Email: vincent.alcazer@chu-lyon.fr

Stéphane Depil, Centre Leon Bérard, Lyon, France.
Email: stephane.depil@lyon.unicancer.fr

Funding information
This work was supported by the INSERM "poste d'accueil" research grant (to V. Alcazer), the Agence de Biomédecine, Foundation ARC pour la Recherche sur le Cancer and Ligue contre le Cancer. This work is supported by a grant of the Wilhelm-Sander-Stiftung

Abstract

Human endogenous retroviruses (HERVs) represent 8% of the human genome. The expression of HERVs and their immune impact have not been extensively studied in Acute Myeloid Leukemia (AML). In this study, we used a reference of 14 968 HERV functional units to provide a thorough analysis of HERV expression in normal and AML bone marrow cells. We show that the HERV retrotranscriptome accurately characterizes normal and leukemic cell subpopulations, including leukemia stem cells, in line with different epigenetic profiles. We then show that HERV expression delineates AML subtypes with different prognoses. We finally propose a method to select and prioritize $CD8^+$ T cell epitopes derived from AML-specific HERVs and we show that lymphocytes infiltrating patient bone marrow at diagnosis contain naturally

Abbreviations: AHR, Active HERVs Regions; AML, Acute Myeloid Leukemia; AZA, Azacytidine; BIC, Bayesian Information Criterion; BMMCs, Bone marrow mononuclear cells; CCLE, Cancer Cell Line Encyclopedia; CNV, Copy-number variation; GDC, Genomic Data Commons; GEO, Gene Expression Omnibus; GMP, granulocyte-monocyte progenitor; GTEX, Genotype-Tissue Expression; HERVs, Human endogenous retroviruses; HSC, Hematopoietic stem cell; LMPP, lymphoid-primed multipotent progenitor; LSC, Leukemic stem cells; LTR, Long-terminal repeat; MILs, Marrow-infiltrating lymphocytes; moDCs, Monocyte-derived dendritic cells; MPP, multipotent progenitor; ORFs, Open-reading frames; PBMCs, Peripheral blood mononuclear cells; PCR, polymerase chain reaction; pHSC, pre-leukemic hematopoietic stem cell; ssGSVA, single sample gene-set variation analysis; TSS, Transcription start site; VST, Variance stabilizing transformation.

Vincent Alcazer and Paola Bonaventura equally contributed to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *American Journal of Hematology* published by Wiley Periodicals LLC.

(no. 2013.086.2) to TH and by the BMBF grant 01ZZ1804B (DIFUTURE) to A.M.N.B.

occurring CD8$^+$ T cells against these HERV epitopes. We also provide in vitro data supporting the functionality of HERV-specific CD8$^+$ T-cells against AML cells. These results show that HERVs represent an important source of genetic information that can help enhancing disease stratification or biomarker identification and an important reservoir of alternative tumor-specific T cell epitopes relevant for cancer immunotherapy.

1 | BACKGROUND

Human endogenous retroviruses (HERVs) represent 8% of the human genome.[1] These sequences are remnants of ancestral germline infections by exogenous retroviruses.[2] The original sequence of a HERV is that of an exogenous retrovirus, with two promoter long-terminal repeat (LTR) sequences surrounding the virus open-reading frames (ORFs): *gag*, *pro*, *pol* and *env*.[3] However, after millions of years of evolution, these ORFs have been deeply altered, and there is currently no description of any autonomous fully infectious HERV.[4]

The long-standing belief is that HERVs are repressed by epigenetic mechanisms and are thus not expressed, or only poorly, in normal tissues.[5] However, recent studies have shown that HERV expression can be detected in a vast range of normal tissues.[6] Different pathological conditions can lead to aberrant HERV expression, as it has now been largely described in auto-immune diseases[7–10] and in cancers,[11–14] where HERVs have been the subject of many studies over the last years. Indeed, it was reported that HERVs could participate in oncogenesis by inducing chromosomal instability, promoting aberrant gene expression with their LTR or by impacting the immune system with their RNA and protein products.[15] HERVs could thus play a prominent role in cancer immunity, increasing tumor immunogenicity by promoting (i) an innate immune response triggered by the viral defense pathway induced by their nucleic acid intermediates, and (ii) an adaptive immune response by forming a pool of tumor-associated antigens.[16]

Acute Myeloid Leukemia (AML) is a heterogeneous disease characterized by the clonal expansion of myeloid progenitor and stem cells.[17] While some AML subtypes are characterized by recurrent genetic translocations or mutations associated with particular prognoses, most AMLs present a normal or complex karyotype, and identifying key factors that predict treatment resistance in these patients represents a major challenge.[17–22] Aside from disease stratification, AML also belongs to malignancies with the lowest mutational burdens,[23] and finding tumor-specific antigens for immunotherapeutic approaches remains very difficult as the frequency of mutations creating neoantigens is expected to be low.[24] In this context, HERV-derived antigens could represent a unique source of alternative tumor-specific antigens (i.e. antigens that do not originate from single nucleotide variations in a coding region[25,26]) that could be exploited for the development of new immunotherapies.

To date, little is known about the expression of HERVs in AML and its relevance as either a biomarker or a therapeutic target. Evidence of HERV-K /HML-2 expression in AML cells was shown as early as 1993 and confirmed in the early 2000s.[27,14] Few studies then focused on HERVs in AML until the late 2010s, with the demonstration that azacytidine (Aza) activates the transcription of different HERVs, potentially contributing to its clinical effects.[28] The exact role of HERVs in Aza therapy is however a matter of debate, with recent evidence arguing in favor of a HERV-independent therapeutic effect.[29] More recently, a link was established between HERVs and the expression of surrounding genes in AML, suggesting a regulatory role of these retroelements.[30] However, few data exist on HERV expression and their immune impact in AML, with studies relying on non-exhaustive quantification methods such as polymerase chain reaction (PCR) and focusing only on a few HERV loci[14] or globally quantifying HERVs at the family level together with transposable elements.[31]

In this study, we thoroughly assessed HERV expression in AML and normal blood and bone marrow cells. Using a recent method to exhaustively quantify the HERV retrotranscriptome in next-generation sequencing data, we show that the latter can accurately define normal and leukemic cell populations, including leukemia stem cells (LSCs) that can be characterized by a 25-HERV signature. We also show that leukemic cells present a distinct epigenetic profile compared to their normal cell counterparts, with a significant correlation between the expression of HERVs located in open chromatin regions and surrounding cancer-associated genes. We then show that the HERV retrotranscriptome can be used to discriminate AML profiles from bulk RNA-seq data, distinguishing known but also new AML subtypes of different prognoses. Finally, we show that HERVs specifically expressed in AML cells represent a reservoir of T cell epitopes able to elicit specific immune responses in AML patients.

2 | METHODS

2.1 | Raw RNA-seq data

Raw RNA-seq data files were accessed from the NCBI Gene Expression Omnibus (GEO) portal, under the accession numbers GSE74246 for the sorted hematopoietic normal and AML cells from Corces et al.,[32] GSE49642, GSE52656, GSE62190, GSE66917, GSE67039 and GSE106272 for the LEUCEGENE data sets. TCGA LAML[33] and BEAT-AML[34] data were accessed from the NCI Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/). Raw data

for the AMLCG cohort[18] were directly provided by the AMLCG group. HLA genotyping of TCGA, BEAT, and AMLCG cohorts was assessed using archasHLA v0.2.0.[35] THP-1 cell line data were accessed from the Broad Institute Cancer Cell Line Encyclopedia (CCLE) portal (https://portals.broadinstitute.org/ccle).

2.2 | HERV and gene expression quantification

HERV expression was quantified using a custom pipeline derived from Telescope.[36] Briefly, RNA-seq reads were aligned to a custom transcriptome using bowtie2 v2.2.1[37] with custom parameters to retain multimaps (−k 100 --very-sensitive-local --score-min "L,0,1.6"). The custom transcriptome consisted in the hg38 reference transcriptome with 14 968 HERV transcriptional units compiled from RepeatMasker annotations.[36] SAM outputs were converted to BAM files using SAMtools v1.4.[38] HERV and gene expression were then calculated using Telescope[36] and HTSeq 0.12.3,[39] respectively. Raw counts were concatenated and normalized independently for each data set using DESEQ2 v1.28.0 with variance stabilizing transformation (VST).[40]

2.3 | ATACseq data

Significant peaks called from ATACseq data analysis were retrieved from the original paper.[32] Briefly, peaks were called using MACS2 and filtered using a custom blacklist. A final set of 590 650 significant peaks were defined among a list of non-overlapping maximally significant 500 bp peaks ranked by their summit significance value. These significant peaks were re-annotated using HOMER with the command "annotatePeaks.pl" and two different references: Gencode v33 only and Gencode v33 with the previously used HERV annotation. Regions containing significant peaks around +/− 1000 or 3000 bp of a HERV TSS were considered to be active HERV regions.

2.4 | Unsupervised hierarchical clustering

For the sorted cells, DESEQ2 VST normalized expression data were directly used for unsupervised hierarchical clustering. Cluster purity was used as an external validation criterion and was calculated by first creating a confusion matrix between assigned cluster number and annotated cell type before adding the maximum values from each row (i.e. assigned cluster) and dividing by the total number of samples.

A benchmark of distances (euclidean, maximum, and pearson) and methods (ward.D2, single, complete, average and centroid) was performed to identify the optimal method leading to the best cluster purity using a pre-defined number of clusters according to the original annotation.

For bulk data sets, DESEQ2 VST normalized expression data were independently calculated and further center-scaled for each data set to correct the potential batch effect. Unsupervised hierarchical clustering was then performed using the average silhouette width and the Bayesian Information Criterion as internal validation markers.

2.5 | Differential ATAC-count analysis

For differential ATAC-count analysis, raw ATAC-seq counts were retrieved from the original paper.[32] Differential expression analysis between each AML population (LSC, pHSC, and Blasts) and their normal counterparts (HSC, GMP, LMPP, and monocytes) was performed using DESEQ2, with cell type as a covariate. Differentially expressed regions surrounding a HERV TSS (+/− 20 000 bp) and with a FDR <5% were retained for the final plot. The rolling mean of 1000 sequential regions, ordered by chromosome location, was then represented.

2.6 | HERVs, genes, and copy number variation correlations

HERVs located in extended AHR (peaks +/− 20 000 bp of a HERV TSS) were selected for correlation analysis. For each HERV, a list of surrounding genes located at +/− 50 000 bp of their TSS was established. Pearson's correlations were calculated between the RNA expression of each HERV and each of its surrounding gene, independently. P-values were corrected with the FDR method. Genes were then annotated using a published list of cancer-related genes from the Cancer Gene Census.[41] The same list of HERVs was then used to perform correlations with CNV from the same cytoband. TCGA LAML CNV data were retrieved from the NCI GDC portal and used as is to calculate Pearson's correlations with HERVs from the same cytoband.

2.7 | Survival analysis

Intensively treated patients with available survival data were retained for the survival analysis. Patients receiving hypomethylating agent therapies or supportive care were discarded from this analysis. Overall survival curves were estimated with the Kaplan–Meier method. Only known prognostic factors (Age, ELN2017, white blood count), batch and clusters were integrated in the final multivariate cox model.

For the second survival analysis (shown in Figure S6), intensively treated HLA-A*02 patients were selected and stratified according to the P1 expression level, cut in terciles. The same covariates were included in a multivariate cox model of overall survival. Survival analyses were performed using the R survival and survminer packages. Cox model proportional-hazards assumptions were assessed using Schoenfeld's test and inspection of residual plots.

2.8 | Cancer hallmark and immune signatures GSVA

For each cancer hallmark,[42] a unique gene signature was established (Table S3) based on The Molecular Signatures Database (MSigDb) Hallmark Gene Set Collection.[43] When not available in MSigDb, hallmark signatures were established from Gene Ontology (GO) signatures,

as previously described.[44] The signature for the immune evasion hallmark was retrieved from Hubert et al.,[45] Individual enrichment scores were calculated from each patient by single sample gene-set variation analysis (ssGSVA),[46] and scaled by study. The mean score for each cluster was then calculated and shown in a radar plot.

Immune signatures were obtained from Thorsson et al.[47] and calculated by ssGSVA for each sample. Unsupervised hierarchical clustering was then performed on study-scaled ssGSVA scores in each cluster.

2.9 | HERV-LSC signature

To establish the HERV-LSC signature, correlations between the expression of each unique HERV and the validated LSC17 score[48] were computed independently in the 4 bulk RNA-seq data sets. 47 HERVs showing a significant correlation (FDR-adjusted p-value <0.05) with the LSC17 score in at least 2 data sets and with concordant results (i.e. correlated in the same direction in each data set) were retained to build signatures. Signatures were calculated as the mean expression of all HERVs pondered (multiplied) by the sum of their Pearson's correlation coefficients. To select a minimal number of HERVs, the resulting 47 HERVs were ranked according to the absolute sum of their Pearson's correlation coefficients, and signatures were iteratively built by removing the last candidate (i.e. the one with the lowest summed correlation). Signatures' performances were then evaluated in the independent testing set of sorted cells from Corces et al.,[32] ROC curves and AUC were drawn and calculated with the plotROC R package. The signature showing the best performances for LSC classification with the lowest number of HERVs was eventually selected, leading to the final 25-HERVs signature.

2.10 | Differential HERV expression analysis

Differential expression analysis was performed using DESEQ2.[40] Raw counts of HERVs and genes from all normal and AML data sets were merged and integrated into the same DESEQ object, using study (i.e. batch) as a covariate in the design formula. Differential expression analysis was performed for all the 4 independent bulk AML data sets and the sorted LSC and pHSC populations against each of the 42 normal tissues. Fold change were shrunk with the apeglm method.[49] Features with a fold change superior to 4 (log2FC >2) and a base mean of at least 1 normalized count per million were considered to be overexpressed.

2.11 | Open-reading frame detection and peptide selection

ORFs were defined using EMBOSS's sixpack v6.6.0.0[50] to translate HERV sequences into the 6 possible frames. ORFs of at least 10 amino-acids were then aligned to a reference of known HERV proteins (Gag, Pro, Pol, Env, Rec, and Np9 from different HERV families). This reference was established from Uniprot,[51] referencing all existing manually reviewed HERV protein sequences, resulting in 71 references (research equation: keyword:"endogenous retrovirus" AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]", last accession: 12/11/2020). Blast was used for the alignment, with optimal parameters for retroviruses (Word size = 3, composition-based statistics, no "low-complexity-region" filter). ORFs aligning with at least 75% identity with a known HERV protein and an *E*-value < 0.01 were considered for further analysis.

Retained ORFs were then screened for predicted HLA-A*02 strong binders using MHCflurry v2.0[52] and netMHCpan 4.1.[53] Peptides with an affinity rank <= 0.5th percentile for both tools were considered to be strong-binders. The human proteome was downloaded on Uniprot (ID: UP000005640) to validate the absence of match before peptide synthesis and in vitro validation.

2.12 | Additive model

Differential expression at the peptide level between normal bone marrow and AML cells was estimated by relocating each peptide in all the potential HERVs with predicted ORFs coding for one of the 262 peptides. Each peptide could thus have one or several differential expression data according to the number of HERVs they were found in. This total number of HERVs was considered by considering the total sum of fold change of each HERV containing the peptide of interest, leading to an estimation of the cumulated fold change per peptide. To avoid prioritization of targets with artificially high fold-changes due to very low expression levels, each fold change was pondered by the base mean of its respective HERV, leading to the final additive model.

2.13 | Biological samples

Bone marrow samples were collected from AML patients at diagnosis at the Centre Hospitalier Lyon Sud in Lyon, France. Sample collection was approved from the institutional review board and ethics committee (20.01.31.72653–21/20_3) and after obtaining the written informed consent of patients, in accordance with the Declaration of Helsinki. BMMCs were obtained by Ficoll density gradient centrifugation (Eurobio, FR, EU) and immediately cryoconserved in fetal bovine serum (FBS) with 10% dimethyl sulfoxide (DMSO).

2.14 | MILs growth

Bone marrow mononuclear cells (BMMCs) were rapidly thawed at 37°C and grown in RPMI medium (Gibco, FR, EU) supplemented with 8% human AB-serum (Etablissement Français du Sang, FR, EU) and high doses (6000 UI/ml) of IL-2 (PROLEUKIN aldesleukine, Novartis Pharma, CH, EU) after a 2-hour resting period. Plates were then incubated for 14 days, with medium replacement when needed.

2.15 | MHC Dextramer® reagents and flow cytometry analysis

After 14 days of growth, cells were washed in 2 ml washing buffer (PBS + 2% FBS + 2 mM EDTA [Sigma Alderich, MI, US]) and stained for 10 min with MHC Dextramer reagents (Immudex®, DK, EU) at room temperature prior to viability and surface marker staining. Washing was performed twice to remove unbound reagents and avoid non-specific binding. An MHC Dextramer carrying non-sense peptide sequences (A*0201/ALIAPVHAV, ref. WB2666, Immudex®, DK, EU) was used as a negative control representing background staining. PBMCs and bone marrow from healthy donors were used as experimental controls. An overall frequency of at least 0.01% of living $CD8^+$ T-cells in the absence of significant background staining (non-sense peptide) was required to consider the positivity of MHC Dextramer staining.

2.16 | Generation of P1-specific $CD8^+$ T-cells

PBMCs from HLA-A*02 healthy donors were obtained by Ficoll density gradient centrifugation (Eurobio, FR, EU). Monocytes were isolated from PBMCs by positive selection of $CD14^+$ cells (Myltenyi, GE, EU). The negative fraction was considered as peripheral blood lymphocytes (PBLs). Cells were frozen in FBS 10% DMSO and kept at −80°C. Monocyte-derived dendritic cells (MoDCs) were generated from 6-day cultures of $CD14^+$ monocytes in complete RPMI medium supplemented with 10% FCS and recombinant human GM-CSF (100 ng/ml) and IL-4 (50 ng/ml). MoDCs were pulsed overnight with the cognate peptide (10 μg/ml) and 10 ng/ml of LPS (Ultrapure LPS, *E. coli*, InVivoGen, FR EU) and washed before co-culture with isolated PBLs for 6 days (MoDCs:T-cells ratio 1:10) in 96-wells plates. After 6 days, T-cells were counted and restimulated with autologous MoDCs for 6 more days.

On day 12, MHC dextramer-positive $CD8^+$ T cells were sorted and expanded on a feeder composed of 35 Gy-irradiated allogeneic PBMCs and B-lymphoblastoic cell lines in a ratio of 10:1. Feeder cells were plated in a 96-well round bottom plate at a concentration of 0.10×10^6 cells per well in RPMI 5% human serum with PHA-L 1.5 μg/ml (Merck KgAa, GE, EU) and IL-2150 IU/ml (Novartis Pharma, CH, EU). Up to 5×10^3 sorted cells were added per well. Cells were cultured for 14 days and medium was replaced when needed with fresh IL-2 enriched with RPMI 5% human serum. At the end of feeding $CD8^+$ T cells were checked for their positivity to dextramer staining (>60% for functionality experiments).

2.17 | Functionality analysis

Dextramer-selected P1-specific $CD8^+$ T cells were co-cultured with the HLA-A*02 AML cell line THP1 (DSMZ, GE, EU) in a ratio 10:1. T2 cells loaded with P1 or the irrelevant HERV-derived peptide (SMDDQLNQL) were used as positive and negative controls of specificity, respectively. After 1 h of co-culture, Golgi Plug (BD, NJ, US) and CD107a antibody (BD, NJ, US) were added into wells and co-cultures were maintained for further 4 h. An anti-human MHC I blocking Ab (Clone W6/32, InVivoMab, BioXcell, NH, US) was added in selected wells for 1 h before co-culture (50 μg/ml) and maintained at 10 μg/ml during the entire time of the co-culture to neutralize the HLA-A*02 dependent T cell activation, as control. Viability (Zombie NearIR, Biolegend, CA, US), extracellular (CD3, BV421/ CD8, FITC both Biolegend, CA, US), intracellular staining (IFN-γ PE, Biolegend, CA, US and TNF-α Pe-Cy7, BD, NJ, US) and fixation were then performed and samples were analyzed on a LSR Fortessa (BD, NJ, US).

3 | RESULTS

3.1 | The HERV retrotranscriptome accurately defines normal hematopoietic cell populations

As a first step, we examined HERV expression in different normal hematopoietic cell populations, assuming that distinct HERV profiles may characterize the main cell types. Using a custom pipeline based on Telescope,[36] we quantified the expression of 14 968 HERV loci in public RNA-seq data from sorted bone-marrow and peripheral blood cell populations from 9 healthy donors (spanning 15 different sorted-cell populations for a total of 49 samples, Figure S1A).[32] Unsupervised hierarchical clustering based on the top 20% most variable HERVs showed a robust classification of normal hematopoietic cell types with a cluster purity of 77.6% and a corrected Rand Index of 0.61 (Figure 1A). The same approach based on genes reached a cluster purity of 65.3% with a corrected Rand Index of 0.47 (Figure 1A).

We then sought to improve the clustering using ATAC-seq data through the analysis of peaks from open chromatin regions. Using the HOMER package, we applied a classic human genome annotation from gencode (v33) to annotate the set of 590 650 significant non-overlapping peaks from open chromatin regions previously defined in bone marrow and peripheral blood cells sorted from healthy donors ($n = 80$ samples).[32] As previously described, unsupervised hierarchical clustering based on promoter elements (peaks between −100 bp and 1000 bp away from a transcription start-site [TSS]) and intergenic elements (peaks more than 1000 bp away from any other feature) significantly improved cluster classification, with a purity reaching 81.8% (Figure S1B). We then re-annotated these significant peaks with a custom reference consisting of the same Gencode annotation concatenated with the previously used 14 968 HERV loci from Repeatmasker. This new annotation showed that 16% of the significant peaks correspond to HERV regions (Figure 1B). One important previously reported finding is that classification based exclusively on intergenic elements (the so-called "distal regulatory elements") is sufficient to classify normal hematopoietic cell populations.[32] Enhanced annotation of these distal regulatory elements revealed an enrichment in HERVs, with up to 37.6% of the top 500 variable intergenic peaks corresponding to a HERV region (Figure 1B). A plot of the total aggregated count from these regions showed a gaussian distribution

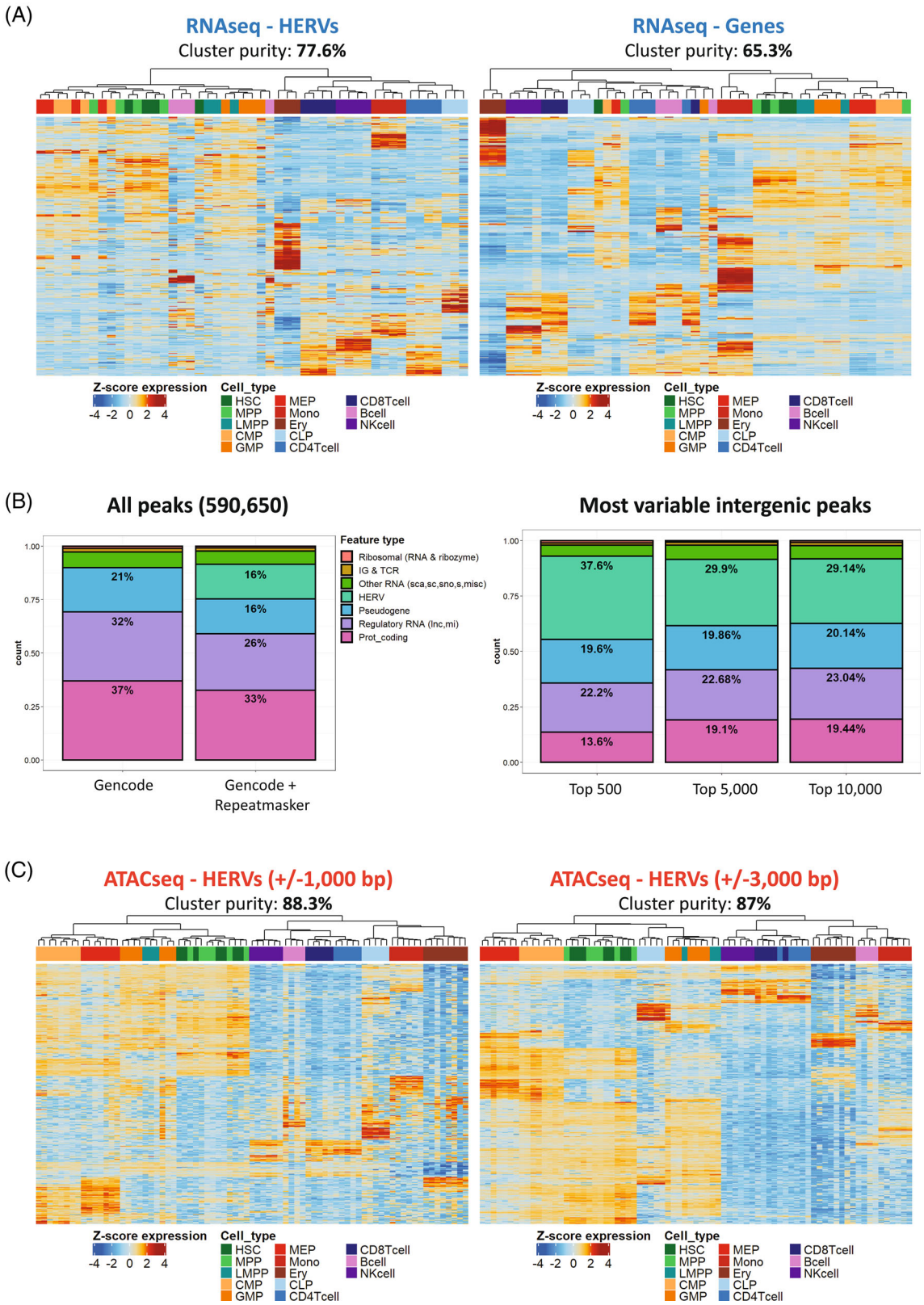


FIGURE 1 Definition of normal hematopoietic cell populations based on the HERV retrotranscriptome. (A) Unsupervised hierarchical clustering of normal hematopoietic cell populations based on HERV (left) or gene (right) expression in RNA-seq. Clustering was performed with the ward. D2 method based on the maximum distance. (B) Annotation of all significant ATAC-seq peaks (left) and top variable intergenic peaks (right) with a custom annotation of Gencode v33 with HERV references from Repeatmasker. (C) Unsupervised hierarchical clustering of normal hematopoietic cell populations based on ATAC-seq peaks from HERV regions (+/− 1000 (left) or 3000 (right) bp from HERVs' TSS). Clustering was performed with the ward. D2 method based on the maximum distance. CLP: Common Lymphoid Progenitor, CMP: Common Myeloid Progenitor, Ery: Erythrocyte, GMP: Granulocyte-Macrophage Progenitor, HSC: Hematopoietic Stem cell, LMPP: lymphoid-primed multipotent progenitor, MEP: Megakaryocyte-Erythroid Progenitor, MPP: Multipotent Progenitor. [Color figure can be viewed at wileyonlinelibrary.com]

surrounding HERV TSS, confirming the good quality of the ATAC-seq signal (Figure S1C). Clustering of samples based on active HERV regions (AHR, defined by peaks surrounding HERVs regions +/− 1000 or 3000 bp) further improved the clustering, reaching 88.3% cluster purity (Figure 1C).

Altogether these results show that the HERV retrotranscriptome can be used to characterize normal immature and mature hematopoietic cell populations. The improved clustering obtained with ATAC-seq data suggests that AHR reflect epigenetic footprints associated with cell differentiation.

3.2 | Acute myeloid leukemia cells show distinct HERV profiles close to their normal cell of origin

We next evaluated how the HERV retrotranscriptome may help to distinguish AML cells. We performed the same clustering approach, integrating the 32 RNA-seq and 45 ATAC-seq bone marrow samples from 12 AML patients at diagnosis to our previous samples (Figure S1A).[32] Unsupervised clustering based on the top 20% variable AHR (+/− 1000 bp from a HERV TSS) in ATAC-seq resulted in a good classification of normal and AML cells, with a slight increase in cluster purity compared to the top 20% most variable intergenic peaks (Figure 2A). Clustering based on HERV expression in RNA-seq yielded comparable results (Figure S2A). Interestingly, leukemic blast cells (blasts) clustered with either monocytes or granulocyte-monocyte progenitor (GMP) cells, LSCs with either GMP or lymphoid-primed multipotent progenitor (LMPP) cells and pre-leukemic hematopoietic stem cells (pHSCs) with either GMP or HSC/multipotent progenitor (MPP) cells, suggesting a clustering with their cell of origin as previously described.[32] Cluster purity based on the original categories of cells did not consider these similarities and was thus a poor indicator of clustering performance in this case. Differential ATAC-count analysis centered on extended AHR (+/− 20 000 bp from a HERV TSS) revealed distinct profiles between AML LSCs, blasts and pHSCs compared to their normal counterparts, with globally a more open chromatin in blasts and a more closed chromatin in LSCs and pHSCs (Figure 2B). To further characterize the role of HERVs in these AHR, we computed correlations between RNA expression of each HERV present in an AHR and its respective surrounding genes located at +/− 50 000 bp. Strikingly, we mostly found positive correlations between HERV expression and their surrounding genes (Figure S2B). Annotation of the genes with a pre-established list of cancer-associated genes from the Cancer Gene Census database[41] revealed several genes positively correlated with HERVs expressed in AHR (Figure 2C). Of note, the highest correlation was found for GATA1 with ERVLB4_Xp11.23b (Pearson's *R*: 0.74, adjusted *p*-value: 8.11e-14) (Table S1). Using TCGA LAML RNA-seq data, we then explored the association between each HERV located in an AHR and gene copy number variation (CNV) on the same cytoband. This highlighted several HERVs correlated both positively and negatively with deletions, and mostly positively with amplifications on the same cytoband (Figure 2D). These results demonstrate that HERV expression profiles differ according to the AML cell type, and suggest that HERVs are associated with gene regulation.

3.3 | HERV expression defines subtypes of AML with distinct cancer hallmarks and outcomes

Having established that the HERV retrotranscriptome characterizes particular cell types, we then wondered whether HERV expression could define distinct AML profiles in bulk RNA-seq data. We explored HERV expression in 4 independent RNA-seq data sets (TCGA, AMLCG, LEUCEGENE, and BEAT), retaining only bone marrow samples from AML patients at diagnosis ($n = 788$) (Table S2). For each data set, we selected the top 2000 most variable HERVs based on the scaled DESEQ2 VST normalized count (Figure S3A). We merged the 4 data sets, keeping only the intersect between each top 2000 candidate HERVs, resulting in 961 variable HERVs conserved across the 4 data sets. Unsupervised hierarchical clustering guided by the average silhouette (Figure S3B) and Bayesian Information Criterion (BIC) evolution defined 9 clusters that were not dependent on the study (Figure 3A and Figure S3C). These 9 clusters were associated with significant differences in overall survival among intensively treated patients (Figure 3B), independently of established prognostic factors such as age, ELN2017 and white blood count, integrated in a multivariate Cox model (Figure 3C). These clusters also presented distinct cancer hallmark profiles (Figure 3D), as assessed by single sample gene-set variation analysis (GSVA) based on cancer hallmark signatures (Table S3). GSVA of immune signatures[47] revealed no clear immune subtype profiles, distinguishing clusters with globally low or high immune scores (Figure 3E).

We then assessed whether these clusters were associated with known recurrent translocations or gene mutations. We found a clear enrichment in inv,[16] t(8;21), and t(15;17) in clusters 1, 7, and 9, respectively (Figure 3F). Other karyotypes (such as complex, poor or intermediate abnormalities) showed no particular association with any cluster, underlining the heterogeneous composition of these groups. Regarding gene mutations, we observed a tendency for an enrichment in RUNX1 mutation in cluster 5 and in TP53 mutation in cluster 3 (Figure 3F).

Focusing on HERVs discriminating each cluster (i.e. HERVs overexpressed in a given cluster compared to all other clusters), we found that clusters 8 and 9 had the highest number of different discriminating HERVs, whereas only few HERVs discriminated clusters 3 and 5 (Figure S3D). HERVs from the HERV-H, ERV-L, MER4, HERV-L, and HERV-K families were the most frequent HERVs to discriminate clusters (Figure S3E). When reported to the total number of HERVs per family, HERV-S, HERV-E, HERV-P, ERV1, and HARLEQUIN families were the most frequently represented (Figure S3F).

To further demonstrate the value of HERV expression as a promising biomarker, we next established a LSC signature based on HERV expression. We calculated correlations between each individual HERV and the previously published LSC17 score[48] in the 4 independent data sets. Forty-seven HERVs with a significant correlation with the LSC17

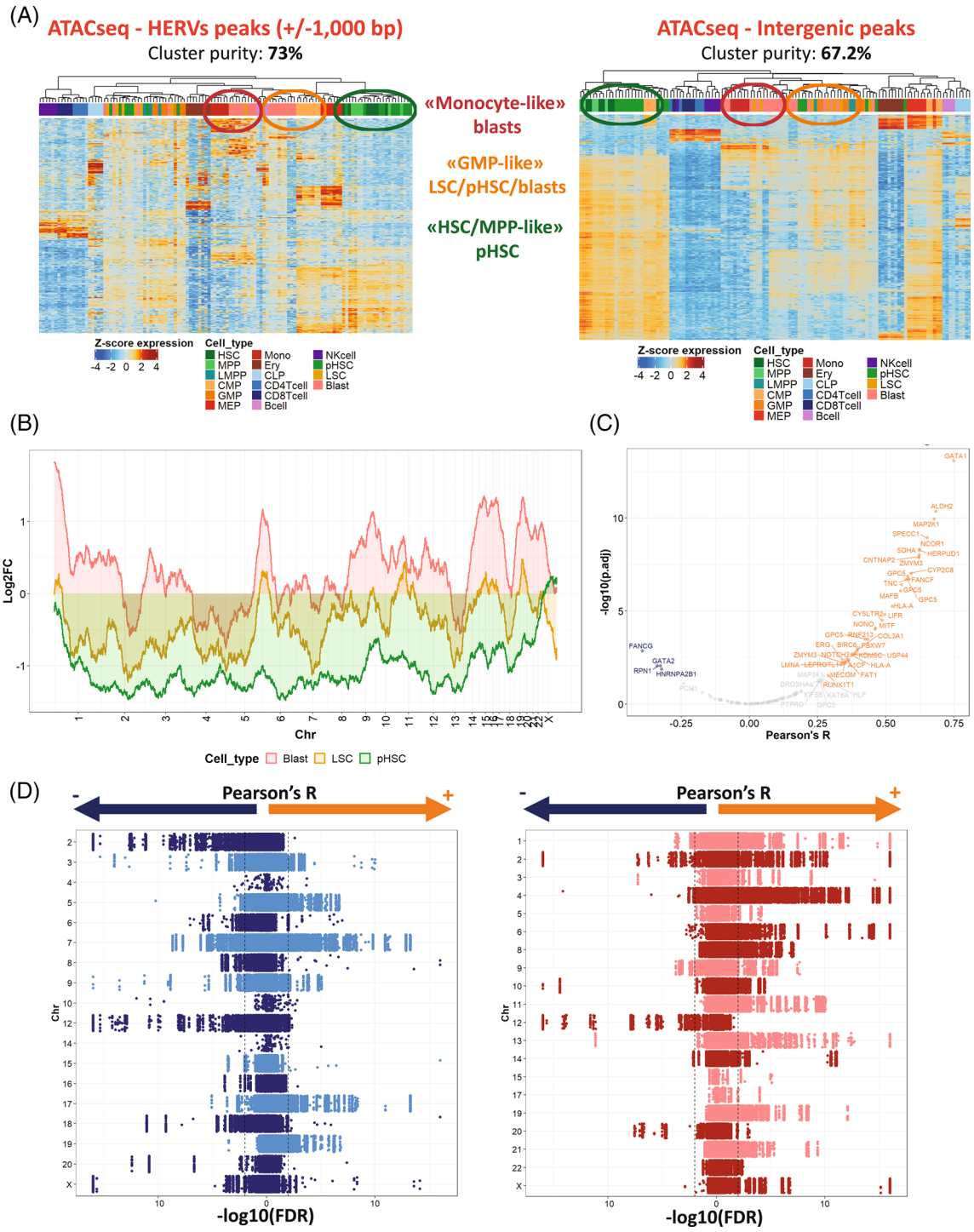


FIGURE 2 AML cells show distinct epigenetic profiles compared to their normal counterpart. (A) Unsupervised hierarchical clustering of normal and leukemic hematopoietic cell populations based on HERV-centered (left) or global intergenic (right) open regions in ATAC-seq. Clustering was performed with the ward. D2 method based on the maximum distance. (B) Differential ATAC-count analysis between AML and normal BM cells. Log2FC between each AML subpopulation (blast, LSC, pHSC) and normal BM cells is shown on the y axis, and chromosomal location on the x axis. Only values with an FDR <0.05 were considered. (C) Correlation between RNA expression of HERVs located in active chromatin regions and the surrounding cancer-associated genes (+/− 50 000 bp). Significant Pearson's R are represented in orange (positive) or blue (negative). *p*-values were corrected using the FDR method. (D) Correlation between RNA expression of HERVs located in active chromatin regions and CNV: deletions (left) and amplifications (right). Individual chromosomes are represented on the y axis, −log10(FDR) on the x axis. Pearson's R were independently calculated for each HERV and all the alterations present on the same cytoband. AML: Acute Myeloid Leukemia, BM: Bone Marrow, CLP: Common Lymphoid Progenitor, CMP: Common Myeloid Progenitor, CNV: Copy Number Variation, Ery: Erythrocyte, FDR: False Discovery Rate, GMP: Granulocyte-Macrophage Progenitor, HSC: Hematopoietic Stem cell, LMPP: lymphoid-primed multipotent progenitor, LSC: Leukemic Stem Cell, MEP: Megakaryocyte-Erythroid Progenitor, MPP: Multipotent Progenitor, pHSC: pre-leukemic Hematopoietic Stem Cell. [Color figure can be viewed at wileyonlinelibrary.com]

score (adjusted p-value < 0.05) in at least 2 independent data sets were used to establish a new LSC score based on HERVs expression only (Table S4, see methods). This 47-HERV LSC signature was further refined to select the minimal number of HERVs allowing a good LSC discrimination. The resulting LSC signature based on 25 different HERVs showed a very good discrimination in the independent validation set of sorted AML cells, separating LSCs from all other cells with an area under the curve (AUC) of 0.80 versus 0.64 for the non-refined signature and 0.76 for the LSC17 score (Figure 3G). Both the full 47-HERV and the 25-HERV LSC signatures showed a significant impact on overall survival, mirroring the LSC17 signature's performances (Figure S3G). Altogether these results show that HERVs represent an important source of genetic information that can be used to define different AML subtypes as well as cell-specific signatures, as highlighted herein for LSCs.

3.4 | AML-specific HERVs contain several open-reading frames representing a source of shared tumor antigens

Considering the specific expression of some HERVs in AML, we evaluated if they may represent a source of shared tumor antigens for immunotherapeutic approaches. To define leukemia-specific targets, we used a data set of more than 1000 normal samples from 42 different sites extracted from the Genotype-Tissue Expression (GTEX) database in addition to the normal bone marrow and peripheral blood samples used above. We performed differential HERV expression analysis between AML and normal tissues, considering transcripts with a shrunken fold change >4 and a base mean of at least 1 normalized count per million as overexpressed. To be considered as AML-specific, each HERV candidate had to be overexpressed in at least 2 independent data sets of AML bone marrow compared to normal bone marrow cells and not overexpressed in any of the 42 normal tissues compared to any AML dataset. These two filters led to the final identification of a set of 125 conserved AML-specific HERVs (Figure S4A). The mean AML expression of these 125 AML-specific HERVs highly exceeded in most of cases the 75th percentile expression in normal tissues, with the exception of testis (Figure 4A). When focusing on the top overexpressed HERVs, we found candidates with a particularly high differential expression, with log2FC up to 17.47 and 16.58 for ERV316A3_1q25.2b in bulk AML cells and LSCs, respectively. Several HERVs had an expression level in LSCs compared to normal bone marrow cells at least as high as other current LSC therapeutic targets (Figure S4B).

To screen for potential HERV-derived epitopes, we next defined all the putative ORFs by translating these 125 AML-specific HERVs into the 6 possible frames. To reduce the number of false positives, ORFs of at least 10 amino-acids were then aligned against a manually annotated database of 71 known existing HERV proteins derived from curated Uniprot references. ORFs with at least a 75% identity with an existing HERV protein and an E-value < 0.01 were selected for epitope screening. Using these criteria, we found that only 33/125 AML-specific HERVs possessed at least 1 ORF (Figure S4C). Two independent recent tools trained on mass spectrometry data were then used for epitope prediction: netMHCpan 4.1[53] and MHC flurry 2.0.[52] HLA-A*02 epitope screening revealed 262 unique peptides predicted as strong HLA-A*02 binders with both independent tools.

To set up a list of peptides to prioritize, we further ranked these results. Considering that epitope immunogenicity is linked to its tumor abundance,[54] we established an additive model considering (i) the mean expression of HERVs containing the peptide sequence, (ii) the fold change compared to normal bone marrow and (iii) the number of different HERVs containing the peptide sequence. We relocated each peptide in all the different HERVs containing its sequence and added the fold changes pondered by the base mean expression of each individual HERV. This additional step filtered-out 158 peptides that were either shared with HERVs present in normal tissue (pondered negatively in our additive model) or overexpressed but with a very low base expression. Among the 104 remaining peptides (Figure 4B), the top expressed candidates were mainly from Gag and Pol proteins of HERV-K/HML-2 family (Figure S4D).

Overall, these results show that HERVs represent a reservoir of potential $CD8^+$ T cell epitopes overexpressed in AML that may represent original therapeutic targets.

3.5 | T-cells specific for HERV-derived epitopes are naturally present in AML patients

We then examined whether the predicted HERV-derived HLA-A*02 epitopes could elicit an immune response in AML patients. We selected 8 different HLA-A*02 epitopes among the top candidates in the additive model, focusing on peptides already identified as immunogenic in our lab (P1, P2, P4, and P6) or peptides preselected in a previous study in solid tumors but for which the immunogenicity has so far not been confirmed (P15, P16, P18, and P20).[55] Bone marrow-infiltrating lymphocytes (MILs) were expanded from bone marrow mononuclear cells (BMMCs) of HLA-A*02 AML patients at diagnosis. High-dose IL-2 stimulation induced expansion of $CD45^+$ cells from 14.4% to 74.3% (mean values) among living cells in 14 days. Using HLA-A*02-peptide dextramers, we screened for the presence of specific $CD8^+$ T cells against the selected epitopes. Strikingly, P1-specific $CD8^+$ T cells were found in 7/10 patients. Specific $CD8^+$ T cell against P15 and P16 were also detected at frequencies >0.01% in 1/10 and 2/10 patients, respectively. No significant dextramer staining was observed on $CD8^+$ T cells expanded from HLA-A*02 normal bone marrow or $CD8^+$ T cells obtained from HLA-A*02 healthy peripheral blood mononuclear cells (PBMCs) (Figure 4C, D, Figure S5).

Based on these results, we decided to evaluate the functionality of P1-specific $CD8^+$ T-cells generated from HLA-A*02 healthy donor's PBMCs. After priming by P1-pulsed autologous monocyte-derived dendritic cells, P1-specific $CD8^+$ T cells were sorted by flow cytometry using dextramer staining and expanded on feeder cells (see methods). To confirm the peptide-specific activation of these expanded T cells, we evaluated the cytokine release following co-

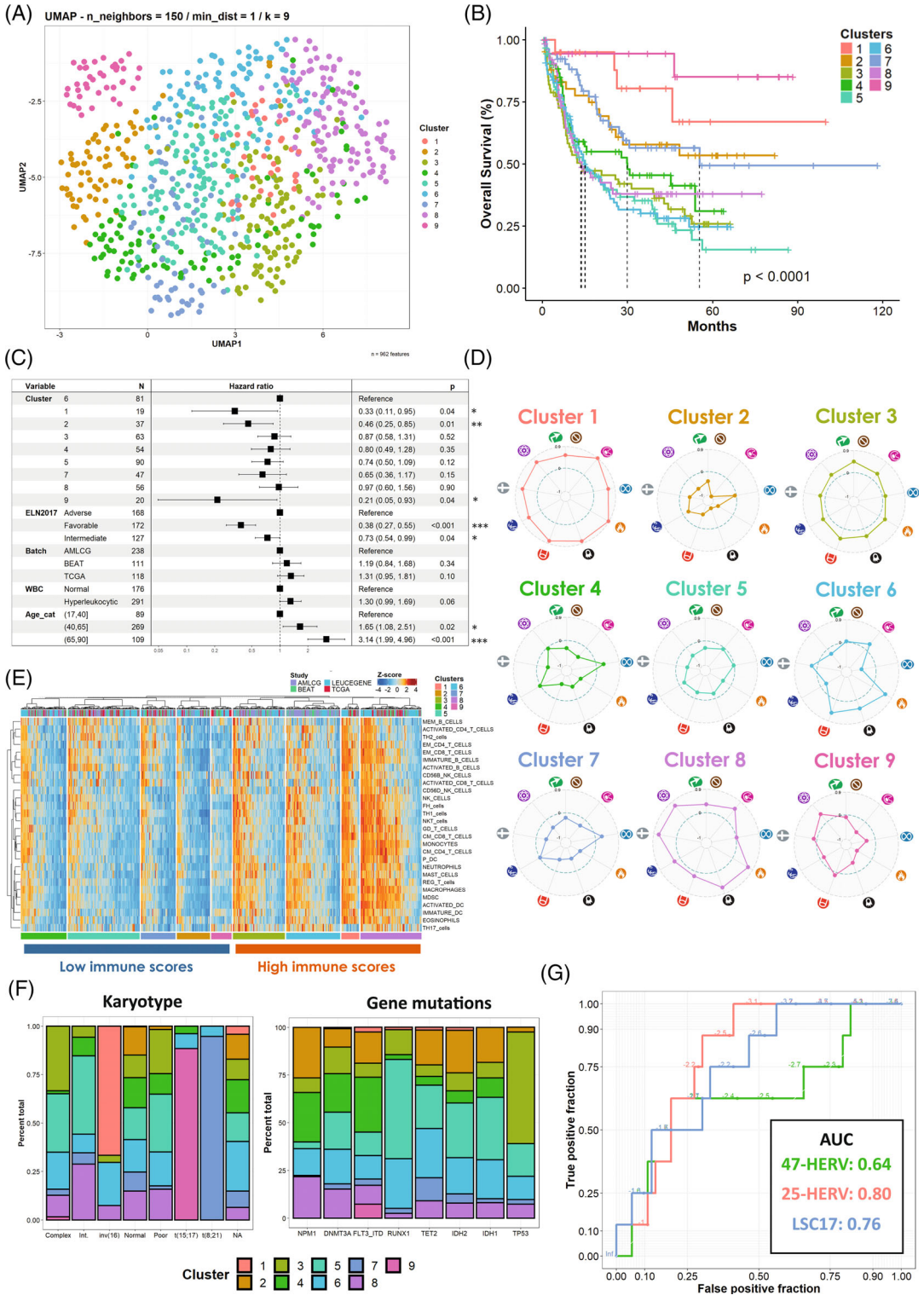


FIGURE 3 The HERV retrotranscriptome defines AML subtypes with distinct prognosis and cancer hallmarks. (A) UMAP representation of the 788 AML patients. Clusters defined by the unsupervised hierarchical clustering approach are shown. (B) Overall survival of intensively treated patients according to the 9 clusters in the whole cohort. (C) Multivariate Cox analysis of overall survival of intensively treated patients. Known risk factor (Age, ELN2017 and WBC), study (batch) and clusters are integrated in the multivariate model. (D) Cancer hallmark profiles of each cluster. Each cancer hallmark is represented by its symbol as defined in.[42] For each cluster, hallmark scores are calculated by ssGSVA. Mean scores are represented on a radar plot. (E) Heatmap of immune signature enrichment. Patients were grouped into clusters before performing unsupervised hierarchical clustering. Z-scores of ssGSVA signature enrichment are represented. (F) Bar chart of genomic alteration (left) and mutation (right) distribution according to clusters. For each category, the total count and the percentage is represented. (G) ROC-curve of LSC classification according to the 47-HERV (green), 25-HERV (red) and LSC 17 (blue) LSC signature. LSC: Leukemic Stem Cell, ROC: Receiver Operating Characteristic Curve, ssGSVA: Single Sample Genes-set Variation Analysis, WBC: White Blood Count. [Color figure can be viewed at wileyonlinelibrary.com]

culture with T2 cells loaded either with P1 or an irrelevant HERV-derived peptide. IFN-γ and TNF-α release was observed only when the target cells were pulsed with P1 and was inhibited by the addition of an anti-HLA class I blocking antibody (Figure 4E and Figure S6A). We then evaluated the functionality of these P1-specific CD8$^+$ T cells against THP-1, a HLA-A*02 AML cell line shown to express P1-containing HERVs (Figure S6B). An increased production of IFN-γ and TNF-α (intracellular staining) and an increase in the percentage of

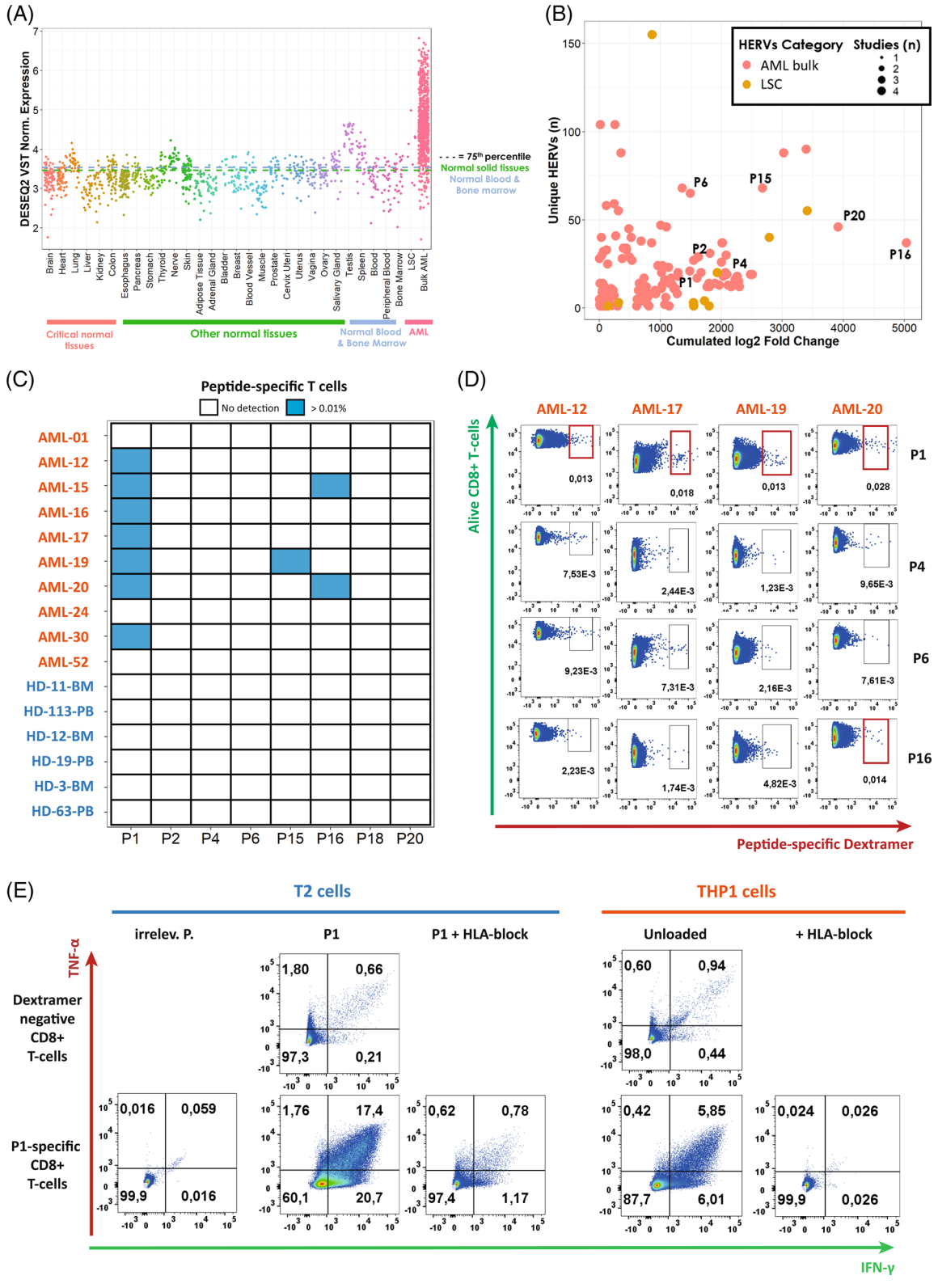


FIGURE 4 Legend on next page.

degranulating (CD107a positive) cells was observed when P1-specific CD8$^+$ T-cells were co-cultured with THP-1 cells (Figure 4E and Figure S6A). These effects were reversed with the anti-HLA-I blocking antibody. Furthermore, no increase in cytokine production was found when unspecific CD8$^+$ T-cells (dextramer-negative CD8$^+$ T-cell fraction sorted and expanded in the same conditions) were incubated with THP-1. Altogether, these results demonstrated a specific and HLA-I restricted activation of P1-specific CD8$^+$ T cells in presence of AML cells.

To further explore the potential impact of P1 expression in AML cells, we assessed the prognostic value of the expression of HERVs containing its sequence among HLA-A*02 patients on the overall cohort. Multivariate Cox analysis revealed that patients with a high expression of P1-containing HERVs in leukemic cells had a better overall survival than patients with a low or intermediate expression, regardless of age, ELN2017, batch and white blood count (Figure S6C). Overall, these results suggest that HERV-derived epitopes elicit a CD8$^+$ T cell response that can be detected in the bone marrow of AML patients. Furthermore, a higher expression of HERVs containing immunogenic epitopes may be associated with a better clinical outcome.

4 | DISCUSSION

Representing 8% of the human genome, HERVs are a great source of genetic information. However, as the results of their analysis are largely dependent on the chosen analytical strategy, this signal can often be missed or misinterpreted. For instance, using a partial database limited to HERVs corresponding to known Uniprot references (56) can occult a major part of the HERV retrotranscriptome, and restricting HERV signal to uniquely mapped reads strongly reduces the signal from conserved, recently integrated families prone to multimap.[57] Consistently, we used a custom-designed pipeline based on Telescope,[36] a recently developed tool allowing the accurate quantification of a complete base of 14 968 HERV loci. One of the main advantages of Telescope resides in its reassignment algorithm to correctly reassign a high number of multimaps and thus keep the highest possible signal from recently integrated and conserved HERV families. This approach provided a thorough analysis of HERV expression in normal and AML cell populations, providing a robust classification of specific cell-types only based on HERV transcriptomes.

Chromatin accessibility represents a major factor of cell differentiation that may capture cell identity more accurately than gene expression in blood cells.[32] Furthermore, HERVs may actively shape the chromatin architecture.[58] In line with these observations, we show here that a very efficient clustering of normal hematopoietic cell populations is obtained with the use of AHR located in distal regulatory elements. In the same way, AHR profiles allowed not only the discrimination of leukemia cells from normal blood cell populations, but also between leukemia cells clustering with their cell-of-origin. Thus, pHSCs, LSCs, and leukemic blast cells displayed a distinct epigenetic profile with distinct AHR compared to their normal counterparts. Using the unique information provided by the HERV retrotranscriptome, we built a signature based on 25 HERVs that allowed a robust classification of LSCs among normal and leukemic bone-marrow cells. An LSC-HERV signature could represent an original tool to either evaluate AML prognosis (i.e. as a surrogate marker of the remaining LSCs) or minimal residual disease during follow-up.

A positive correlation was found between HERV expression from AHR and their surrounding genes, including cancer-associated genes. This may reflect the presence of these elements in actively transcribed DNA regions. Alternatively, HERVs may exert a regulatory role in gene transcription in these regions, as already described.[59] In this context, it is tempting to speculate that some of these HERVs may act as enhancers for AML oncogenes, as reported recently by Deniz et al.[30] Interestingly, LSCs and pHSCs showed fewer AHR compared to blasts. This result is consistent with a previous report demonstrating that transposable elements are silenced in LSCs.[31] This may be due to the fact that silencing of HERVs might promote LSC immune escape, because HERVs have been described to induce both innate and adaptive immune responses.[16] However, as we show herein, some HERVs containing conserved ORFs are still overexpressed in LSCs compared to normal cells, suggesting that it may be possible to therapeutically boost an immune response against HERV epitopes in LSCs.

Relying only on HERV expression, we managed to retrieve particular AML subtypes but also to define new subtypes within normal and

FIGURE 4 HERVs as a source of shared epitopes in AML. (A) Scatter plot showing the mean expression of the 125 AML-specific HERVs across normal and AML tissues. Dotted lines represent the 75th percentile expression of normal solid and hematopoietic tissues. (B) Scatter plot of individual peptides in the additive model. For each peptide, cumulative log2FC between AML cells and normal BM (x axis) and number of unique HERVs containing the peptide sequence (y axis) is represented. Cumulative log2 FC were pondered by the base mean expression of the corresponding HERV. (C) Summary table of specific CD8$^+$ T cell responses found among MILs in patients. An overall frequency of at least 0.01% of living CD8$^+$ T-cells in the absence of significant background staining (MHC Dextramer with a non-sense peptide A*0201/ALIAPVHAV) was required to consider the positivity of MHC Dextramer staining. (D) Representative dextramer results of 4 AML patients. Bone marrow CD8$^+$ T cell populations are gated among single living cells after 14 days of expansion. CD8 staining is represented on the y axis, dextramer staining is represented on the x axis. Significant results are shown in red. (E) Functionality analysis of P1-specific CD8$^+$ T-cells. IFN-γ and TNF-α production are shown. Dextramer-selected P1-specific CD8$^+$ T cells were expanded for 14 days before being co-cultured with T2 or THP-1 cells in a ratio of 10:1. Intracellular (IFN-γ, TNF-α) and extracellular (CD107a) staining were performed after 5 hours of co-culture. Dextramer-negative fraction of CD8$^+$ T-cells was expanded with the same protocol and used as negative control (data representative of 3 independent experiments). AML: Acute Myeloid Leukemia, Factor FC: Fold Change, FDR: False Discovery Rate, IFN: Interferon, LSC: Leukemic Stem Cell, MILs: Marrow Infiltrating Lymphocytes, P1: Peptide 1, TNF: Tumor-necrosis. [Color figure can be viewed at wileyonlinelibrary.com]

complex karyotype groups. Importantly, the prognostic differences found across these AML subtypes were independent of the currently used ELN2017 score, based on the karyotype and mutated genes. These results suggest that HERVs could be used to improve risk stratification and treatment resistance prediction in patients with no genetic or molecular abnormalities associated with well-defined prognosis and resistance profiles.

Several studies have evaluated the possibility to predict survival or response to intensive therapy using gene-expression.[18,48,60–63] However, most of these studies do not account for major cofounding factors (such as treatment type and stem cell transplantation) or do not contain a validation cohort. When performed, well-made studies using validation cohorts hardly reach a concordance statistic of 0.8, most of them being around 0.7.[64,65] More recently, Vadakekolathu J. et al. proposed an immune score based on IFN-γ-related genes that was able to predict response to flotetuzumab (a CD3xCD123 bispecific antibody) with an AUC of 0.84.[22] Our results suggest that the systematic association of HERVs with gene expression could greatly improve gene expression-based predictors.

Finding specific tumor antigens in AML remains a major challenge for immunotherapeutic approaches. Because AML belongs to malignancies with the lowest mutational burdens,[23] the frequency of mutations creating neoantigens is expected to be low. In this context, HERV-derived antigens represent a unique source of alternative tumor-specific epitopes[25,26] that could be exploited for the development of new immunotherapies. Our bioinformatics-based approach allowed us to identify several $CD8^+$ T cell epitopes from AML-specific HERVs, that is, HERVs expressed at high levels in AML cells and never expressed in normal tissue. For practical reasons, we provided a proof-of-concept using HLA-A*02, the most common human MHC class I. The proposed pipeline can easily be adapted to other haplotypes, and it will be of interest to evaluate other HLA alleles to propose a broader population coverage.

The presence of $CD8^+$ T cells specific for HERV epitopes in the bone marrow of AML patients at diagnosis confirmed that these epitopes are naturally processed and are immunogenic. Using a proteogenomic approach, Ehx et al. recently showed that most MHC-class I tumor-specific antigens from primary AML samples originate from non-coding regions, encompassing introns and endogenous retroelements.[26] Our results substantiate a recent report from Saini et al. showing that HERVs represent a reservoir of $CD8^+$ T cell epitopes in myeloid malignancies.[66] In this latter study, a high-throughput epitope screening method was set up based on barcode dextramers. It would be interesting to combine both approaches to obtain an optimized epitope detection pipeline associated with a more systematic functional screening method. Of note, the P1 epitope (FLQFKTWWI), which seemed to be particularly immunogenic in our study and in another in the context of solid tumors[67] was screened by Saini et al. but not detected with the barcoded dextramers. This discrepancy could be explained by the use of peripheral blood samples or a lower sensitivity in the absence of prior expansion of bone marrow T-cells,[68] especially in AML, a disease characterized by a massive leukemic cell infiltration in the bone marrow.

In conclusion, our study unveils the HERV retrotranscriptome as a powerful tool to characterize and classify specific cell populations and to provide new tumor-specific epitopes for the development of immunotherapeutic strategies. This approach, demonstrated here in AML, could easily be extended to any disease to provide a better risk-stratification, establish relevant predictive signatures for therapeutic responses or identify original therapeutic targets.

5 | CONCLUSIONS

The HERVs retrotranscriptome represent an important source of genetic information that can be used to enhance disease stratification or identify cell populations, as shown here with leukemic stem cells. Moreover, HERVs represent an important reservoir of alternative tumor-specific T cell epitopes that can be identified and prioritized using our approach.

AUTHOR CONTRIBUTIONS

V. Alcazer: Conceptualization, data curation, formal analysis, funding acquisition, methodology, investigation, resources, software, validation, visualization, writing-original draft. **P. Bonaventura:** Data curation, formal analysis, methodology, resources, investigation, validation, visualization, writing – review & editing. **E. Michel, V. Mutez, C. Fabres**: Methodology, resources, investigation, validation, visualization, writing – review & editing. **L. Tonon, A. M. N. Batcha:** Software, investigation, writing – review & editing. **A. Viari:** Resources, project administration, **G. Sauvageau, K. H. Metzeler, W. Hiddemann T. Herold, N. Chuvin, Y. Estornes, R. Boulos:** Resources, writing – review & editing. **C. Caux:** Methodology, resources, writing – review & editing. **S. Depil:** Conceptualization, funding acquisition, methodology, resources, validation, project administration, writing – review & editing.

ACKNOWLEDGMENTS

The results/or some of the results obtained in this publication are based on data generated by the Leucegene group essentially located at IRIC in Montreal, Canada, and supported by Genome Canada and Genome Québec. This study thus relied on human AML specimens provided by the BCLQ, Montreal, QC. The authors thank François Mallet and Nicolas Dulphy for their feedback and help throughout the study, and Brigitte Manship for her help with English editing.

CONFLICT OF INTEREST

SD is founder and chairman of ErVaccine Technologies. VA and PB are consultants for ErVaccine Technologies. EM, NC, YE, RB, and VM are employees of ErVaccine Technologies. The other authors have no conflict of interests to disclose.

CLINICAL TRIAL REGISTRATION

None.

PATIENT CONSENT STATEMENT

None.

DATA AVAILABILITY STATEMENT

The data analyzed in this study were obtained from GEO at GSE74246, GSE49642, GSE52656, GSE62190, GSE66917, GSE67039, and GSE106272, and from the NCI Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/). THP-1 cell line data were accessed from the CCLE portal (https://portals.broadinstitute.org/ccle).

ORCID

Vincent Alcazer https://orcid.org/0000-0003-1843-6286
Nicolas Chuvin https://orcid.org/0000-0002-2138-2330
Rasha Boulos https://orcid.org/0000-0003-1607-2394
Yann Estornes https://orcid.org/0000-0003-2687-290X

REFERENCES

1. Consortium IHGS. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):35057062.
2. Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol*. 2019;17(6):355-370.
3. Vargiu L, Rodriguez-Tomé P, Sperber GO, et al. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology*. 2016;13(1):7.
4. Kassiotis G, Stoye JP. Immune responses to endogenous retroelements: taking the bad with the good. *Nat Rev Immunol*. 2016;16(4):207-219.
5. Alcazer V, Bonaventura P, Depil S. Human Endogenous Retroviruses (HERVs): shaping the innate immune response in cancers. *Cancers*. 2020;12(3):610.
6. Larouche JD, Trofimov A, Hesnard L, et al. Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Med*. 2020;12(1):1-16.
7. Hervé CA, Lugli EB, Brand A, Griffiths DJ, Venables PJW. Autoantibodies to human endogenous retrovirus-K are frequently detected in health and disease and react with multiple epitopes. *Clin Exp Immunol*. 2002;128(1):75-82.
8. Mameli G, Astone V, Arru G, et al. Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not human herpesvirus 6. *J Gen Virol*. 2007;88(1):264-274.
9. Brudek T, Christensen T, Aagaard L, Petersen T, Hansen HJ, Møller-Larsen A. B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity. *Retrovirology*. 2009;6(1):104.
10. Li W, Lee MH, Henderson L, et al. Human endogenous retrovirus-K contributes to motor neuron disease. 2015;7(307):307ra153.
11. Ma W, Hong Z, Liu H, et al. Human endogenous retroviruses-K (HML-2) expression is correlated with prognosis and Progress of hepatocellular carcinoma. *Biomed Res Int*. 2016;2016:1-9.
12. Strissel PL, Ruebner M, Thiel F, et al. Reactivation of codogenic endogenous retroviral (ERV) envelope genes in human endometrial carcinoma and prestages: emergence of new molecular targets. *Oncotarget*. 2012;3(10):1204-1219.
13. Johanning GL, Malouf GG, Zheng X, et al. Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype. *Sci Rep*. 2017;7:41960.
14. pil S, Roche C, Dussart P, Prin L. Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. *Leukemia*. 2002;16(2):254-259.
15. Burns KH. Transposable elements in cancer. *Nat Rev Cancer*. 2017;17(7):415-424.
16. Attermann AS, Bjerregaard AM, Saini SK, Grønbæk K, Hadrup SR. Human endogenous retroviruses and their implication for immunotherapeutics of cancer. *Ann Oncol*. 2018;29(11):2183-2191.
17. De Kouchkovsky I, Abdul-Hay M. Acute myeloid leukemia: a comprehensive review and 2016 update. *Blood Cancer J*. 2016;6(7):e441.
18. Herold T, Jurinovic V, Batcha AMN, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica*. 2018;103(3):456-465.
19. Onecha E, Ruiz-Heredia Y, Martínez-Cuadrón D, et al. Improving the prediction of acute myeloid leukaemia outcomes by complementing mutational profiling with ex vivo chemosensitivity. *Br J Haematol*. 2020;189(4):672-683.
20. Tazi Y, Arango J, Zhou Y, et al. A unified classification and risk stratification algorithm to support clinical decisions in acute myeloid leukemia. Abstr S133 EHA2021. *Virtual Congr*. 2021.
21. 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.
22. Vadakekolathu J, Minden MD, Hood T, et al. Immune landscapes predict chemotherapy resistance and immunotherapy response in acute myeloid leukemia. *Sci Transl Med*. 2020;12(546):eaaz0463.
23. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-421.
24. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science*. 2015;348(6230):69-74.
25. Smith CC, Selitsky SR, Chai S, Armistead PM, Vincent BG, Serody JS. Alternative tumour-specific antigens. *Nat Rev Cancer*. 2019;19(8):465-478.
26. Ehx G, Larouche JD, Durette C, et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity*. 2021;54(4):737-752.e10.
27. Brodsky I, Foley B, Haines D, Johnston J, Cuddy K, Gillespie D. Expression of HERV-K proviruses in human leukocytes. *Blood*. 1993;81(9):2369-2374.
28. Tobiasson M, Abdulkadir H, Lennartsson A, et al. Comprehensive mapping of the effects of azacitidine on DNA methylation, repressive/permissive histone marks and gene expression in primary cells from patients with MDS and MDS-related disease. *Oncotarget*. 2017;8(17):28812-28825.
29. Kazachenka A, Young GR, Attig J, Kordella C, Lamprianidou E, Zoulia E, et al. EPigenetic therapy of myelodysplastic syndromes connects to cellular differentiation independently of endogenous retroelement derepression. *Genome Med*. 2019;11(1):86.
30. Deniz Ö, Ahmed M, Todd CD, Rio-Machin A, Dawson MA, Branco MR. Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat Commun*. 2020;11(1):3506.
31. Colombo AR, Zubair A, Thiagarajan D, Nuzhdin S, Triche TJ. Ramsingh G. suppression of transposable elements in leukemic stem cells. *Sci Rep*. 2017;7(1):7029.
32. Corces MR, Buenrostro JD, Wu B, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016;48(10):1193-1203.
33. The Cancer Genome Atlas Research Network. Genomic and Epigenomic landscapes of adult De novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
34. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*. 2018;562(7728):526-531.
35. Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R. arcasHLA: high-resolution HLA typing from RNAseq. *Bioinforma Oxf Engl*. 2020;36(1):33-40.

36. Bendall ML, de Mulder M, Iñiguez LP, et al. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput Biol*. 2019;15(9):e1006453.
37. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9(4):357-359.
38. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
39. Anders S, Pyl PT, Huber W. HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169.
40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
41. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*. 2018;18(11):696-705.
42. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-674.
43. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database Hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425.
44. Loeffler-Wirth H, Kreuz M, Hopp L, et al. A modular transcriptome map of mature B cell lymphomas. *Genome Med*. 2019;11(1):27.
45. Hubert M, Gobbini E, Couillault C, et al. IFN-III is selectively produced by cDC1 and predicts good clinical outcome in breast cancer. *Sci Immuno*. 2020;5(46):eaav3942.
46. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*. 2013;14(1):7.
47. Thorsson V, Gibbs DL, Brown SD, et al. The immune landscape of cancer. *Immunity*. 2018;48(4):812-830.e14.
48. Ng SWK, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. 2016;540(7633):433-437.
49. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 2019;35(12):2084-2092.
50. Madeira F, Park Y, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. 2019;47(W1):W636-W641.
51. Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506-D515.
52. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst*. 2020;11(4):418-419.
53. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 2020;48(W1):W449-W454.
54. Wells DK, van Buuren MM, Dang KK, et al. Key parameters of tumor epitope immunogenicity revealed through a Consortium approach improve Neoantigen prediction. *Cell*. 2020;183(3):818-834.e13.
55. Bonaventura P, Alcazer V, Mutez V, et al. Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy. *Sci Adv*. 2022;8(4):eabj3671.
56. Mayer J, Blomberg J, Seal RL. A revised nomenclature for transcribed human endogenous retroviral loci. *Mob DNA*. 2011;2(1):7.
57. Iñiguez LP, de Rougvie M, Stearrett N, et al. Transcriptomic analysis of human endogenous retroviruses in systemic lupus erythematosus. *Proc Natl Acad Sci*. 2019;116(43):21350-21351.
58. Zhang Y, Li T, Preissl S, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet*. 2019;51(9):1380-1388.
59. Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*. 2008;9(5):397-405.
60. Mosquera Orgueira A, Peleteiro Raíndo A, Cid López M, et al. Personalized survival prediction of patients with acute Myeloblastic leukemia using gene expression profiling. *Front Oncol*. 2021;11:1018.
61. Jiang F, Mao Y, Lu B, Zhou G, Wang J. A hypoxia risk signature for the tumor immune microenvironment evaluation and prognosis prediction in acute myeloid leukemia. *Sci Rep*. 2021;11(1):14657.
62. Lai Y, Sheng L, Wang J, Zhou M, OuYang G. A novel 85-gene expression signature predicts unfavorable prognosis in acute myeloid leukemia. *Technol Cancer Res Treat*. 2021;20:15330338211004932.
63. Warnat-Herresthal S, Perrakis K, Taschler B, et al. Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics. *iScience*. 2020;23(1):100780.
64. Estey E, Gale RP. How good are we at predicting the fate of someone with acute myeloid leukaemia? *Leukemia*. 2017;31(6):1255-1258.
65. Hu F, Wang Y, da Wang W, Gale RP, Wu B, Liang Y. Improving prediction accuracy in acute myeloid leukaemia: micro-environment, immune and metabolic models. *Leukemia*. 2021;35(11):3073-3077.
66. Saini SK, Ørskov AD, Bjerregaard AM, et al. Human endogenous retroviruses form a reservoir of T cell targets in hematological cancers. *Nat Commun*. 2020;11(1):5660.
67. Rakoff-Nahoum S, Kuebler PJ, Heymann JJ, et al. Detection of T lymphocytes specific for human endogenous retrovirus K (HERV-K) in patients with seminoma. *AIDS Res Hum Retroviruses*. 2006;22(1):52-56.
68. D'Ippolito E, Wagner KI, Busch DH. Needle in a haystack: the Naïve repertoire as a source of T cell receptors for adoptive therapy with engineered T cells. *Int J Mol Sci*. 2020;21(21):8324.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Alcazer V, Bonaventura P, Tonon L, et al. HERVs characterize normal and leukemia stem cells and represent a source of shared epitopes for cancer immunotherapy. *Am J Hematol*. 2022;97(9):1200-1214. doi:10.1002/ajh.26647