Genes, Chromosomes and Cancer

WILEY

RESEARCH ARTICLE OPEN ACCESS

Discovery of Cis-Regulatory Mechanisms via Non-Coding Mutations in Acute Lymphoblastic Leukemia

Efe Aydın[1] | Eleanor L. Woodward[1] | Gladys Telliam Dushime[1] | Rebeqa Gunnarsson[1] | Henrik Lilljebjörn[1] | Larissa H. Moura-Castro[1] | Thoas Fioretos[1,2,3] | Bertil Johansson[1,2] | Kajsa Paulsson[1] | Minjun Yang[1]

[1]Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, Lund, Sweden | [2]Department of Clinical Genetics, Pathology, and Molecular Diagnostics, Office for Medical Services, Region Skåne, Lund, Sweden | [3]Clinical Genomics Lund, Science for Life Laboratory, Lund University, Lund, Sweden

Correspondence: Kajsa Paulsson (kajsa.paulsson@med.lu.se) | Minjun Yang (minjun.yang@med.lu.se)

Received: 17 December 2024 | **Revised:** 6 February 2025 | **Accepted:** 14 March 2025

Funding: This study was supported by grants from the Swedish Childhood Cancer Fund (PR2020-0033 (M.Y.), TJ2020-0024 (M.Y.), PR2021-0005 (B.J.), and PR2021-0016 (K.P.)); the Crafoord Foundation (20230778 (M.Y.) and 20240747 (M.Y.)); the Swedish Cancer Fund (232 694 Pj (B.J.) and 222 062 Pj (K.P.)); Governmental Funding of Clinical Research within the National Health Service (B.J. and K.P.); and the Swedish Research Council (2020-01164 (B.J.) and 2020-00997 (K.P.)).

Keywords: B-cell precursor acute lymphoblastic leukemia | cis-regulatory elements | leukemogenesis | multi-omics | non-coding mutations

ABSTRACT

The non-coding genome, constituting 98% of human DNA, remains largely unexplored, yet holds potential for identifying new biomarkers and therapeutic targets in acute lymphoblastic leukemia (ALL). In this study, we conducted a systematic analysis of recurrent somatic non-coding single nucleotide variants (SNVs) in pediatric B-cell precursor (BCP) ALL. We leveraged whole genome sequencing (WGS) data from 345 pediatric BCP ALL cases, representing all major genetic subtypes and identified 346 mutational hotspots that harbored somatic SNVs in at least three cases. Through the integration of paired RNA sequencing along with published ChIP-seq and ATAC-seq data, we found 128 non-coding hotspots associated with differentially expressed genes nearby, which were enriched for cis-regulatory elements, demonstrating the effectiveness of multi-omics integration in distinguishing pathogenic mutations from passengers. We identified one mutational hotspot that was associated with increased expression of the leukemia-associated gene *NRAS* in three primary ALLs. Micro-C analysis in the leukemia cell line demonstrated interactions between the hotspot region and *NRAS* regulatory elements. Dual luciferase assays indicated that the mutations disrupted regulatory interactions and CRISPR-mediated deletion of the region significantly upregulated *NRAS*, confirming the hypothesized regulatory link. Altogether, we provide new insights into the functional roles of non-coding mutations in leukemia.

1 | Introduction

B-cell precursor acute lymphoblastic leukemia (BCP ALL) is the most common pediatric cancer [1]. Although overall survival (OS) rates now exceed 90% for several BCP ALL genetic subtypes in contemporary treatment protocols [2], conventional treatment is toxic and accumulating evidence shows long-term health effects in survivors [3]. Furthermore, the prognosis after relapse is still dismal, with survival rates dropping to 50%–60% after the first relapse [4, 5]. Further intensification of chemotherapy is not likely to improve the outcome since it will result in increased frequencies of adverse events [6, 7]. Thus, new drugs that target specific genetic changes need to be developed.

Kajsa Paulsson and Minjun Yang contributed equally.

For years, the coding genome has been extensively analyzed, leading to the discovery of novel subtypes and new putative targets for the treatment of BCP ALL [8]. However, research on the non-coding genome has been limited, and little is known about its role in leukemogenesis. Non-coding regions account for approximately 98% of the human genome and harbor the majority of constitutional and somatic variants [9, 10]. Compared to the protein-coding genome, the functionality of the non-coding regions is driven by more complex and indirect mechanisms, making it harder to differentiate between driver and passenger events. Both somatic and germline variants in non-coding genomic regions can influence gene expression through various mechanisms. These include altering the activities of promoters and enhancers, modifying how transcription factors (TFs) bind within cis-regulatory elements (CREs), and changing the local chromatin structure [10]. For instance, somatic mutations in the *GFI1B* enhancer impair GATA2 binding and lead to reduced *GFI1B* expression in acute myeloid leukemia (AML) [11]. Similarly, risk alleles in *ARID5B* intron 3 reduce *ARID5B* expression in patients with high-hyperdiploid acute lymphoblastic leukemia (ALL) by disrupting RUNX3 binding and consequently diminishing RUNX3-dependent *ARID5B* expression [12]. Additionally, non-coding variants within genomic silencers can disrupt normal gene repression by altering chromatin interactions. These disruptions may result in the loss of regulatory elements that typically suppress gene activity, potentially changing the expression of the target gene and playing a pivotal role in the pathogenesis of cancer [13].

In this study, we systematically analyzed recurrent non-coding somatic variants to explore the cis-regulatory architecture in pediatric BCP ALL. By integrating multi-omics data, we dissected the landscape of non-coding mutations in BCP ALL and demonstrated that non-coding mutation clusters that are found near differentially expressed genes are enriched for CREs, highlighting the regulatory potential of non-coding events. In addition, through these events, we identified a novel non-coding region associated with *NRAS* expression. In summary, we report novel recurrent events in pediatric BCP ALL and provide new insights into the regulation of leukemia-associated genes.

2 | Methods

2.1 | Patient Data

A total of 345 BCP ALL cases from two different cohorts were included (SI Table S1). The Lund cohort consisted of 98 cases, of which 34 have been previously published [14, 15], whereas the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative cohort consisted of 247 cases (https://portal.gdc.cancer.gov/projects; dbGAP accession number phs000464, ALL phase 2 discovery, and expansion cohort). Notably, whereas the former cohort was based on material being available for analysis and thus close to population-based, the TARGET cohort was enriched for high-risk patients (https://gdc.cancer.gov/content/target-all-publications-summary). In total, cases belonged to the following genetic subtypes based on the International Consensus and 5th WHO edition classifications [16]: high hyperdiploidy (HeH, $n=75$; 21.7%), *ETV6::RUNX1* ($n=37$; 10.7%), *BCR::ABL1-like* ($n=31$; 9%), *TCF3::PBX1* ($n=28$; 8.1%), *PAX5* alteration ($n=17$; 4.9%), *BCR::ABL1* ($n=10$; 2.9%), *ETV6::RUNX1*-like ($n=8$; 2.3%), hypodiploidy (<46 chromosomes) ($n=8$; 2.3%), *ZNF384* rearrangement ($n=7$; 2%), *KMT2A* rearrangement ($n=4$; 1.2%), *MEF2D* rearrangement ($n=4$; 1.2%), *DUX4* rearrangement ($n=3$; 0.9%), *HLF* rearrangement ($n=2$; 0.6%), *ZNF384* rearrangement-like ($n=2$; 0.6%), intrachromosomal amplification of chromosome 21 (iAMP21) ($n=2$; 0.6%), and *PAX5* P80R mutation ($n=1$; 0.3%). The remaining 106 cases (30.7%) did not harbor any subtype-defining genetic changes and were hence classified as B-other. Of the 345 BCP ALLs, 174 (50.4%) were males and 171 (49.6%) females. The median age at diagnosis was 5 years (range 0–17 years). Informed consent was obtained according to the Declaration of Helsinki and the study was approved by the Swedish Ethical Review Authority (application no. 2023-01550-01).

2.2 | Whole Genome Sequencing Data Analysis

For the complete genomics data generated by the TARGET program, somatic variants were identified by the TARGET WGS analysis pipeline (https://www.cancer.gov/ccg/research/genome-sequencing/target). The data was further filtered for Somatic Score ≥ 0 and number of unique reads of the mutated allele > 10 [14]. Illumina WGS sequencing libraries for matched diagnostic and remission samples from BCP ALL patients diagnosed and treated at Skåne University Hospital, Sweden, were constructed using the TruSeq Nano DNA sample preparation kit (Illumina, San Diego, CA, USA). Paired-end sequencing (2×150 bp) was performed at ~60× coverage for diagnostic and ~30× for remission samples. The paired-end fastq files were aligned to the human genome build hg19 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/) using BWA [17]. Somatic variants were detected by Mutect, Mutect2, and MuSE [18, 19]. Variants that were present in the outputs of at least two of the programs were kept for further analysis. Single nucleotide variants (SNVs) with variant allele frequencies (VAF) <0.2 were removed to keep only the major clone mutations. VCF files for each individual sample were converted to a combined MAF format file for further analysis. To reduce false-positive findings, hypermutated samples ($n=10$) were removed from the MAF file [20]. Three standard deviations from the median were used as a cut-off. Single base substitution (SBS, https://cancer.sanger.ac.uk/signatures/sbs/) signature analyses were performed using MutationalPatterns [21] and the proportions of COSMIC SBS signatures that have previously been associated with sequencing artifacts or ultraviolet light exposure (SBS7a,b,c,d and SBS43–SBS60) were calculated for each sample [22]. Samples carrying a fraction of these signatures higher than 35% were intended to be discarded, but no such proportion was found for any of the non-hypermutated cases. Variants within coding regions were filtered out. The remaining variants were examined on gnomAD version 3.1.2 for their presence in the reference population and variants with VAF higher than 0.0001 in gnomAD were removed from the analysis [23]. For one cluster, additional filtering was applied since this SNV showed a high frequency in other databases in dbSNP [24] but not in gnomAD. Recurrency was determined using the BEDtools cluster algorithm with the distance argument set to 200 base pairs [25].

Mutational hotspots identified in at least three different samples were defined as recurrent and selected for downstream analysis. These hotspots are represented within their hg38 locations in the study. Transformation of the assembly was carried out using the liftOver function from the rtracklayer package [26].

2.3 | Integration of Multi-Omics Data

The paired-end RNA sequencing (RNA-seq) fastq files from both cohorts were aligned to the human genome build hg19 using the STAR 2-pass mapping pipeline and the reads count of genes was quantified by RSEM [27]. RNA-seq data from the two cohorts was corrected for batch effects using ComBat-seq [28]. A two-sided Mann–Whitney *U* test with a $p < 0.1$ threshold was applied on the RSEM values to identify differentially expressed genes within a 1 Mb distance of each mutational cluster.

ChIP-Seq and ATAC-seq data obtained from primary BCP ALL samples were downloaded from the Blueprint Epigenome Project [29] using the European Nucleotide Archives repository (http://ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/GRCh38/bone_marrow/). Additionally, the ATAC-seq data generated by Barnett et al. were also used [30]. H3K4me1 ($n = 8$), H3K4me3 ($n = 10$), H3K27me3 ($n = 4$), H3K27ac ($n = 10$), and ATAC-seq ($n = 156$) were selected as marks of interest. Histone mark peaks were merged and compared against the mutational cluster regions. A similar approach was followed for the ATAC-seq data. Start and end positions of the histone peaks were extended by 100 bases and compared against the clusters. To investigate if these marks were enriched within hotspots with differentially expressed genes nearby, Fisher's exact test was performed. A similar procedure was carried out to check for enrichment on specific TFs, using clustered ENCODE data from lymphoblast cell lines [9]. Enhancer information for regions of interest was extracted from EnhancerAtlas 2.0 [31]. NIH Roadmap Epigenomics Program's 15-state chromatin modeling data for the GM12878 lymphoblast cell line was also used to annotate the chromatin state of the clusters [32]. Mutational clusters with differentially expressed genes in the vicinity were manually examined using the additional information generated from integrated multi-omics. PERFECTOS-APE was used to predict differential TF binding scores due to hotspot mutations (https://opera.autosome.org/perfectosape).

2.4 | Micro-C Assay and Chromatin Structure Data Mining

Micro-C was done on the BCP ALL cell line REH (positive for the *ETV6*::*RUNX1* fusion). Sequencing libraries were constructed using the Dovetail Micro-C kit following the manufacturer's protocol (Cantata Bio, Scotts Valley, CA, USA). A total of three million cells and 0.25 μL of MNase Enzyme Mix were used. Briefly, in-nucleus fixation with disuccinimidyl glutarate (DSG) and formaldehyde was performed, followed by digestion with micrococcal nuclease. Cells were lysed using sodium dodecyl sulfate (SDS). Chromatin capture beads were used to rescue the cross-linked chromatin fragments. Following proximity ligation and crosslink reversal, the DNA was purified and libraries were constructed with Illumina-compatible adaptors. High-throughput paired-end sequencing was performed at the Center for Translational Genomics, Lund University, using the Illumina NovaSeq 6000 platform, generating 3.2 billion 2 × 150 bp read pairs. Micro-C sequencing data was processed using the 4DN Hi-C data processing pipeline (https://data.4dnucleome.org/resources/data-analysis/hi_c-processing-pipeline). The FAN-C toolkit was used for data analysis and visualization of .cool files [33]. Plots were generated at 5 kb resolution, using ICE-normalized contact matrix as input [34]. In addition, previously published promoter capture Hi-C data for the GM12878 cell line [35] were used to investigate the chromatin interactions between the gene promoters and mutational hotspots.

2.5 | Luciferase Assays

We integrated three mutations from the selected hotspot region into a single luciferase vector with Hotspot-miniCMV-fireflyLuciferase sequence along with its wild type counterpart (SI Table S2). An internal control vector with CMV promoter and renilla luciferase was used for normalization (VectorBuilder, Chicago, IL, USA). Two plasmid vector mixes were made. These included the firefly luciferase vectors combined with renilla luciferase vector at a 20:1 M ratio. Plasmids were mixed with 1×10^6 REH cells resuspended in BTXpress electroporation solution (BTX, Holliston, MA, USA). Transfection was carried out by electroporation using a BTX ECM 830 (BTX, Holliston, MA, USA) at 125 V and a 5 ms pulse length. Electroporated cells were incubated in RPMI media with 20% FBS for 24 h. Luciferase signals were measured, as described by the manufacturer, using the Dual-Glo Luciferase Assay System (Promega, Madison, WI, USA) on a GloMax 96 microplate luminometer (Promega, Madison, WI, USA).

For data analysis, the ratio of firefly/renilla luciferase was recorded for each sample and the Shapiro–Wilk test was performed to assess normality. Each sample was checked for equality of variances. The two-sided Mann–Whitney *U* test was used to assess the difference between wild type and mutated inputs.

2.6 | CRISPR/Cas9 Induced Deletions

Two guide RNAs (#1; TTAAAGCAAGCTGCCAGAGG, #2; TCTGCCTTCACTGTAACTGA) were mixed with cas9 using GFP and RFP tags, respectively, and a 1:1 M ratio to create a Ribonucleoprotein complex (RNP). Electroporation was carried out using the same protocol as Luciferase assays with the addition of an electroporation enhancer. Cells were incubated in 20% FBS RPMI media for 24 h. Guide RNAs, cas9, and enhancer were obtained from IDT (IDT, Coralville, IA, USA).

REH transfected cell replicates were pooled in 1.5 mL Eppendorf, washed with sort media (PBS Ca^{2+} Mg^{2+} 2 mM EDTA 5% FBS) and centrifuged for 5 min at 300 × g 4°C. The cell pellet was resuspended in 300 μL, passed through a 30 μm nylon filter (400 μL additional to wash the filter). DAPI solution was added before acquisition to exclude dead cells. Single positive GFP REH cells and single positive RFP REH cells were used for compensation and to define positive gates. REH transfected live cells (DAPI negative) were sorted using Aria llu (BD, Franklin Lakes, NJ, USA) with purity sort mode; RFP positive cells were sorted

from GFP positive gate selection. GFP+/RFP+ sorted cells were washed and resuspended in REH culture media (RPMI 10% FBS) for further amplification.

Sorted cells were collected after 48 h and qPCR was run directly on cell lysate using Cells-to-C_T 1-step TaqMan Kit (Thermo Fisher, Waltham, MA, USA) and ABI 7500 Real-Time PCR System (Thermo Fisher, Waltham, MA, USA). Primer-probe assays are obtained from IDT (Coralville, IA, USA) with the following assay IDs: Hs.PT.58.1271059 (*NRAS*), Hs.PT.58v.27737538 (*GUSB*). Results were obtained using the relative quantification method with *GUSB* as a reference gene.

To check for the accuracy of deletion, we used full-in primers targeting the putative deletion (Forward: TATGCTTACT TCTGGCGAGGTT, Reverse: TAGAAGGCCAGACTTTAGCT GTG) and a control primer set targeting the *GPR15* region from another study to check for a reference copy number [36]. DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany).

3 | Results

3.1 | Computational Pipeline Identifies 346 Mutational Hotspots in Non-Coding Regions

After filtering for hypermutation on the WGS data of 345 BCP ALL cases, 335 cases were retained (Figure 1A). As a follow-up, the burden of noise associated with COSMIC SBS signatures was controlled for, and none of the remaining cases were found to possess a proportion of those signatures higher than the expected threshold. A total of 258,371 somatic SNVs (median 675, range 15–3,103, SI Table S3) were identified. Intergenic and intronic regions harbored the vast majority, accounting for 44.5% and 43.2% of the SNVs, respectively, while SNVs in the protein-coding parts constituted less than 2% of the total number of SNVs. We identified 346 mutational hotspots involving non-coding SNVs in at least three samples (median 3 cases, range 3–14 cases; SI Table S4). Of these, 151 were found in cases from both the Lund and the TARGET cohorts (Figure 1B, Supplementary Table S4). An analysis of whether hotspots were more common in certain chromosomes showed a median of 11.11 (range 6–19) hotspots per 100 Mb across each chromosome (Figure 1C,D). Using two standard deviation distance from the median as the significance cut-off, we found enrichment of hotspot presence in chromosome 4. On average, five protein-coding genes were found within 1 Mb distance of a mutational hotspot (range 0–86). When examining the presence of these genes in the COSMIC cancer gene census [22], no significant enrichment was found.

We also assessed whether the identified hotspots exhibited subtype specificity using Fisher's exact test ($p < 0.05$). To enhance statistical power, we included only subtypes represented by more than five samples (HeH, *BCR::ABL1*, *BCR::ABL1*-like, *ETV6::RUNX1*, *ETV6::RUNX1*-like, *PAX5* alteration, *TCF3::PBX1*, hypodiploidy, *ZNF384* rearrangement). Of the 346 hotspots analyzed, 68 showed significant enrichment for at least one subtype. Next, we examined the presence of COSMIC genes

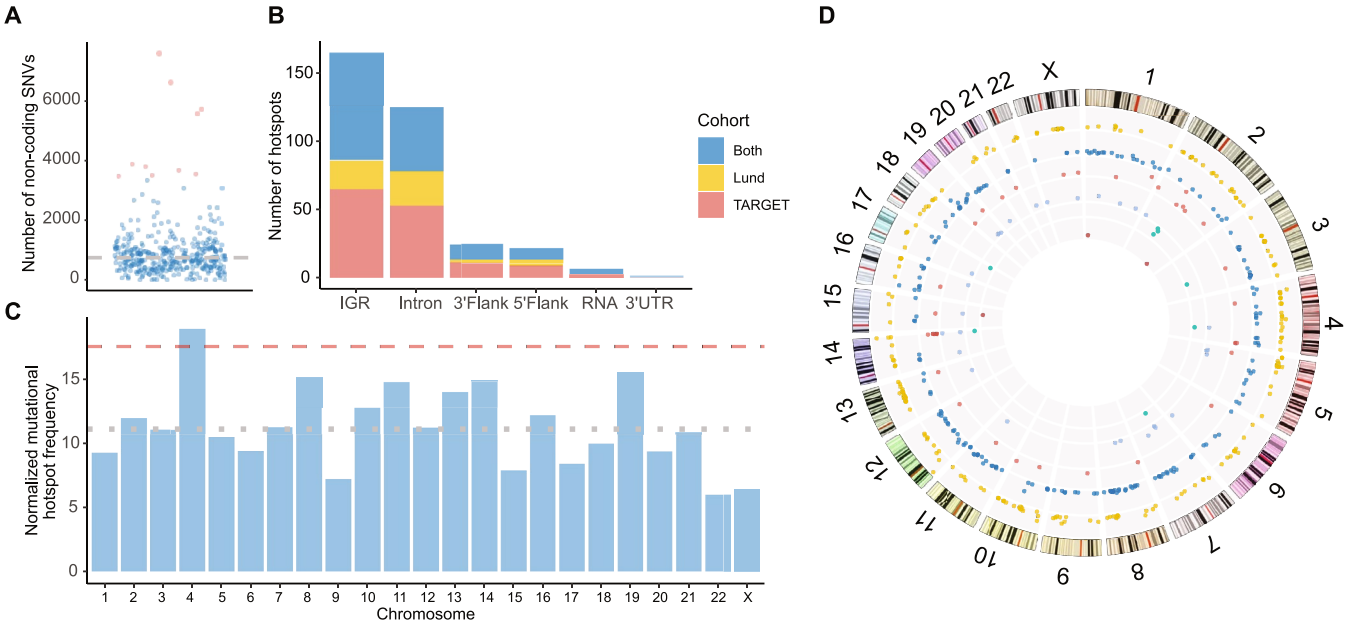

FIGURE 1 | Overview of non-coding somatic single nucleotide variants (SNVs) in 345 cases of pediatric B-cell precursor acute lymphoblastic leukemia (BCP ALL). (A) Variants per sample displayed for the 345 BCP ALL cases included in this study. Gray dashed line represents median number of mutations per sample for the retained cases ($n = 741$). Samples with variant counts higher than three standard deviations from the median were considered hypermutated and removed from further analysis ($n = 10$). These are color-coded in red. (B) Bar plots showing the distribution of detected hotspots based on their variant classifications. As expected, most of the non-coding hotspots are located in intergenic regions (IGR) and introns. (C) Number of hotspots in each chromosome normalized by their lengths in base pairs. Gray dashed line represents the median and red shows the +2 standard deviation threshold. Chromosome 4 was enriched for hotspots. (D) Circos plot displaying the presence of hotspots on each chromosome. Outer to inner circle: Chromosome numbers, chromosome ideograms, scatter plot for non-coding hotspots located within intergenic region, intronic region, 3′Flank, 5′Flank, RNA gene, and 3′UTR. Abbreviation: UTR: untranslated regions.

near these hotspots and found that subtype-specific hotspots were more likely to be associated with COSMIC genes nearby (odds ratio = 1.81, $p = 0.038$, Fisher's exact test).

3.2 | RNA-Seq Integration With Multi-Layered Epigenetic Data Reveals Functional Roles for Non-Coding Hotspots

Matched RNA-seq data were available in 263 cases. After the exclusion of clusters that had none or only one case with RNA-seq data, 301 hotspots remained. Of these, 277 included at least one gene within 1 million base pair distance. Further analysis of the mRNA levels of nearby genes showed differential expression patterns for at least one gene in 128 (46%) hotspots.

To validate the robustness of these findings, we compared the presence of epigenetic marks in hotspots with and without differentially expressing genes nearby, using ChIP-Seq and ATAC-seq data from the primary ALL samples of the BluePrint Epigenome project [29] and in Barnett et al. [30]. Significant enrichments were found for H3K4me1, H3K27ac, and ATAC-seq peaks, indicating cis-regulatory roles for some of these hotspots (Figure 2A). Due to the enrichment of these enhancer-associated marks, we hypothesized that if these regions were indeed enhancers, they should also be enriched for annotated enhancers on the GM12878 cell line, since it represents a lymphoblast, and variability between cell lineage is stronger in the target gene pairing of enhancers but not in enhancers themselves [34]. Therefore, we repeated the previous analysis with the inclusion of annotated GM12878 enhancers from EnhancerAtlas 2.0 [31]. As expected, a strong enrichment ($p = 0.013$) was found. We further evaluated these 128 hotspots to check if they carried specific TF motifs by incorporating available ENCODE [31] data from all available lymphoblast cell lines, but no TFs were found to be enriched (Figure 2A).

3.3 | Non-Coding Hotspot on Chromosome 1 is Associated With Upregulation of *NRAS*

We manually curated the filtered data to prioritize candidates with possible cis-regulatory roles in leukemogenesis for further analysis. First, we selected hotspots near at least one differentially expressed COSMIC gene and excluded those with fewer than three RNA-seq samples to enhance accuracy. Next, we removed single-base-pair hotspots to reduce potential noise or undetected germline variants. Finally, we focused on hotspots

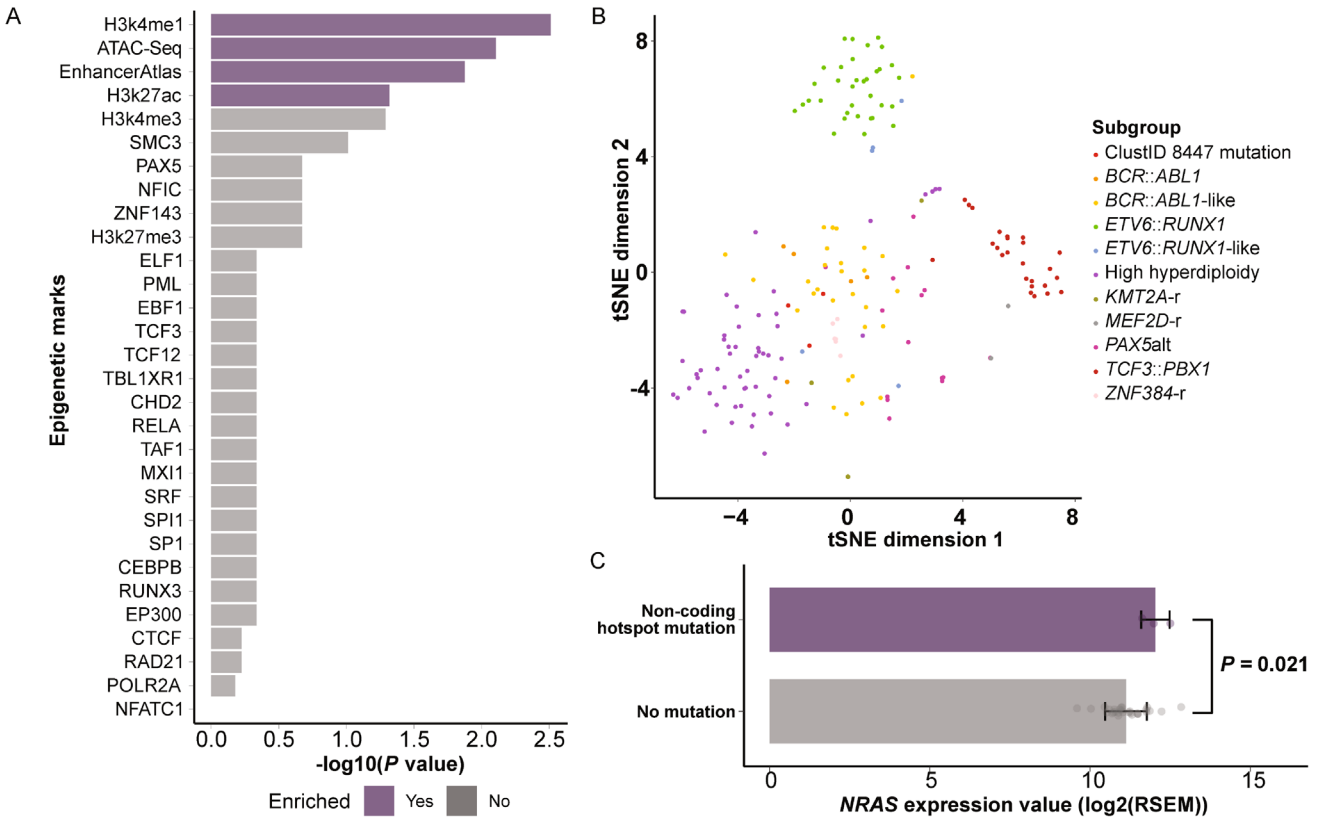

FIGURE 2 | Results of multi-omics integration on hotspots. (A) Enrichment analysis of epigenetic marks. *X* axis −log10(*p* values), and the *Y*-axis lists the names of the epigenetic marks. ATAC-seq and Histone marks represent primary sample data while the rest belongs to GM12878 studies. *p* value threshold of 0.05 were used to determine enrichment. (B) tSNE plot generated by using the top 5000 varying genes in RNA-seq data. Samples with hotspot 8447 mutations show similar expression patterns to the *CRLF2* rearranged subtype. For ease of visualization, only subtypes with sample size greater than three were included. (C) Bar graph showing the expression levels of *NRAS* of samples carrying hotsot 8447 mutations compared against wild type samples from *BCR::ABL1-like* subtype. The *X*-axis represents normalized expression values calculated by the RSEM program. Error bars indicate standard errors of the mean. *p* values are derived from unpaired, two-tailed Mann–Whitney *U* tests.

with subtype-specific enrichment, resulting in the prioritization of four hotspots (ClustIDs: 8447, 66401, 199001, and 223524; Supplementary Table S4). Among them, we further focused on hotspot 8447 due to its close proximity to *NRAS*, a well-established oncogene in ALL.

The hotspot 8447 included three different somatic SNVs occurring in an intronic region of *TSPAN2* and spanning 143 base pairs (G>A chr1:115085508, T>G chr1:115085583, and C>T chr1:115085651, respectively). Two of these samples were classified as *BCR::ABL1*-like and carried *CRLF2* aberrations, while the third case belonged to the B-other group. However, the transcriptional profile of the B-other sample was also similar to the *BCR::ABL1-like* subgroup (Figure 2B). All affected patients were female, aged 7, 11, and 14 at diagnosis. One patient relapsed after 678 days. WGS of the relapse sample confirmed the persistence of the non-coding hotspot mutation, with VAF of 0.46 at diagnosis and 0.36 at relapse, respectively.

To evaluate the expression of genes in the vicinity of this hotspot, differential gene expression analysis was performed and genes within 1 Mb distance of the hotspot were investigated ($n = 16$). This analysis revealed a significant upregulation of *NRAS* ($p = 0.017$, two-sided Mann–Whitney *U* test; Figure 2C), *AMPD1* ($p = 0.041$), and *CSDE1* ($p = 0.032$) in samples carrying these hotspot mutations, compared to *BCR::ABL1* and *BCR::ABL1-like cases*. Considering that *NRAS* is one of the most common driver genes in cancer, we focused our further investigations on this gene. None of the hotspot samples carried coding mutations on the *NRAS* gene.

3.4 | Silencer Role of Hotspot 8447 on *NRAS* Expression

To investigate the presence of regulatory elements within hotspot 8447, dual-luciferase reporter assays were performed with the inclusion of wild type, mutant, and empty vectors. Wild type vector generated a significantly higher reporter expression compared to empty vector ($p = 0.000037$, two-sided Mann–Whitney *U* test, Figure 3A), indicating the presence of regulatory elements within the target sequence. However, the strength

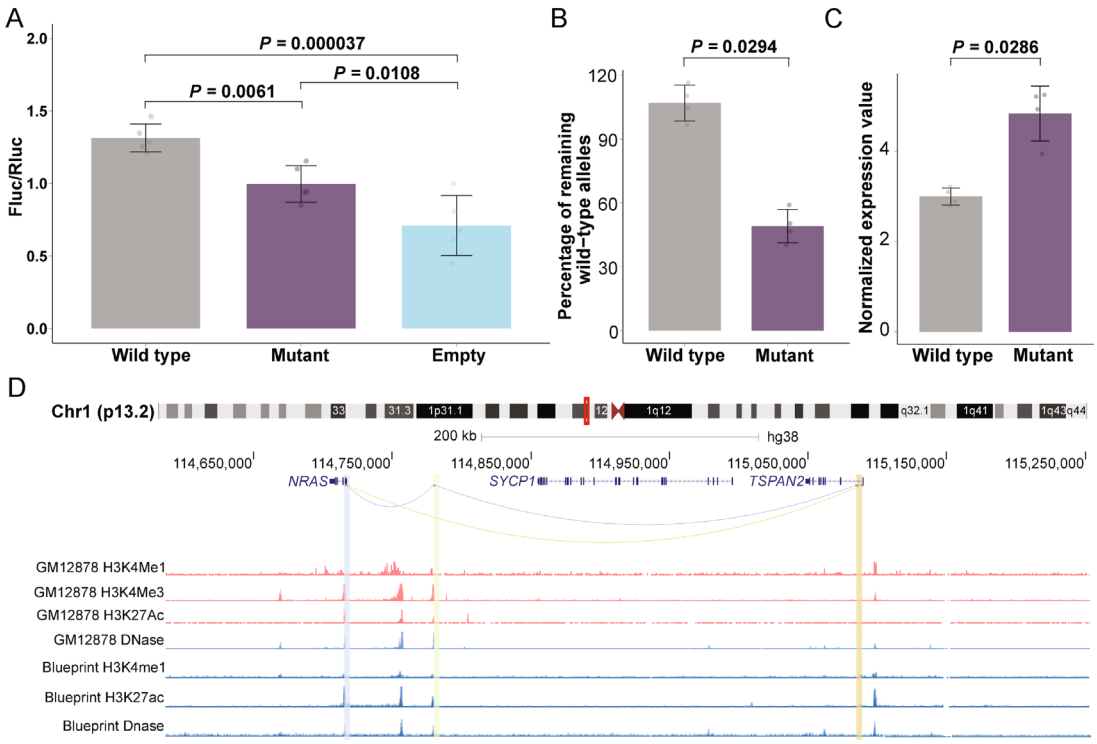

FIGURE 3 | Non-coding mutations affect *NRAS* expression by disrupting normal gene repression. (A) Dual luciferase reporter assays were conducted to assess functionality. Plasmid vectors carried single nucleotide variants seen in the hotspot. The displayed data represents the average of five replicates. *p* values were generated via two-sided Mann–Whitney *U* test. (B, C) Comparison of (B) percentage of remaining wild-type alleles on reference and targeted sites after CRISPR experiment and (C) normalized *NRAS* expression between the wild type Reh cell line and the Reh cell line with deletions of the hotspot 8447 mutation introduced by CRISPR-Cas9 knockout. The displayed data represents the average of four replicates. *p* values were derived from unpaired, two-tailed Mann–Whitney *U* tests performed on relative quantification results. (D) Integration of epigenetic marks and chromatin data at the hotspot 8447 mutation region. The top panel shows the chromatin interactions observed from Micro-C data of the REH cell line and promoter capture Hi-C data of the GM12878 cell line [35], along with the epigenetic marks H3K4Me1, H3K4Me3, H2K427Ac, and ATAC-seq signal from the GM12878 cell line and epigenetic marks H3K4Me1, H2K427Ac, and ATAC-seq signal from ALL patients that are derived from the Blueprint project [29]. Purple curves illustrate the chromatin loops identified using Micro-C data from the REH cell line and the green curve represents the chromatin loop identified through promoter capture Hi-C data from the GM12878 cell line [35]. Regions marked in dark yellow represent hotspots, gray represents the target gene and green represents regions that are both interacting with the hotspot and reported as an enhancer of target gene for GM12878. Error bars indicate standard errors of the mean. Abbreviations: Fluc: firefly luciferase; Rluc: renilla luciferase.

of the regulatory function seemed to be significantly reduced when mutations were introduced ($p = 0.0061$, two-sided Mann–Whitney *U* test, Figure 3A). These findings suggest a disruptive function of the hotspot mutations on regulatory interactions that could be explained by either a weak enhancer or silencer function of the original three-dimensional (3D) chromatin, since loss of function in both cases would lead to an increased expression of the putative target. To further address this, we checked the chromatin state annotation of this region for GM12878, using NIH Roadmap Epigenomics Program's 15 state chromatin model [32]. This region was reported as a weakly repressed polycomb state, indeed suggesting a repressor role for the region.

None of the TF marks integrated in this study were present in the hotspot region. However, due to the limited nature of these experiments—such as the representativeness of lymphoblast cell lines, the number of TFs studied, and potential sensitivity issues of ChIP-seq—we also conducted a computational prediction for differential TF binding using PERFECTOS-APE https://opera.autosome.org/perfectosape for hotspot 8447. This analysis identified 66 motifs with differential binding predictions due to hotspot mutations. Notably, 54 of these (82%) suggested a loss of binding (Supplementary Table S5), indicating a potential disruption of the region's inherent functionality.

To further examine the effect of the hotspot region on the expression of *NRAS*, located approximately 368 kb upstream of the mutational cluster, we created ~300 base pair deletions by targeting both ends of the hotspot region in the REH cell line using the CRISPR/Cas9 assay. After sorting the transfected cells (GFP+/RFP+) and confirming the sgRNA accuracy via qPCR amplification of the targeted region ($p = 0.0294$, two-sided Mann–Whitney *U* test; Figure 3B), we measured *NRAS* expression. The deleted samples showed increased *NRAS* expression ($p = 0.0286$, two-sided Mann–Whitney *U* test; Figure 3C), consistent with a repressor function of the hotspot region.

To evaluate a possible connection between the hotspot 8447 and *NRAS* through a cis-regulatory network, we performed Micro-C in the *ETV6::RUNX1*-positive REH cell line and examined the promoter capture Hi-C data from the GM12878 cell line [35]. We observed chromatin interactions between the hotspot region and both the *NRAS* promoter and an *NRAS* enhancer reported in GM12878 [35], as evidenced by the Micro-C data (Figure 3D, Supplementary Figure S1). In line with this, the promoter capture Hi-C data confirmed chromatin interactions between the *NRAS* promoter and the hotspot 8447 region in the GM12878 cell line, indicating a cis-regulatory role of the hotspot region on *NRAS* expression in healthy lymphoblasts. Taken together, these data indicate that disrupting the hotspot region would lead to a loss of *NRAS* repression, suggesting that the mutated region acts as a silencer of *NRAS* expression.

4 | Discussion

Genomic variants in non-coding regions can affect the expression of tumor-related genes by altering promoter and enhancer activities, modifying TF binding, and changing chromatin interactions. In this study, we explored the non-coding genome of BCP ALL through a systematic examination of recurring non-coding SNVs. We utilized multi-omics data from different studies to distinguish pathogenic mutations from passenger mutations and identified 346 non-coding mutational hotspots, including 128 associated with changes in the expression of nearby genes. The pipeline applied in this study generated a significant enrichment of histone modifications H3K4me1 and H3K27ac at hotspots affecting gene expression, suggesting the presence of cis-regulatory mechanisms involved in non-coding mutational hotspots in BCP ALL.

Our examination of genes close to identified hotpots showed no enrichment of known COSMIC genes around hotspots, indicating most of these variants were passenger mutations. The scarcity of non-coding drivers compared to their protein coding counterparts is in line with the results from a pan-cancer WGS analysis [37]. An explanation for this could be related to the potentially fine-tuning impact on gene expression for non-coding variants compared with coding variants, which may change protein structure or affect mRNA levels. Another potential explanation could be the diverse ways in which a mutational hotspot is defined, influenced by various factors such as the appropriate distance of adjacent genes, the size of the base pair window used to identify recurrent mutations [38], and the frequency threshold for a mutation to be considered part of a hotspot.

Yet, despite adjusting for these parameters, the core one-dimensional approach remains unchanged, since a hotspot is traditionally defined by mutations occurring on a selected window of a DNA sequence. We believe this approach may be inadequate for capturing the complexity of 3D chromatin interactions, which involve multiple TFs and chromatin remodeling elements interacting with cis-regulatory sequences. For instance, mutations occurring at distant locations in the DNA sequence may appear unrelated from a one-dimensional perspective, but if they participate in a regulatory loop, they could be integral to the same functional network. Thus, these hypothetical mutations could arguably be considered as members of the same hotspot from a 3D perspective. This introduces the potential for incorporating 3D-based mechanisms in gene dysregulation, deviating from the conventional model. We believe further studies leveraging the high-resolution chromatin capture technologies will prove important to address this issue.

The RAS family genes are among the most frequently mutated in human cancers. In pediatric BCP ALL, approximately 45% of cases exhibit coding mutations in RAS pathway genes, and these mutations are linked to relapse and drug resistance [39, 40]. Furthermore, recent studies also indicate that the association between RAS genes and leukemia may not be limited to mutation status. Koschut et al. demonstrated that inhibition of wild type RAS expression was sufficient to reduce cell growth of *BCR::ABL1-like* ALL and that alternative mechanisms in the absence of established mutations can induce RAS activation [41]. Similarly, RAS overexpression has also been found to be associated with adverse prognosis in AML, independent of mutation status [42].

Here, we identified a non-coding mutational hotspot associated with increased *NRAS* expression in BCP ALL cases without coding *NRAS* mutations. By Micro-C, we found that this hotspot region could interact with a known *NRAS* enhancer and the

NRAS promoter, and that CRISPR-induced deletions of it led to upregulation of the gene. Our data thus demonstrate a novel CRE, acting as a silencer, located 305 kb from the *NRAS* gene, that regulates *NRAS* expression in BCP ALL. Notably, this CRE could not be identified by integrated epigenetic markers from healthy lymphoblasts or unmatched primary samples. This demonstrates the limitations of common annotation approaches and highlights the leukemogenic potential of undiscovered non-coding regions.

Unraveling the regulatory roles of non-coding mutations may have significant potential for clinical use in the near future. As new targeted therapies emerge, understanding how CRE alterations affect gene expression could help predict patient responses to inhibitors. These alterations may reveal alternative therapeutic targets for otherwise undruggable oncogenic drivers. Incorporating non-coding mutations into risk stratification models could improve prognostic accuracy and guide personalized treatment decisions.

In conclusion, this study sheds light on the regulatory functions of the non-coding genome in BCP ALL and provides new insights into *NRAS* regulation. Our findings underline the importance of multi-omics integration in the evaluation of pathogenic non-coding variants. We demonstrate that previously unknown regulatory regions within the transcriptional network of leukemia-associated genes can be identified through recurrent non-coding variation patterns and emphasize the necessity of including non-coding somatic variants in genomic studies.

Author Contributions

K.P. and M.Y. planned and supervised the study. E.A., E.L.W., G.T.D., R.G., L.H.M.-C., K.P., and M.Y. performed experiments and analyzed data. H.L., T.F., and B.J. provided clinical data and samples and analyzed data. E.A., K.P., and M.Y. wrote the article with input from all authors.

Acknowledgments

This study was supported by grants from the Swedish Childhood Cancer Fund, grant numbers PR2020-0033 (MY), TJ2020-0024 (MY), PR2021-0005 (BJ), and PR2021-0016 (KP); the Crafoord Foundation 20230778 (MY) and 20240747 (MY); the Swedish Cancer Fund, grant numbers 23 2694 Pj (BJ) and 22 2062 Pj (KP); Governmental Funding of Clinical Research within the National Health Service (BJ and KP); and the Swedish Research Council, grant numbers 2020-01164 (BJ) and 2020-00997 (KP). The authors would like to acknowledge Clinical Genomics Lund, SciLifeLab, and Center for Translational Genomics (CTG), Lund University, for providing expertise and service with sequencing and analysis.

Ethics Statement

Informed consent has been obtained from patients and/or their parents/legal guardians according to the Declaration of Helsinki and the study has been approved by the Swedish Ethical Review Authority (application no. 2023-01550-01). Access to patient files and biobanked samples was approved by Region Skåne, Sweden (KVB application no. 317-23).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The results published here are in part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) Initiative (phs000464). The TARGET data used for this analysis is available at https://portal.gdc.cancer.gov/projects. Information about TARGET can be found at https://ocg.cancer.gov/programs/target. WGS data from the Lund cohort has been deposited to European Genome Archive (EGA, https://ega-archive.org/) under accession number EGAD00001010103.

References

1. S. W. Brady, K. G. Roberts, Z. Gu, et al., "The Genomic Landscape of Pediatric Acute Lymphoblastic Leukemia," *Nature Genetics* 54, no. 9 (2022): 1376–1389, https://doi.org/10.1038/s41588-022-01159-z.

2. U. Norén-Nyström, M. K. Andersen, G. Barbany, et al., "Genetic Subtypes and Outcome of Patients Aged 1 to 45 Years Old With Acute Lymphoblastic Leukemia in the NOPHO ALL2008 Trial," *Hemasphere* 7, no. 5 (2023): e883, https://doi.org/10.1097/HS9.0000000000000883.

3. D. A. Mulrooney, G. Hyun, K. K. Ness, et al., "The Changing Burden of Long-Term Health Outcomes in Survivors of Childhood Acute Lymphoblastic Leukaemia: A Retrospective Analysis of the St Jude Lifetime Cohort Study," *Lancet Haematology* 6, no. 6 (2019): e306–e316, https://doi.org/10.1016/S2352-3026(19)30050-X.

4. S. P. Hunger and E. A. Raetz, "How I Treat Relapsed Acute Lymphoblastic Leukemia in the Pediatric Population," *Blood* 136, no. 16 (2020): 1803–1812, https://doi.org/10.1182/blood.2019004043.

5. T. Oskarsson, S. Söderhall, J. Arvidson, et al., "Relapsed Childhood Acute Lymphoblastic Leukemia in the Nordic Countries: Prognostic Factors, Treatment and Outcome," *Haematologica* 101, no. 1 (2016): 68–76, https://doi.org/10.3324/haematol.2015.131680.

6. B. Lund, A. Åsberg, M. Heyman, et al., "Risk Factors for Treatment Related Mortality in Childhood Acute Lymphoblastic Leukaemia," *Pediatric Blood & Cancer* 56, no. 4 (2011): 551–559, https://doi.org/10.1002/pbc.22719.

7. K. Schmiegelow, M. F. Levinsen, A. Attarbaschi, et al., "Second Malignant Neoplasms After Treatment of Childhood Acute Lymphoblastic Leukemia," *Journal of Clinical Oncology* 31, no. 19 (2013): 2469–2476, https://doi.org/10.1200/JCO.2012.47.0500.

8. C. H. Pui, "Precision Medicine in Acute Lymphoblastic Leukemia," *Frontiers in Medicine* 14, no. 6 (2020): 689–700, https://doi.org/10.1007/s11684-020-0759-8.

9. ENCODE Project Consortium, "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature* 489, no. 7414 (2012): 57–74, https://doi.org/10.1038/nature11247.

10. E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein, "Role of Non-Coding Sequence Variants in Cancer," *Nature Reviews Genetics* 17, no. 2 (2016): 93–108, https://doi.org/10.1038/nrg.2015.17.

11. B. He, P. Gao, Y. Y. Ding, et al., "Diverse Noncoding Mutations Contribute to Deregulation of cis-Regulatory Landscape in Pediatric Cancers," *Science Advances* 6, no. 30 (2020): eaba3064, https://doi.org/10.1126/sciadv.aba3064.

12. J. B. Studd, J. Vijayakrishnan, M. Yang, G. Migliorini, K. Paulsson, and R. S. Houlston, "Genetic and Regulatory Mechanism of Susceptibility to High-Hyperdiploid Acute Lymphoblastic Leukaemia at 10p21.2," *Nature Communications* 8, no. 1 (2017): 14616, https://doi.org/10.1038/ncomms14616.

13. Y. Cai, Y. Zhang, Y. P. Loh, et al., "H3K27me3-Rich Genomic Regions can Function as Silencers to Repress Gene Expression via Chromatin Interactions," *Nature Communications* 12, no. 1 (2021): 719, https://doi.org/10.1038/s41467-021-20940-y.

14. K. Paulsson, H. Lilljebjörn, A. Biloglav, et al., "The Genomic Landscape of High Hyperdiploid Childhood Acute Lymphoblastic Leukemia," *Nature Genetics* 47, no. 6 (2015): 672–676, https://doi.org/10.1038/ng.3301.

15. E. L. Woodward, M. Yang, L. H. Moura-Castro, et al., "Clonal Origin and Development of High Hyperdiploidy in Childhood Acute Lymphoblastic Leukaemia," *Nature Communications* 14, no. 1 (2023): 1658, https://doi.org/10.1038/s41467-023-37356-5.

16. A. S. Duffield, C. G. Mullighan, and M. J. Borowitz, "International Consensus Classification of Acute Lymphoblastic Leukemia/Lymphoma," *Virchows Archiv* 482, no. 1 (2023): 11–26, https://doi.org/10.1007/s00428-022-03448-8.

17. H. Li and R. Durbin, "Fast and Accurate Short Read Alignment With Burrows-Wheeler Transform," *Bioinformatics* 25, no. 14 (2009): 1754–1760, https://doi.org/10.1093/bioinformatics/btp324.

18. K. Cibulskis, M. S. Lawrence, S. L. Carter, et al., "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples," *Nature Biotechnology* 31, no. 3 (2013): 213–219, https://doi.org/10.1038/nbt.2514.

19. Y. Fan, L. Xi, D. S. Hughes, et al., "MuSE: Accounting for Tumor Heterogeneity Using a Sample-Specific Error Model Improves Sensitivity and Specificity in Mutation Calling From Sequencing Data," *Genome Biology* 17, no. 1 (2016): 178, https://doi.org/10.1186/s13059-016-1029-6.

20. M. H. Bailey, C. Tokheim, E. Porta-Pardo, et al., "Comprehensive Characterization of Cancer Driver Genes and Mutations," *Cell* 173, no. 2 (2018): 371–385, https://doi.org/10.1016/j.cell.2018.02.060.

21. F. Blokzijl, R. Janssen, R. van Boxtel, and E. Cuppen, "MutationalPatterns: Comprehensive Genome-Wide Analysis of Mutational Processes," *Genome Medicine* 10, no. 1 (2018): 33, https://doi.org/10.1186/s13073-018-0539-0.

22. J. G. Tate, S. Bamford, H. C. Jubb, et al., "COSMIC: The Catalogue of Somatic Mutations in Cancer," *Nucleic Acids Research* 47, no. D1 (2019): D941–D947, https://doi.org/10.1093/nar/gky1015.

23. K. J. Karczewski, L. C. Francioli, G. Tiao, et al., "The Mutational Constraint Spectrum Quantified From Variation in 141,456 Humans," *Nature* 581, no. 7809 (2020): 434–443, https://doi.org/10.1038/s41586-020-2308-7.

24. S. T. Sherry, M. H. Ward, M. Kholodov, et al., "dbSNP: The NCBI Database of Genetic Variation," *Nucleic Acids Research* 29, no. 1 (2001): 308–311, https://doi.org/10.1093/nar/29.1.308.

25. A. R. Quinlan and I. M. Hall, "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features," *Bioinformatics* 26, no. 6 (2010): 841–842, https://doi.org/10.1093/bioinformatics/btq033.

26. M. Lawrence, R. Gentleman, and V. Carey, "Rtracklayer: An R Package for Interfacing With Genome Browsers," *Bioinformatics* 25, no. 14 (2009): 1841–1842, https://doi.org/10.1093/bioinformatics/btp328.

27. B. Li and C. N. Dewey, "RSEM: Accurate Transcript Quantification From RNA-Seq Data With or Without a Reference Genome," *BMC Bioinformatics* 12 (2011): 323, https://doi.org/10.1186/1471-2105-12-323.

28. Y. Zhang, G. Parmigiani, and W. E. Johnson, "ComBat-Seq: Batch Effect Adjustment for RNA-Seq Count Data," *NAR Genomics and Bioinformatics* 2, no. 3 (2020): lqaa078, https://doi.org/10.1093/nargab/lqaa078.

29. J. H. Martens and H. G. Stunnenberg, "BLUEPRINT: Mapping Human Blood Cell Epigenomes," *Haematologica* 98, no. 10 (2013): 1487–1489, https://doi.org/10.3324/haematol.2013.094243.

30. K. R. Barnett, R. J. Mobley, J. D. Diedrich, et al., "Epigenomic Mapping Reveals Distinct B Cell Acute Lymphoblastic Leukemia Chromatin Architectures and Regulators," *Cell Genomics* 3, no. 12 (2023): 100442, https://doi.org/10.1016/j.xgen.2023.100442.

31. T. Gao and J. Qian, "EnhancerAtlas 2.0: An Updated Resource With Enhancer Annotation in 586 Tissue/Cell Types Across Nine Species," *Nucleic Acids Research* 48, no. D1 (2020): D58–D64, https://doi.org/10.1093/nar/gkz980.

32. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, et al., "Integrative Analysis of 111 Reference Human Epigenomes," *Nature* 518, no. 7539 (2015): 317–330, https://doi.org/10.1038/nature14248.

33. K. Kruse, C. B. Hug, and J. M. Vaquerizas, "FAN-C: A Feature-Rich Framework for the Analysis and Visualisation of Chromosome Conformation Capture Data," *Genome Biology* 21, no. 1 (2020): 303, https://doi.org/10.1186/s13059-020-02215-9.

34. M. Imakaev, G. Fudenberg, R. P. McCord, et al., "Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization," *Nature Methods* 9, no. 10 (2012): 999–1003, https://doi.org/10.1038/nmeth.2148.

35. B. Mifsud, F. Tavares-Cadete, A. N. Young, et al., "Mapping Long-Range Promoter Contacts in Human Cells With High-Resolution Capture Hi-C," *Nature Genetics* 47, no. 6 (2015): 598–606, https://doi.org/10.1038/ng.3286.

36. J. Hoebeeck, R. van der Luijt, B. Poppe, et al., "Rapid Detection of VHL Exon Deletions Using Real-Time Quantitative PCR," *Laboratory Investigation* 85, no. 1 (2005): 24–33, https://doi.org/10.1038/labinvest.3700209.

37. E. Rheinbay, M. M. Nielsen, F. Abascal, et al., "Analyses of Non-Coding Somatic Drivers in 2,658 Cancer Whole Genomes," *Nature* 578, no. 7793 (2020): 102–111, https://doi.org/10.1038/s41586-020-1965-x.

38. S. W. Piraino and S. J. Furney, "Identification of Coding and Non-Coding Mutational Hotspots in Cancer Genomes," *BMC Genomics* 18, no. 1 (2017): 17, https://doi.org/10.1186/s12864-016-3420-9.

39. I. S. Jerchel, A. Q. Hoogkamer, I. M. Ariës, et al., "RAS Pathway Mutations as a Predictive Biomarker for Treatment Adaptation in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia," *Leukemia* 32, no. 4 (2018): 931–940, https://doi.org/10.1038/leu.2017.303.

40. M. Messina, S. Chiaretti, J. Wang, et al., "Prognostic and Therapeutic Role of Targetable Lesions in B-Lineage Acute Lymphoblastic Leukemia Without Recurrent Fusion Genes," *Oncotarget* 7, no. 12 (2016): 13886–133901, https://doi.org/10.18632/oncotarget.7356.

41. D. Koschut, D. Ray, Z. Li, et al., "RAS-Protein Activation but not Mutation Status is an Outcome Predictor and Unifying Therapeutic Target for High-Risk Acute Lymphoblastic Leukemia," *Oncogene* 40, no. 4 (2021): 746–762, https://doi.org/10.1038/s41388-020-01567-7.

42. J. D. Zhou, D. M. Yao, X. X. Li, et al., "KRAS Overexpression Independent of RAS Mutations Confers an Adverse Prognosis in Cytogenetically Normal Acute Myeloid Leukemia," *Oncotarget* 8, no. 39 (2017): 66087–66097, https://doi.org/10.18632/oncotarget.19798.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.