Single-cell Rapid Capture Hybridization sequencing to reliably detect isoform usage and coding mutations in targeted genes

Hongke Peng, Jafar S. Jabbari, Luyi Tian, et al.

Genome Res. published online January 10, 2025
Access the most recent version at doi:10.1101/gr.279322.124

Single-cell Rapid Capture Hybridization sequencing reliably detects isoform usage and coding mutations in targeted genes

Authors: Hongke Peng[1,2], Jafar S. Jabbari [1,2], Luyi Tian[1,2,8], Changqing Wang[1,2], Yupei You[1,2], Chong Chyn Chua[1,2,3,4], Natasha S. Anstee[1,2], Noorul Amin[1,2], Andrew H. Wei[1,2,5], Nadia M. Davidson[1,2], Andrew W. Roberts[1,2,5], David C. S. Huang[1,2], Matthew E. Ritchie[1,2,*], Rachel Thijssen[1,2,6,7,*]

[1]*The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia*

[2]*Department of Medical Biology, University of Melbourne, Melbourne, Australia*

[3]*Monash Haematology, Monash Health, Melbourne Australia*

[4]*Clinical Haematology, Northern Health, Melbourne, Australia*

[5]*Department of Clinical Haematology, Royal Melbourne Hospital and Peter MacCallum Cancer Centre, Melbourne, Australia*

[6]*Department of Hematology, Amsterdam UMC, Amsterdam, the Netherlands.*

[7]*Cancer Center Amsterdam, Cancer Biology & Immunology, Amsterdam, the Netherlands.*

[8]*Current Address: Guangzhou Laboratory, Guangdong, China*

**Joint senior author*

Correspondence to R. Thijssen; r.thijssen@amsterdamumc.nl at the Department of Hematology, Amsterdam UMC, De Boelelaan 1117, Amsterdam, The Netherlands.

Running title: scRaCH-seq to detect isoforms and mutations

Abstract

Single-cell long-read sequencing has transformed our understanding of isoform usage and the mutation heterogeneity between cells. Despite unbiased in-depth analysis, the low sequencing throughput often results in insufficient read coverage thereby limiting our ability to perform mutation calling for specific genes. Here, we developed a **s**ingle-**c**ell **Ra**pid **C**apture **H**ybridization **seq**uencing (scRaCH-seq) method that demonstrated high specificity and efficiency in capturing targeted transcripts using long-read sequencing, allowing an in-depth analysis of mutation status and transcript usage for genes of interest. The method includes creating a probe panel for transcript capture, using barcoded primers for pooling and efficient sequencing via Oxford Nanopore Technologies platforms. scRaCH-seq is applicable to stored and indexed single-cell cDNA which allows analysis to be combined with existing short-read RNA-seq datasets. In our investigation of *BTK* and *SF3B1* genes in samples from patients with chronic lymphocytic leukaemia (CLL), we detected *SF3B1* isoforms and mutations with high sensitivity. Integration with short-read scRNA-seq data revealed significant gene expression differences in *SF3B1*-mutated CLL cells, though it did not impact the sensitivity of the anti-cancer drug venetoclax. scRaCH-seq's capability to study long-read transcripts of multiple genes makes it a powerful tool for single-cell genomics.

Introduction

Single-cell sequencing technologies have revolutionized our understanding of cell state and function (Tang et al. 2009; Sandberg 2014). These approaches enable the comprehensive characterization of individual cells, revealing cancer biology and tumour heterogeneity (Stewart et al. 2020; Wang et al. 2020; Liu et al. 2022; Mustachio and Roszik 2022; Tian et al. 2022; Wang et al. 2022; Nagler and Wu 2023). While single-cell RNA sequencing (scRNA-seq) has been widely used for transcriptomic profiling of individual cells, it has limitations in calling mutations and quantifying isoform usage due to the 5' and 3' bias induced by the fragmentation step in the library preparation protocol and the widespread use of short-read technology for sequencing samples. Consequently, linking the single-cell transcriptome to the mutation status of cancer cells becomes a challenge. Understanding the clonal and non-clonal mechanisms of selection and adaptation in response to therapeutic pressure is of importance. Moreover, the influence of isoform usage on cell states further underscores the need for comprehensive methodologies. In response to these challenges, alternative methods have emerged to link the transcriptome profile to isoform usage, mutations, and translocations in full-length transcripts at a single-cell level (Wu and Schmitz 2023).

In recent years, several high-throughput methods have been developed to enable single-cell long-read sequencing. These approaches typically involve barcoding of cDNA using existing methods such as 10x Genomics or Drop-seq and sequencing the indexed full-length cDNA on platforms such as Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) (Lebrigand et al. 2020; Joglekar et al. 2023). While these unbiased single-cell long-read sequencing methods can detect a larger number of isoforms at a single-cell level, the lower overall sequencing level per cell reduces the ability to accurately quantify isoform usage and mutation calling. To overcome this challenge, other methods were developed to target genes of interest, but these methods often require primer spike-in with the 10x Genomics protocol and primer panel design optimization (Nam et al. 2019; Griffin et al. 2023). This limits the ability to target multiple genes of interest or to perform long-read sequencing on already amplified single-cell indexed cDNA.

The growing interest in high throughput single-cell multi-omic methods to enhance our understanding of cancer complexity led us to develop a method that can flexibly link genotypes in single cells to pre-existing short-read transcriptome data generated by various 10x Genomics protocols. CITE-seq combines transcriptome with cell surface protein expression and 10x Genomics multiome combines transcriptome with chromatin accessibility data (Stoeckius et al. 2017). Our single-cell Rapid Capture Hybridization sequencing (scRaCH-seq) method enables the capture of multiple transcripts from pre-indexed and stored cDNA independent of the 10x Genomics kit used. This approach uses biotinylated probes to target genes of interest and Streptavidin beads for on-target transcript enrichment. It was first detailed in Thijssen *et al.* (Blood 2022), to capture the transcriptional landscape of 17 BCL2 family genes in patients with chronic lymphocytic leukaemia (CLL) (Thijssen et al. 2022). That study investigated the resistance mechanisms observed in chronic lymphocytic leukaemia cells after treatment with the BCL2 inhibitor venetoclax treatment. Using scRaCH-seq, we successfully identified the *BCL2 G101V* mutation, a novel *PMAIP1* transcript and a unique *BAX* isoform across different venetoclax-relapsed samples at a single-cell level. Despite the first application of scRaCH-seq to that research, a detailed method description, data analysis and method evaluation has not been published.

Here, we provide a comprehensive description of the scRaCH-seq method in which we target the transcripts of 2 genes *SF3B1* and *BTK* as illustrative examples. The aim of this study is to demonstrate the utility and accuracy of scRaCH-seq in detecting gene-specific mutations and splicing events in single cells, with a focus on probe capture efficiency, artefact diagnostics and mutation/deletion calling.

Results

Assessing data quality and probe-capture efficiency

A step-by-step guide to probe panel design (**Supplemental Fig S1**) was created to facilitate the application of scRaCH-seq together with a tailored analysis pipeline that extends the *FLAMES* (**F**ull-**L**ength **A**nalysis of **M**utations and **S**plicing) R package (Tian et al. 2021) and incorporates additional read integrity checks and steps to remove background noise. After scRaCH-seq was performed (**Figure 1**), a total of 43,762,097 raw reads were acquired through Nanopore sequencing across 21 samples, with a read length distribution spanning from 0 to 4,000bp (**Figure 2A**). The *FLAMES* demultiplexing process was subsequently able to recover 28,573,271 reads that could be attributed to specific cells. Despite an overall 35% read loss during this step, the length distribution of the retained reads remained unaltered (**Figure 2A**). After demultiplexing, sufficient reads (~1,000,000) were preserved for each sample (**Figure 2B**). We also tested Flexiplex (Cheng et al. 2024) and the reads retained for downstream processing were similar. Next demultiplexed reads without TSO sequences and poly(A) tails were discarded following the quality control step which eliminated around 49% of reads, primarily those with lengths < 1500bp (**Figure 2A**). A total of 14,442,490 reads were retained for subsequent read alignment and feature counting. Despite some read loss during demultiplexing and read integrity checks, we achieved an average saturation rate exceeding 75% for the target genes across all samples (**Supplemental Fig S2A-B**). Additionally, there was an increase in UMIs as more transcripts were sequenced for each sample (**Supplemental Fig S2C-D**). scRaCH-seq exhibited lower read loss during both demultiplexing and read integrity checks compared to single-cell full-length transcript sequencing (scFLT-seq), an unbiased single-cell long-read sequencing method (**Supplemental Fig S2E-G**).

scRaCH-seq and our *FLAMES* pipeline effectively captured the majority of cell barcodes identified by *CellRanger* in scRNA-seq data (**Figure 2B**). After quality control, 92.6% of all the reads retained aligned successfully to the targeted genes *SF3B1* and *BTK* (**Figure 2B, Supplemental Fig S3A**). Duplicated reads were collapsed based on isoforms. In concordance with on-target reads, a high

number of UMIs were associated with *SF3B1* and *BTK*, with the majority of captured cells harbouring the target genes (**Figure 2C**). Most of the off-target UMIs arose from single reads (**Supplemental Fig S3B**), indicating the presence of low levels of background signal in the data. The captured off-target genes include *HSPD1*, which is near *SF3B1*, and *TIMM8A*, located in close proximity to the *BTK* gene (**Figure 2D**). These were likely read-through transcripts. Additionally, some other top off-target transcripts, such as *RPL13*, were captured due to shared sequences with the target transcripts (**Supplemental Fig S3C**). However, another group of off-target transcripts, such as *CD74*, were neither located near the target genes nor shared similar sequences with them. This type of off-target transcript is likely attributable to sequencing background. Per-cell UMI counts for the target genes in the corresponding short-read data were evaluated, revealing a higher per-cell UMI capture in the scRaCH-seq data for *SF3B1* and *BTK* (**Figure 2E**). In a separate scRaCH-seq experiment targeting 17 BCL2 family genes in CLL cells, similar UMI counts were observed in scRaCH-seq data compared to short-read data (**Supplemental Fig S4A**). In another experiment using acute myeloid leukaemia (AML) cells with similar read length distribution as the CLL samples (**Figure 2A**, **Supplemental Fig S4B**) we successfully captured the transcripts of 24 genes (**Supplemental Fig S4C**). The AML sample was processed using the 10x Genomics Chromium single-cell 3' kit. These findings indicate the high efficiency of scRaCH-seq in capturing the transcripts of target genes observed in matched single-cell 5' and 3' short-read datasets via long-read sequencing.

Next, we compared the gene expression of *SF3B1* and *BTK* in the scRaCH-seq data with the short-read dataset. Similar to the short-read gene expression data, scRaCH-seq showed that *BTK* expression was detected in the CLL/B cell population, whereas *SF3B1* was expressed across all cell types (**Figure 2F-H**)

Isoform detection by scRaCH-seq

Given that scRaCH-seq operates by capturing transcripts of target genes with probes, it enables the exploration of isoform usage for the targeted genes. In our dataset, we incorporated the CLL and B cells from matched samples from CLL patients at diagnosis, after venetoclax relapse, in patients who

relapsed and subsequently received a BTK inhibitor (BTKi) and healthy donor samples (Thijssen et al. 2022). The primary *BTK* isoform (*BTK-201*, ENST00000308731) captured by scRaCH-seq is a protein-coding isoform with 19 exons, exhibiting high expression across all samples (**Figure 3A**). The other four transcripts among the top 5 captured isoforms by scRaCH-seq are novel isoforms not catalogued previously (**Figure 3A**). Transcripts #2 and #3 lack the last four exons compared to *BTK-201*, with transcript #2 having a 16-base pair longer exon 10 compared to transcript #3. Novel transcript #4 lacks exon 10 compared to the *BTK*-201 transcripts. While all five *BTK* transcripts detected by scRaCH-seq have a poly(A) tail in the 3' UTR, transcript #5, comprising only three exons, has a different starting point at the 5' UTR. Consequently, it remains uncertain whether *BTK* transcript #5 is an artefact resulting from truncation at the 5' end. No evidence of differential *BTK* transcript usage was found across the different sample groups (**Figure 3A**). Among the *BTK* transcripts, four of them shared a nearby TSS with the primary *BTK* isoform, except for *BTK* transcript #5 (**Figure 3A**). However, *BTK* transcript #5 also shared a nearby TSS with another annotated isoform (**Supplemental Fig S5A**). Additionally, TSS signals were detected in this region in the CAGE database (FUNTOM6) (**Supplemental Fig S5B**).

In our investigation of the *SF3B1* isoforms, a previously unidentified *SF3B1* transcript emerged as the predominant form, characterized by the absence of the last 14 exons in the 3' end (**Figure 3B).** Given that this transcript is not the canonical *SF3B1* transcript expressed in the majority of cells, including healthy donor samples, we undertook a comprehensive assessment to determine its authenticity. This transcript does not appear to be an artefact induced by Nanopore sequencing, since it initiates at exon 1 and concludes with a poly(A) tail. To further validate its presence, we examined the distribution of *SF3B1* reads in scRaCH-seq data and compared it to matched bulk RNA-seq and scFLT-seq data for sample CLL2-RB. A consistent pattern in the *SF3B1* read distribution between scRaCH-seq and scFLT-seq was observed, with a higher coverage of exons 1–11 compared to that of exons 12–25 (**Supplemental Fig S6A**). However, when comparing single-cell sequencing data (scRaCH-seq and scFLT-seq) to bulk RNA-seq, bulk RNA-seq data exhibited a relatively stable read coverage across all exons (**Supplemental Fig S6B).** This hints at a potential bias introduced by the 10x Genomics

protocol into the composition of *SF3B1* transcripts. Upon closer examination, we discovered that 86% of single-cell reads mapped to exon 11 contained three cytosine (C) to thymine (T) mismatches following the drop in read coverage, indicating that these reads were created due to the unexpected binding of 10x poly(A) primers (**Supplemental Fig S6C)**. This artefact is reflected in the read length distribution plot with a higher proportion of reads shorter than the canonical isoforms of SF3B1 (**Fig. 2A**). These artefacts were subsequently removed from the scRaCH-seq data, resulting in the elimination of the novel *SF3B1* transcript which lacks the last 14 exons (**Supplemental Fig S6D**).

After removal of the artefacts, high expression levels of *SF3B1* transcript #1 (*SF3B1-211*, ENST00000487698), transcript #2 (a novel transcript containing exons 1 – 16) and transcript #3 (*SF3B1-201*, ENST00000335508) were consistently observed across multiple samples (**Figure 3C**). The expression of transcripts #4 (a novel transcript) and #5 (*SF3B1-204*, ENST00000414963) was relatively low (**Figure 3C**). In contrast to *BTK*, a clear difference in *SF3B1* isoform usage was evident between CLL cells from patient samples at screening and those from other CLL stages or healthy donor B cells (**Figure 3C**). Specifically, *SF3B1* transcript #1 was highly expressed in screening samples, whereas transcripts #2 and #3 were highly expressed in venetoclax-relapsed and healthy donor samples. In contrast to the CLL and B cells, *SF3B1* transcript #3 was the dominant transcript expressed by the T cells from the CLL patients at different stages and healthy donors (**Supplemental Fig S7**). To investigate why the screening samples exhibit higher expression of transcript #1, we linked isoform usage with gene expression. The CLL and B cells from all samples were isolated and re-clustered, revealing distinct clusters of the different stages, including one B cell cluster (C9), four screening clusters (C1, C3, C4, C12), three relapsed clusters (C0, C2, C8), and five clusters shared by screening and relapsed CLL cells (C5, C6, C7, C10, C11) (**Figure 3D**). *SF3B1* transcript #1 was specific to the screening CLL cells from cluster 3 (**Figure 3D**). Meanwhile, transcript #3, the canonical *SF3B1* isoform, was expressed by all CLL and B cells (**Figure 3D**). Subsequently, single-cell differential gene expression analysis between all the screening clusters (C1, C3, C4, C12) was performed. For this analysis, other clusters were excluded to eliminate confounding factors related to

treatment stages. CLL cells in C3 exhibited a high expression level of *DDX17*, a gene involved in almost all RNA metabolism processes(Xu et al. 2022) and *PARP15*, a negative regulator of transcription (Ryu et al. 2015) (**Figure 3E**).

Calling *SF3B1* mutations using scRaCH-seq data

Besides isoform detection, scRaCH-seq has the ability of linking the mutation status to cell state and differential gene expression. Among the CLL patients treated with venetoclax, *SF3B1* mutation was detected in 5 samples by whole exome sequencing (WES) (Thijssen et al. 2022) which was used as a ground truth to validate the scRaCH-seq data. For the identification of *SF3B1* mutations, we aggregated the scRaCH-seq data from all the samples, revealing many *SF3B1* transcripts with single-nucleotide polymorphism (SNP) (**Figure 4A**). The three most prominent SNPs were #5, #6, and #8. While SNP #6 and #8 exhibited a low frequency (~20%) across all samples, SNP #5 was unique to four specific samples. SNP #5 corresponds to the SF3B1 K700E mutation (Chr2:197402110 T→C), a mutation frequently observed at a sub-clonal level in CLL patients (Wan and Wu 2013; Landau et al. 2015; Landau et al. 2017). The K700E mutation can interrupt the recruitment of *SF3B1* to the correct 3' end branch site and consequently result in alternative splicing (Zhang et al. 2019). We identified the SF3B1 K700E mutation in CLL3 and CLL26 in both screening (S) and venetoclax-relapsed (R) samples (**Supplemental Fig S8A**). This mutation was also confirmed by the WES data from these samples(Thijssen et al. 2022). In scRaCH-seq, the K700E mutation was consistently observed with an overall frequency of 45.0% across the four samples, encompassing 87,674 UMI counts at the location of the SF3B1 K700E mutations and 48,224 UMI counts containing the reference base (Thymine) (**Figure 4B**). Although SNP #1 and #3 were specific to four samples and occurred at a high frequency, the overall UMI coverage for these altered transcripts was low (**Figure 4A-B**). A comparison of scRaCH-seq efficiency in capturing the SF3B1 K700E mutation with whole transcriptome short-read sequencing (scRNA-seq) and long-read sequencing (scFLT-seq) revealed that scRaCH-seq significantly increased the capture of cells carrying the mutation and enriched for transcripts (**Figure**

4C). scRaCH-seq obtains a 14-fold enrichment of reads for SF3B1 K700E compared to unbiased scFLT-seq, while approximately 168,000 cells were sequenced versus 1,300 cells.

Subsequently, we isolated cells harbouring SF3B1 K700E mutant reads from the scRaCH-seq data and visualized their distribution on the UMAP layout, which contained all patient sample cells (**Figure 5A**). Most cells with the mutant reads clustered within the CLL-cell group, consistent with the findings from WES. However, a small number of cells with mutant reads were detected within the non-CLL clusters, including the B cell cluster from a healthy donor, T cell cluster, and monocyte cluster (**Figure 5A and Supplemental Fig S8B**). Given that WES established the absence of *SF3B1* mutations in non-CLL cells, we investigated whether these non-CLL cells with *SF3B1* mutant reads might be attributed to false positives induced by doublets. Comparative analysis of UMI counts and gene counts revealed that non-CLL cells with *SF3B1* mutant reads exhibited similar values to CLL cells with *SF3B1* mutants (**Figure 5B**). Moreover, expression of multiple lineage markers within the non-CLL cells with *SF3B1* mutant reads was not observed, suggesting that these cells were not doublets (**Figure 5B**). In the *SF3B1* wild-type samples confirmed by WES, we identified altered reads with a read fraction of less than 1% (**Figure 5C),** indicative of reads containing sequencing errors (Sereika et al. 2022). To mitigate these false positives in both wild-type samples and non-CLL clusters, we established a threshold based on the abundance of *SF3B1* mutant transcripts detected per cell. This resulted in the exclusion of 80.7% of cells with *SF3B1* mutant transcripts in wild-type samples (**Figure 5C**). Furthermore, by raising the threshold to more than 2 *SF3B1* mutant transcripts, 93.4% of false positives in wild-type samples were successfully eliminated (**Figure 5C**). Consistent with false positives observed in wild-type samples, most *SF3B1*-mutant cells in non-CLL clusters exhibited low counts of altered reads (**Figure 5D**). Applying the threshold of more than 2 mutant reads per cell for the identification of a true positive mutant *SF3B1* cell resulted in the removal of most false positives in non-CLL clusters (**Figure 5E**). Employing the threshold of more than 2 mutant transcripts based on abundance was then established as a standard quality control step to eliminate false positives during mutation calling. The transcript with altered *SF3B1* #1 was observed in matched samples from CLL3 and CLL5 (**Supplemental Fig S8C**). This alteration was observed in both CLL

clusters and non-CLL clusters in the UMAP (**Supplemental Fig S8D**). These observations suggest that the C-to-G alteration in the *SF3B1* 3' UTR is likely a SNP present at a baseline level in specific patient samples. Notable, even with the threshold in place, we identified SNP #8, a C→T alteration at Chr2:197421097, in all samples and across all cell types (**Supplemental Fig S8E-F**).

Detecting the 6-base pair deletion in *SF3B1* among CLL cells

In addition to the SF3B1 K700E point mutation, WES also confirmed the presence of a 6-bp deletion (6bp-del) at Chr2:197402104:197402109, resulting in an SF3B1 KVR700**R mutation in CLL17 at a subclonal level (Thijssen et al. 2022). The original *FLAME*S pipeline was only designed to identify point mutations, so to address this limitation, we developed a supplementary script dedicated to detect the specific deletion directly in the FASTQ files of all samples. To minimize potential false positives, we applied the same criterion as we did for point mutation calling, necessitating at least two transcripts with the deletion per cell across all samples. Notably, cells with the 6bp deletion were exclusively observed in sample CLL17-R (**Figure 6A**), with a predominant clustering within the CLL group (**Figure 6B**), aligning with WES findings and underscoring the reliability of the deletion counting method. Subsequently, single-cell differential gene expression analysis was performed to identify gene expression changes between CLL cells with the *SF3B1* deletion and wild-type cells across all venetoclax-relapsed (R) samples. The analysis revealed a total of 377 DEGs with a $|\log_2 FC| > 0.5$ and adjusted p-value < 0.05. Among these, 152 genes showed significant upregulation, while 225 genes were downregulated in CLL cells with the KVR700**R mutation (**Figure 6C**). Similarly, we identified 147 differentially expressed genes (FDR < 0.05) between K700E mutant and wild-type CLL cells, comprising 28 upregulated genes and 119 downregulated genes in SF3B1 K700E CLL cells (**Supplemental Fig S9A**). Given the sub-clonal nature of the SF3B1 KVR700**R mutation in CLL17, DE analysis between mutant and wild-type *SF3B1* CLL cells of CLL17 was also conducted. We did not observe any DEGs (**Figure 6D**), implying that the identified DEGs between cells carrying the *SF3B1* 6-bp deletion and wild-type cells are likely relapsed patient-specific rather than mutation-driven.

Altered splicing in CLL cells with *SF3B1* mutations

It has been reported that the inhibition of RNA splicing can enhance the sensitivity of venetoclax in a mouse model bearing acute myeloid leukaemia (AML) (Wang et al. 2023). Therefore, the question remains if *SF3B1* mutation and downstream altered splicing events can contribute to venetoclax resistance in patients with CLL. The current scRaCH-seq probe panel was designed for two target genes to detect isoform usage and mutation calling. However, scFLT-seq from the same samples offers a broader overview of all genes at the transcript level, albeit with lower read coverage per gene(Thijssen et al. 2022). Interrogating the single-cell full-length transcriptomic data showed altered splicing in the CLL cells with *SF3B1* mutation including altered ‘3 splicing of the *TPT1* gene as an example (**Supplemental Fig S10A-B**). This is coherent with previous reports (Tang et al. 2020; Cortés-López et al. 2023), however, we found no evidence of altered splicing transcripts specific to the relapsed *SF3B1* mutated samples (**Supplemental Fig S10C**). Furthermore, no altered splicing of the BCL2 family genes was observed in *SF3B1* mutated CLL cells. This finding, in conjunction with the shared enriched pathways uncovered by scRNA-seq data (**Supplemental Fig S9B-C**), suggests that *SF3B1* mutations may not play a significant role in the development of acquired venetoclax resistance in CLL cells.

Discussion

Cell identity and function could be influenced by the alternative splicing of transcripts which will result in a substantial number of transcript isoforms. However, a significant portion of alternative transcripts will not be detected through high-throughput sequencing-based single-cell RNA-seq methods due to the short read length and the inherent bias toward 3′ or 5′ ends of the transcripts (Joglekar et al. 2023). To address this limitation, we developed scRaCH-seq, demonstrating high specificity and efficiency in capturing targeted long-read transcripts. This method provides an in-depth analysis of transcript usage of genes of interest, adding another layer to the existing single-cell short-read RNA-seq data. Integration of unique barcode primers specific to ONT allowed pooling and

efficient sequencing of multiple samples on the Nanopore platform. scRaCH-seq's capability to capture multiple transcripts simultaneously enhances isoform usage detection and facilitates the discovery of transcripts with mutations, providing a valuable tool across diverse research fields. Our study demonstrated the efficient capture of transcripts <5kb from pre-indexed and stored cDNA, depending on which genes were targeted. While this study focused on the capture of 2 genes that are highly expressed, we demonstrated the efficient capture of the transcripts of 17 and 24 genes, with the potential for larger gene sets without the need for primer optimization. Notably, scRaCH-seq demonstrated concordance with scRNA-seq (10x Genomics) in capturing genes, suggesting its reliability in transcriptomic profiling.

We optimized the *FLAMES* pipeline for scRaCH-seq data by incorporating read integrity checks to eliminate 3'end-truncated reads and setting a quality control threshold of at least two mutant reads to reduce false positives. Additionally, a deletion counting script was created for identifying known deletions. This allowed the detection of the SF3B1 K700E mutation and the 6-base pair deletion in *SF3B1* in particular samples, both of which were previously confirmed by WES. Meanwhile, the incidence of false positives stemming from sequencing errors was minimal after the filtering based on the abundance of mutant reads detected in CLL cells. While single-cell DNA sequencing, such as Mission Bio Tapestri, provides insight into clonality upon drug resistance (Thompson et al. 2022), it lacks transcriptome profiling per cell. In contrast, scRaCH-seq offers short-read whole transcriptomic data and mutation status if the mRNA of the targeted gene is expressed. By integrating the short-read scRNA-seq data, we observed multiple significant DEGs between SF3B1 K700E/KVR700**R and wild-type CLL cells. The subclonal nature of SF3B1 KVR700**R in CLL17 allowed us to study the mutation's impact at a single-cell level. No DEGs were observed between SF3B1 KVR700**R and wild-type CLL cells from sample CLL17. However, scFLT-seq demonstrated that the SF3B1 KVR700**R mutation increased the expression of novel isoforms in CLL17. Our single-cell data revealed that the altered gene expression between *SF3B1*-mutated and wild-type CLL cells was primarily driven by patient specificity. This emphasizes the importance of caution when drawing

conclusions based on bulk RNA-seq data comparing wild-type and mutant samples from different patients.

Venetoclax is an effective therapy for CLL patients with the potential to induce long remissions. However, we showed that multiple mechanisms resulting in the deregulation of apoptotic genes could occur at a polyclonal level in venetoclax-relapsed patient samples (Thijssen et al. 2022). Now by incorporating scRaCH-seq, we can add another layer of information and study the effect of an *SF3B1* mutation on venetoclax sensitivity. While it was observed that inhibiting RNA-splicing could enhance venetoclax sensitivity in AML (Wang et al. 2023), we demonstrated that *SF3B1* mutated venetoclax-relapsed CLL cells did not express novel isoforms of the BCL2 family members that could impact venetoclax sensitivity. We observed differential transcript usage of *SF3B1* between screening and venetoclax-relapsed samples. *SF3B1*-211 was found to be specific to a cluster of screening samples but the specific role of this *SF3B1* isoform remains unknown. The CLL cells with this *SF3B1*-211 isoform expressed lower levels of ribosomal genes. These findings suggest a potential connection between the usage of *SF3B1*-211 transcript and the senescence stage in CLL cells, possibly through its impact on gene transcription activity and protein synthesis.

Lastly, we observed a consistent *SF3B1* artefact across all samples in the scRaCH-seq and scFLT-seq data. Comparison of read coverage between scRaCH-seq, scFLT-seq, and bulk RNA-seq data revealed that this artefact resulted from the unexpected binding of the 10x reverse transcriptase primer, rather than being a technical issue with the scRaCH-seq protocol. This unexpected binding of the 10x primers did not impact the results of scRNA-seq since scRNA-seq operates at the gene level, counting all reads aligned to *SF3B1*. However, it could pose a problem for scFLT-seq and scRaCH-seq, which distinguish reads at the transcript level. Therefore, when identifying novel transcripts in scRaCH-seq or scFLT-seq data, it is advisable to conduct a thorough examination of potential unexpected binding sites of 10x primers within the corresponding gene.

In conclusion, scRaCH-seq provides an innovative strategy for studying long-read transcripts from pre-indexed cDNA, holding promise for advancing gene expression studies and unravelling complex

biological processes. scRaCH-seq can be scaled to the number of target transcripts. We demonstrated a 14-fold read enrichment in capturing a mutation compared to unbiased long-read sequencing, while 21 samples were processed on a single PromethION flow cell versus 1 sample on a flow cell for unbiased long-read sequencing. Its potential adaptability for PacBio sequencing further extends its utility, with the primary advantage lying in linking transcript usage and mutation status at a single-cell level. The approach is both cost-effective, high throughput and flexible, which is achieved by leveraging a wide range of widely used 10x single-cell protocols, making scRaCH-seq applicable for large-scale studies to comprehensively characterize cellular heterogeneity. It holds potential for integration into single-cell spatial data, representing a powerful advancement in single-cell genomics for understanding cellular heterogeneity in development, disease, and other biological processes.

Materials and methods

Samples

The full-length cDNA of 18 peripheral blood samples from chronic lymphocytic leukaemia (CLL) patients and 3 healthy donors (HD) were used(Thijssen et al. 2022). 10x Genomics was performed using the Chromium Next GEM (Gel Bead-In Emulsions) Single Cell V(D)J (v1.1, 10x Genomics, cat#PN-1000165) according to the manufacturer's instructions. The indexed full-length surplus cDNA(Thijssen et al. 2022) was stored and used as input for this scRaCH-seq experiment (**Figure 1**).

Probe panel design

Probes were designed to capture the indexed cDNA of target genes. Each probe was 120 base pairs (bp) in length, ensuring robust coverage of the targeted regions. First, a FASTA file was generated for all the isoforms of the genes of interest. The FASTA file containing all Ensembl-annotated exons from all the isoforms was compiled and this was achieved by extracting the corresponding sequences from the human genome GRCh38. This FASTA file served as the foundation for probe design. Next, exon sequences shorter than 120 bp were merged with the preceding and subsequent exons to create

concatenations exceeding 120 bp, resulting in an updated FASTA file with all reads over 120 bp. Using the custom FASTA file, a probe panel was generated to cover each base in the input, employing https://sg.idtdna.com/pages/tools/xgen-hyb-panel-design-tool. For probe selection, a strategy was implemented to cover every 1000 bp of an exon sequence with a single probe (**Supplemental Fig S1**). Probes were selected based on their GC content (GC%) to ensure consistent efficiency during their hybridization with cDNA. The average GC% of the redundant probe panel was used as a baseline. The probe with GC% closest to the average GC% was chosen from the group of probes covering a thousand-base region. For concatenated exons, probes fully covering the entire sequence of the original short exon were selected (**Supplemental Fig S1).** Duplicated probes in the selected panel were removed to finalize the probe panel. A pre-fixed R script (Team 2021) detailing the code for the probe panel design process as well as a step-by-step instruction is available at the following GitHub repository: https://github.com/HongkePn/RaCHseq_Probe_Design.

Single-cell Rapid Capture Hybridization sequencing (scRaCH-seq)

Amplification of cDNA

The surplus 10x indexed cDNA was amplified to increase the cDNA yield for the hybridization step. To obtain (1.5-2 µg) cDNA, 2-10ng of cDNA was used as input in 5×50µl reactions. For a 50µl reaction using 10× Genomics 3' or 5' cDNA, a mixture containing 10µl 5× PrimeSTAR buffer (Takara, #R050B), 4µl dNTP (2.5mM; Takara, #R050B), 1µl Partial R1 – CTACACGACGCTCTTCCGATCT (10µM), 1µl T5′ PCR Primer IIA – AAGCAGTGGTATCAACGCAGAG (10µM), cDNA (5-10ng), nuclease-free water (33µl-volume cDNA), 1µl PrimeSTAR GXL DNA Polymerase (Takara, #R050B) was made. The thermocycler was programmed as follows: 98°C for 30sec, 9cyles of 98°C for 10sec, 65°C for 15sec and 68°C for 8min, 1cycle of 68°C for 10min. The concentration of the resulting product was assessed, and if it fell below 10 ng/µl, additional cycles were carried out as needed. The pooled amplified cDNA was cleaned up with 0.6× AMPure XP or SPRIselect beads and taken up in 50µl 5× diluted Elution Buffer (EB; Qiagen) in nuclease-free water (Thermo Fisher).

We optimised scRaCH-seq so that it can also be applied to 10x Genomics multiome cDNA. A mixture containing 2µl cDNA (2ng), 25µl 2× KAPA HiFi master mix (Roche, #KK2602), 1µl Partial R1 – CTACACGACGCTCTTCCGATCT (10µM), 1µl TSOp1 – TGGTATCAACGCAGAGTACATGGG (10µM), 21µl EB buffer was made. The thermocycler needs to be programmed as follows: 98°C for 3min, 8 cycles of 98°C for 15sec, 64°C for 20sec and 72°C for 7min, and 1 cycle of 72°C for 10min.

Hybridization

The amplified cDNA sample (1.5-2µg) was dried using a DNA vacuum concentrator (speed vac) together with 1µl of 1000µM IDT Blocking Oligos: Poly (A): TTTTTTTTTTTTTTTTTTTTTTTTTTTTT/3InvdT/, PR1: CTACACGACGCTCTTCCGATCT, SO: AAGCAGTGGTATCAACGCAGAGTAC. Subsequently, the dried sample was taken up in the following mixture: 8.5µl of 2× Hybridization Buffer, 2.7µl Hybridization Buffer Enhancer (xGen IDT, cat#1080577) and 1.8µl nuclease-free water. The sample was incubated at 95°C for 10min to denature the cDNA. To this denatured sample, 4µl of custom probe panel was added and incubated at 65°C for 16 hours to facilitate probe hybridization with the cDNA.

Probe Pulldown

We used the xGen IDT Lockdown Hybridization and Wash Kit. The sample was added directly to the dried washed beads (100µl of Dynabeads M-270 Streptavidin) and incubated at 65°C for 45min. Every 10min the sample and beads were mixed. The captured cDNA was then thoroughly washed, and the bead plus sample mixture was resuspended in 50µl EB.

Amplification of Captured cDNA Sample

The captured cDNA with beads was amplified using LA Taq DNA Polymerase Hot-Start (Takara, # RR042B) in 4×50µl reactions. For a 50µl reaction, a mixture containing 5µl 10x PrimeSTAR buffer, 4µl dNTP (2.5mM), 1µl FPSfilA: ACTAAAGGCCATTACGGCCTACACGACGCTCT TCCGATCT

(10µM), 1µl RPSfilBr: TTACAGGCCGTAATGGCCAAGCAGTGGTATCAACGCAGAGTA (10µM), 12.5µl cDNA on beads, 26.1µl nuclease-free water, 0.4µl LA Taq DNA Polymerase was made. The thermocycler was programmed as follows: 95°C for 2min, 12cyles of 95°C for 20sec, and 68°C for 10min, 1cycle of 72°C for 10min. After amplification, the cDNA was cleaned up with 0.6× AMPure XP or SPRIselect beads and taken up in 30µl EB.

Nanopore library preparation and sequencing

The Oxford Nanopore Technology (ONT) SQK-PCB111.24 kit was used to index the samples. Since the cDNA will have the ONT overhang, we started from the "Selecting for full-length transcripts by PCR" step in the protocol with 5µl (1ng/µl) scRaCH-seq library, 0.75µl Unique Barcode Primer, 6.75µl Nuclease-free water and 12.5µl 2× LongAmp Hot Start Taq Master Mix. The thermocycler was programmed as follows: 95°C for 30sec, 5 cycles of 95°C for 15sec, 62°C for 15sec and 65°C for 1min, and 1 cycle of 65°C for 6min. After the PCR, the sample was cleaned up with 0.6× beads. For the *BTK* and *SF3B1* gene capture, 21 samples were pooled together and sequenced on PromethION R9.4.1 flow cell (**Figure 1**).

The scRaCH-seq libraries can also be sequenced on a PromethION R10.4.1 flow cell and using the SQK-NBD114.24 Native Barcoding Kit 24 to index the samples and pool them together. 12µl of 200fmol of the scRaCH-seq library was combined with 0.1µl Diluted DNA Control Sample (DCS) and the other protocol components, resulting in a 15µl mixture. This mixture was incubated at 20°C for 10min and 65°C for 10min.

Short-read data analysis

Data processing

We used *cellranger* (v5.0.0) and *bcl2fastq* (v2.19.1) to pre-process our short-read sequencing data. The percentage of mitochondrial gene counts was calculated using the `PercentageFeatureSet` function in *Seurat* (v4.0.5)(Stuart et al. 2019). This calculation targeted genes starting with the regex

pattern "MT-". For each sample, the `isOutlier` function from *Scater* (v1.20.0)(McCarthy et al. 2017) was used to identify low-quality cells. These were defined as cells deviating by more than 3 median absolute deviations (MADs) from the median in terms of unique molecular identifiers (UMIs, both higher and lower), detected gene numbers (higher and lower), and mitochondrial gene expression (higher). Cells identified as low-quality were then excluded from further analysis. Library size normalization and $\log_2(x + 1)$ transformation were performed using the `NormalizeData` function in *Seurat*. This step was followed by the identification of highly variable genes (HVGs) using the `FindVariableFeatures` function in *Seurat*, adhering to default parameters. These HVGs were then scaled and centred with the `ScaleData` function in *Seurat*.

UMAP by gene expression

We performed principal component analysis (PCA) by applying the `RunPCA` function in *Seurat* to scaled HVGs (n=2000) with the default parameter settings. To remove the unwanted variability due to batch effects, we applied *Harmony* (v0.1.0) for batch correction (Korsunsky et al. 2019). For cell clustering analysis, we employed the shared nearest neighbour (SNN) method implemented in *Seurat*. To construct the SNN graph, we used the `FindNeighbors` function with the first 20 Harmony corrected PCA embeddings and k.param = 20. For visualization, we generated a Uniform Manifold Approximation and Projection (UMAP) using the `RunUMAP` function with the first 20 *Harmony*-corrected PCs. To identify cell clusters, we executed the `FindClusters` function on the SNN graph, employing the original Louvain algorithm with default parameters. To identify the doublets, we removed the cluster that exhibited the expression of multiple immune cell lineage markers and high RNA contents. Finally, we applied the same *Seurat* functions on the filtered data to re-cluster cells and update the UMAP.

Single-cell differential expression analysis

To perform single-cell differential expression analysis, we employed the `FindMarkers` function of *Seurat*, using the parameters configured as test.use = "MAST", logfc.threshold = 0 and min.pct = 0.

Pseudo-bulk differential expression analysis

The count matrix for gene expression was aggregated using the `aggregateAcross` function incorporated in *Scater*. Subsequently, the Likelihood ratio test, implemented in *edgeR* (v3.34.0) was applied to identify differentially expressed genes (DEGs)(Robinson et al. 2010).

Gene set enrichment analysis

We downloaded the C2 canonical pathways and Hallmark pathway collections using *msigdbr* (v7.5.1)(Liberzon et al. 2015). All genes were then ordered in a decreasing fashion based on their log-fold changes. This ordered list of genes was then used as input for the `GSEA` function in *ClusterProfiler* (v4.4.4), which was tested against the gene set collections using default parameters(Wu et al. 2021).

scRaCH-seq data analysis

We developed a customized *FLAMES* (v1.3.4) pipeline to conduct the scRaCH-seq analysis (**Figure 1**). A detailed description of the original *FLAMES* pipeline can be found in the Methods paper that introduces single-cell Full-Length Transcriptome sequencing (scFLT-seq) (Tian et al. 2021).

Data pre-processing (STEP 01)

We base called the raw data of nanopore sequencing using the *Guppy* software (v3.1.5) with the "dna_r9.4.1_450bps_sup_prom.cfg" configuration, resulting in the generation of FASTQ files. Subsequently, we used the "find_barcode" function in *FLAMES* to demultiplex the FASTQ files, which was accomplished by cross-referencing the cell barcodes identified in the corresponding scRNA-seq data. We allowed 2 base pairs of edit distance for the cell barcode matching.

Read integrity check (STEP 02)

We added an extra step of reads integrity check to the *FLAMES* pipeline. The demultiplexed reads underwent an examination to confirm the presence of all essential components. These components include a cell barcode, a unique molecular identifier (UMI), a template switch oligo (TSO) sequence, and a poly(A) tail at the end of the transcript. Users can specify the TSO sequence in the "find_barcode" function in *FLAMES* and filter reads that are missing the specified TSO sequence.

Point mutation calling (STEP 03)

The reads were first aligned to the human genome GRCh38 (downloaded from GENCODE) using the *FLAMES* function "minimap2_align". We then used the mutation calling function in *FLAMES* (v0.1) with the default parameters to identify the point mutations, by comparing the pile-up reads against the gene of interest. For SF3B, the gene region Chr2: 197,388,515 – 197,435,079 from the human genome GRCh38 was selected. For each sample, the positions with alteration frequencies ranging from 10% to 90% were identified as mutations. The allele frequency of identified mutations was then counted and subsequently summarized to generate a mutation count matrix, with the rows corresponding to mutations and the columns named after the cell barcodes.

Deletion calling (STEP 03)

We incorporated an additional step specifically to detect multiple base pair deletions in genes of interest. We first aligned the reads which passed the integrity check to the human genome GRCh38, using *minimap2* with parameters set as "-ax splice --junc-bonus 1 -k14 --secondary=no --junc-bed"(Li 2018). We then compared the pile-up from the reads to the reference by checking the alignments at the specific chromosomal position. For the *SF3B1* 6bp deletion previously identified by WES, read alignments were checked at the Chr2:197,402,104 - 197,402,109 position to identify the presence of the 6bp deletion. The reads with the multiple bp deletion were merged by UMIs and subsequently counted for each cell to generate a count matrix, with rows representing the *SF3B1* 6bp-deletion/WT and column names corresponding to the cell barcodes. We updated *FLAMES* to provide functions for detection of multiple base pair deletions along with point mutations at both piled-up bulk level ("find_variants") and single-cell level ("sc_mutations") to simplify such analyses.

Isoform detection (STEP 04)

We employed the *FLAMES* function “find_isoform” to summarize the alignment for making isoform assembly. The isoforms that exhibited <10bp variance in their splicing junctions and <100bp variance in their start or end sites were merged. Following this, we re-aligned the reads that passed the integrity check to the isoform assembly to generate the isoform count matrix, with the rows corresponding to isoforms and the columns named after the cell barcodes.

Data integration (STEP 05)

We combined the isoform and mutation count matrices with the processed single-cell short-read data to construct a comprehensive multi-assay *Seurat* object, using the 'CreateAssayObject' function. The scRaCH-seq data were used to identify cell groups with specific isoforms or mutations, followed by short-read differential expression analysis between these groups.

Differential transcript usage analysis

For scRaCH-seq data, we merged the per-cell counts of target gene transcripts by samples to make pseudo-bulk transcript data. We calculated the frequency of different transcripts for our genes of interest *SF3B1* and *BTK* respectively and then displayed the top 5 transcripts with the highest frequencies across all samples for both target genes. For scFLT-seq data, cells from patients harbouring *SF3B1* mutations/deletions were combined per sample to make pseudo-bulk transcript data for the *SF3B1*-mutated group. Cells of wild-type screening/venetoclax-relapsed/healthy donor samples were merged to construct screening-wt/relapsed-wt/healthy-wt pseudo-bulk transcript data for the *SF3B1*-wild type group. We then applied the function “diffSpliceDGE” in *edgeR* to identify the differential transcript usage between groups (*SF3B1*-mutated, *SF3B1*-wt) using pseudo-bulk data as replicates. The scFLT-seq data and whole exome sequencing data involved in this study can be access using the EGA accession number EGAS00001005815 (Thijssen et al. 2022).

Data Access

All raw sequencing data generated in this study have been submitted to the European Genome-phenome Archive (EGA; https://ega-archive.org/) under accession number EGAD50000000235.

The probe panel design code is available at https://github.com/HongkePn/RaCHseq_Probe_Design. The complete scRaCH-seq analysis code is available at https://github.com/HongkePn/scRaCHseq_data_analysis. We have also included a copy of the code in the Supplemental Material of this paper (Supplemental Code).

Competing Interest Statement

The authors declare no competing interests.

Acknowledgements

We thank Stephen Wilcox, Sarah MacRaild, the WEHI SCORE team and core facilities (Genomics) for their support. Illustrations were created with BioRender.com. This work was supported by fellowships and grants from the Australian National Health and Medical Research Council: Program Grants 1113133 to D.C.S.H.; Synergy 2011139 to A.H.W., A.W.R. and D.C.S.H.; Fellowship 1156024 to D.C.S.H.; Ideas Grant 2013478 to D.C.S.H. and R.T; Investigator 2017257 to M.E.R. and 1174902 to A.W.R, the Leukemia & Lymphoma Society of America (Specialized Center of Research (SCOR) grant 7015-18 to A.W.R. and D.C.S.H.), The Australian Research Council (Discovery Project 200102903 to M.E.R.), Leukaemia Foundation (grant 2012526 to R.T.), Victorian Cancer Agency (ECRF21014 fellowship to R.T.), the University of Melbourne (MIRS and MIFRS scholarships to H.P.), the Chan Zuckerberg Initiative DAF (an advised fund of Silicon Valley Community Foundation; grant number 2019-002443 to M.E.R.), and the Amsterdam UMC Fellowship (R.T.). This work was made possible through Victorian State Government Operational Infrastructure Support and the Australian Government NHMRC IRIISS.

617 *Author contribution:* H.P., D.C.S.H., M.E.R. and R.T. designed the study. A.H.W. and A.W.R., were
618 responsible for patient care and recruited patient samples. H.P., J.S.J., C.C.C., N.S.A. and R.T.
619 acquired the data. H.P., L.T., C.W., Y.Y., N.A., N.M.D. and M.E.R. analysed and interpreted the data.
620 C.W. updated *FLAMES*. H.P., M.E.R. and R.T. wrote the first version of the manuscript; all authors
621 reviewed the data and contributed to critical revision of the manuscript
622

References

Cheng O, Ling MH, Wang C, Wu S, Ritchie ME, Goke J, Amin N, Davidson NM. 2024. Flexiplex: a versatile demultiplexer and search tool for omics data. *Bioinformatics* **40**.

Cortés-López M, Chamely P, Hawkins AG, Stanley RF, Swett AD, Ganesan S, Mouhieddine TH, Dai X, Kluegel L, Chen C et al. 2023. Single-cell multi-omics defines the cell-type-specific impact of splicing aberrations in human hematopoietic clonal outgrowths. *Cell Stem Cell* **30**: 1262-1281.e1268.

Griffin GK, Booth CAG, Togami K, Chung SS, Ssozi D, Verga JA, Bouyssou JM, Lee YS, Shanmugam V, Hornick JL et al. 2023. Ultraviolet radiation shapes dendritic cell leukaemia transformation in the skin. *Nature* **618**: 834-841.

Joglekar A, Foord C, Jarroux J, Pollard S, Tilgner HU. 2023. From words to complete phrases: insight into single-cell isoforms using short and long reads. *Transcription* doi:10.1080/21541264.2023.2213514: 1-13.

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**: 1289-1296.

Landau DA, Sun C, Rosebrock D, Herman SEM, Fein J, Sivina M, Underbayev C, Liu D, Hoellenriegel J, Ravichandran S et al. 2017. The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nat Commun* **8**: 2185.

Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, Kluth S, Bozic I, Lawrence M, Bottcher S et al. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**: 525-530.

Lebrigand K, Magnone V, Barbry P, Waldmann R. 2020. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nature Communications* **11**: 4025.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**: 417-425.

Liu T, Liu C, Yan M, Zhang L, Zhang J, Xiao M, Li Z, Wei X, Zhang H. 2022. Single cell profiling of primary and paired metastatic lymph node tumors in breast cancer patients. *Nat Commun* **13**: 6823.

McCarthy DJ, Campbell KR, Lun AT, Wills QF. 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**: 1179-1186.

Mustachio LM, Roszik J. 2022. Single-Cell Sequencing: Current Applications in Precision Onco-Genomics and Cancer Therapeutics. *Cancers (Basel)* **14**.

Nagler A, Wu CJ. 2023. The end of the beginning: application of single-cell sequencing to chronic lymphocytic leukemia. *Blood* **141**: 369-379.

Nam AS, Kim KT, Chaligne R, Izzo F, Ang C, Taylor J, Myers RM, Abu-Zeinah G, Brand R, Omans ND et al. 2019. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**: 355-360.

R Core Team. 2021. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL https://www.R-project.org/

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Ryu KW, Kim D-S, Kraus WL. 2015. New Facets in the Regulation of Gene Expression by ADP-Ribosylation and Poly(ADP-ribose) Polymerases. *Chemical Reviews* **115**: 2453-2481.

Sandberg R. 2014. Entering the era of single-cell transcriptomics in biology and medicine. *Nat Methods* **11**: 22-24.

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* **19**: 823-826.

Stewart CA, Gay CM, Xi Y, Sivajothi S, Sivakamasundari V, Fujimoto J, Bolisetty M, Hartsfield PM, Balasubramaniyan V, Chalishazar MD et al. 2020. Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* **1**: 423-436.

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**: 865-868.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902.e1821.

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**: 1438.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**: 377-382.

Thijssen R, Tian L, Anderson MA, Flensburg C, Jarratt A, Garnham AL, Jabbari JS, Peng H, Lew TE, Teh CE et al. 2022. Single-cell multiomics reveal the scale of multilayered adaptations enabling CLL relapse during venetoclax therapy. *Blood* **140**: 2127-2141.

Thompson ER, Nguyen T, Kankanige Y, Markham JF, Anderson MA, Handunnetti SM, Thijssen R, Yeh PS, Tam CS, Seymour JF et al. 2022. Single-cell sequencing demonstrates complex resistance landscape in CLL and MCL treated with BTK and BCL2 inhibitors. *Blood Adv* **6**: 503-508.

Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C et al. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22**: 310.

Tian Y, Li Q, Yang Z, Zhang S, Xu J, Wang Z, Bai H, Duan J, Zheng B, Li W et al. 2022. Single-cell transcriptomic profiling reveals the tumor heterogeneity of small-cell lung cancer. *Signal Transduct Target Ther* **7**: 346.

Wan Y, Wu CJ. 2013. SF3B1 mutations in chronic lymphocytic leukemia. *Blood* **121**: 4627-4634.

Wang E, Pineda JMB, Kim WJ, Chen S, Bourcier J, Stahl M, Hogg SJ, Bewersdorf JP, Han C, Singer ME et al. 2023. Modulation of RNA splicing enhances response to BCL2 inhibition in leukemia. *Cancer Cell* **41**: 164-180.e168.

Wang L, Mo S, Li X, He Y, Yang J. 2020. Single-cell RNA-seq reveals the immune escape and drug resistance mechanisms of mantle cell lymphoma. *Cancer Biol Med* **17**: 726-739.

Wang X, Nissen M, Gracias D, Kusakabe M, Simkin G, Jiang A, Duns G, Sarkozy C, Hilton L, Chavez EA et al. 2022. Single-cell profiling reveals a memory B cell-like subtype of follicular lymphoma with increased transformation risk. *Nat Commun* **13**: 6772.

Wu S, Schmitz U. 2023. Single-cell and long-read sequencing to enhance modelling of splicing and cell-fate determination. *Comput Struct Biotechnol J* **21**: 2373-2380.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L et al. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**: 100141.

Xu K, Sun S, Yan M, Cui J, Yang Y, Li W, Huang X, Dou L, Chen B, Tang W et al. 2022. DDX5 and DDX17—multifaceted proteins in the regulation of tumorigenesis and tumor progression. *Frontiers in Oncology* **12**.

Zhang J, Ali AM, Lieu YK, Liu Z, Gao J, Rabadan R, Raza A, Mukherjee S, Manley JL. 2019. Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Mol Cell* **76**: 82-95.e87.

Figure legends

Figure 1 Schematic of scRaCH-seq approach and *FLAMES* pipeline for single-cell long-read data analysis.

scRaCH-seq can be incorporated in the standard high-throughput single-cell RNA-seq experiments (left side). After single-cell isolation, mRNA is converted into cDNA and barcoded. Only some of this amplified indexed cDNA is used for short-read library preparation and Illumina sequencing. The surplus cDNA will be stored and can be used for scRaCH-seq. For scRaCH-seq, a probe panel will be designed for genes of interest. The biotinylated probes will be hybridized with amplified cDNA overnight. The probes and target genes will be captured with Streptavidin beads, washed, and amplified. The enriched target long-read transcripts will be sequenced on the Nanopore platform. With the *FLAMES* pipeline, FASTQ files were demultiplexed by cross-referencing the cell barcodes identified in scRNA-seq data (STEP 01). Next, the demultiplexed reads will undergo an integrity check and the reads that possess a UMI, a TSO sequence, and poly(A) tails are retained (STEP 02). Reads were aligned to the GRCh38 reference genome to construct a transcript assembly. Concurrently, the piled-up reads were compared against GRCh38 to identify base alterations and deletions, generating a comprehensive mutation/deletion matrix (STEP 03). The reads will be realigned to the transcript assembly for quantification (STEP 04). Single-cell short-read gene expression data was used to cluster the cells, based on the conventional analysis pipeline for scRNA-seq. The transcript usage and mutation/deletion information were then aligned with the single-cell gene expression data. The bridge connecting these datasets was the shared cell barcodes, ensuring the gene expression profile at a transcript level (STEP 05).

Figure 2 scRaCH-seq is efficient in capturing enriched genes of interest.

A) A graph showing the losses in read counts due to demultiplexing and integrity checks, depicting the distribution of read lengths ranging from 0 to 5,000 base pairs. The canonical isoform of *BTK* (2,027 bp) and *SF3B1* (2,225 bp) are marked on the plot. The four shades of grey, ranging from light

to dark, correspond to raw reads, demultiplexed reads, reads with TSO, and reads with also a poly(A) end.

B) Bar plots showing the loss of reads after demultiplexing (top panel), barcodes detected in long-read sequencing data (middle panel) and counts of off-target and on-target reads for each sample (bottom panel).

C) Violin plot showing the $\log_{10}$ UMI counts of collapsed *BTK* (red), *SF3B1* (orange) and off-target transcripts (purple) based on isoform detected by scRaCH-seq (left panel). Violin plot showing the $\log_{10}$ counts of cells possessing collapsed transcripts of *BTK* (red), *SF3B1* (orange) and off-target genes (purple) (right panel).

D) Dot plot showing the $\log_{10}$ counts of off-target genes with the top 10 off-target transcripts specifically marked.

E) Dot plot showing the per-cell UMIs of *SF3B1* (red) and *BTK* (orange), detected by scRNA-seq (X-axis) and scRaCH-seq (Y-axis) for all samples.

F) UMAP projection of peripheral blood mononuclear cells from CLL patients or healthy donors and clustering based on short-read gene expression.

G) Violin plot showing the expression of CD14 (monocyte marker), CD19 (B/CLL cell marker) and CD3 (T cell marker) per cell cluster (Figure 2F).

H) UMAP projection of *SF3B1* (left) and *BTK* (right) gene-level expression detected by scRNA-seq (purple) or scRaCH-seq (green).

Figure 3 scRaCH-seq reveals different *SF3B1* isoform usage by CLL cells from patients undergoing different treatments.

A) Heatmap showing *BTK* isoform usage (rows) per sample (columns) in the B/CLL cells. The top 5 *BTK* transcripts are illustrated on the right with novel transcripts highlighted in yellow. Samples are grouped in screening (S; green), venetoclax relapsed (VEN-R/R; pink), venetoclax relapsed and subsequently on BTKi (VEN-RB/RB; bright yellow) and healthy donor (HD; turquoise).

B) Illustration of the dominant *SF3B1* transcript identified and characterized by the absence of the last 14 exons in the 3' end (highlighted in purple).

C) Heatmap showing *SF3B1* isoform usage (rows) per sample (columns) in the B/CLL cells. The top 5 *SF3B1* transcripts are illustrated on the right with novel transcripts highlighted in yellow.

D) UMAP projection of CLL/B cells from CLL patients or healthy donors and clustering based on short-read gene expression. The samples from CLL patients at screening (S; green), venetoclax relapsed (R; pink), venetoclax relapsed and subsequently on BTKi (RB; bright yellow) and healthy donors (HD; turquoise) are highlighted in the UMAP. The expression of the *SF3B1* transcript #1 and #3 are overlaid as red dots in the UMAP (Figure 3B).

E) Dot plot showing marker genes that distinguish the CLL cells present in the 4 clusters shared by the screening samples (C1, C3, C4 and C12 in Figure 3C). Cluster 3 with *SF3B1* transcript #1 usage is indicated in bold.

Figure 4 Mutation calling with scRaCH-seq and *FLAMES*.

A) Dot plot showing the *SF3B1* point mutations identified using *FLAMES* (left panel). After consolidating reads with UMIs, the reads from all the samples were aligned to the GRCh38 reference genome. The size of each dot in the plot corresponds to the proportion of altered transcripts.

B) A graphical representation highlights the precise locations of these high-frequency (>10%) alterations within the *SF3B1* gene. The bottom graph shows the frequency of altered *SF3B1* transcripts in each sample.

C) Bar plots showing the cells with a SF3B1 K700E mutation (left panel) or transcripts (UMIs) with SF3B1 K700E mutation (right panel) detected by short-read scRNA-seq, long-read scFLT-seq or long-read scRaCH-seq.

Figure 5 The SF3B1 K700E mutation is detected by scRaCH-seq.

A) UMAP projection of cells carrying the SF3B1 K700E mutation (red).

B) Violin plot showing the UMI counts, gene counts (Features), expression of mitochondria genes (MT%), expression of CD3 (T cell marker), CD14 (monocyte marker) and CD19 (B cell marker) per cell type cluster.

C) Bar plot showing the abundance of mutant transcripts. Left plot showing cells per sample with >0 transcripts (UMI) of mutant SF3B1 K700E. Middle plot showing cells per sample with >1 transcripts (UMIs) of *SF3B1* mutation. Right plot showing cells per sample with >2 transcripts (UMIs) of *SF3B1* mutation. Samples with confirmed SF3B1 K700E mutation by WES are highlighted in dark grey.

D) Bar plot showing the distribution of the abundance of SF3B1 K700E mutation detected in CLL (pink) and non-CLL (blue) cells. The vertical dashed line indicates the criterion of >2 *SF3B1* mutation transcripts.

E) UMAP projection of cells carrying >2 SF3B1 K700E mutation transcripts (red).

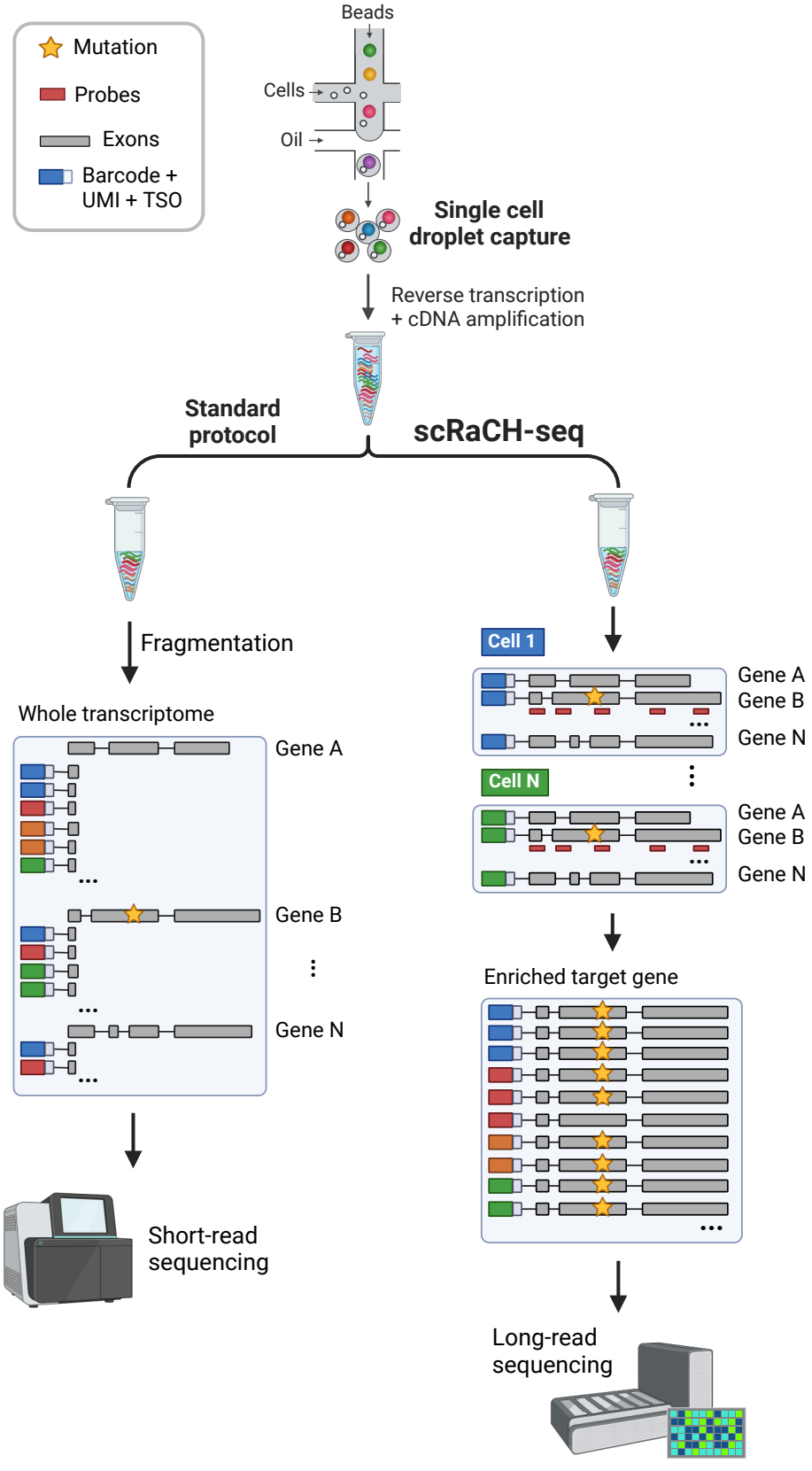
Figure 6 The SF3B1 KVR700R alteration is detected by scRaCH-seq.**

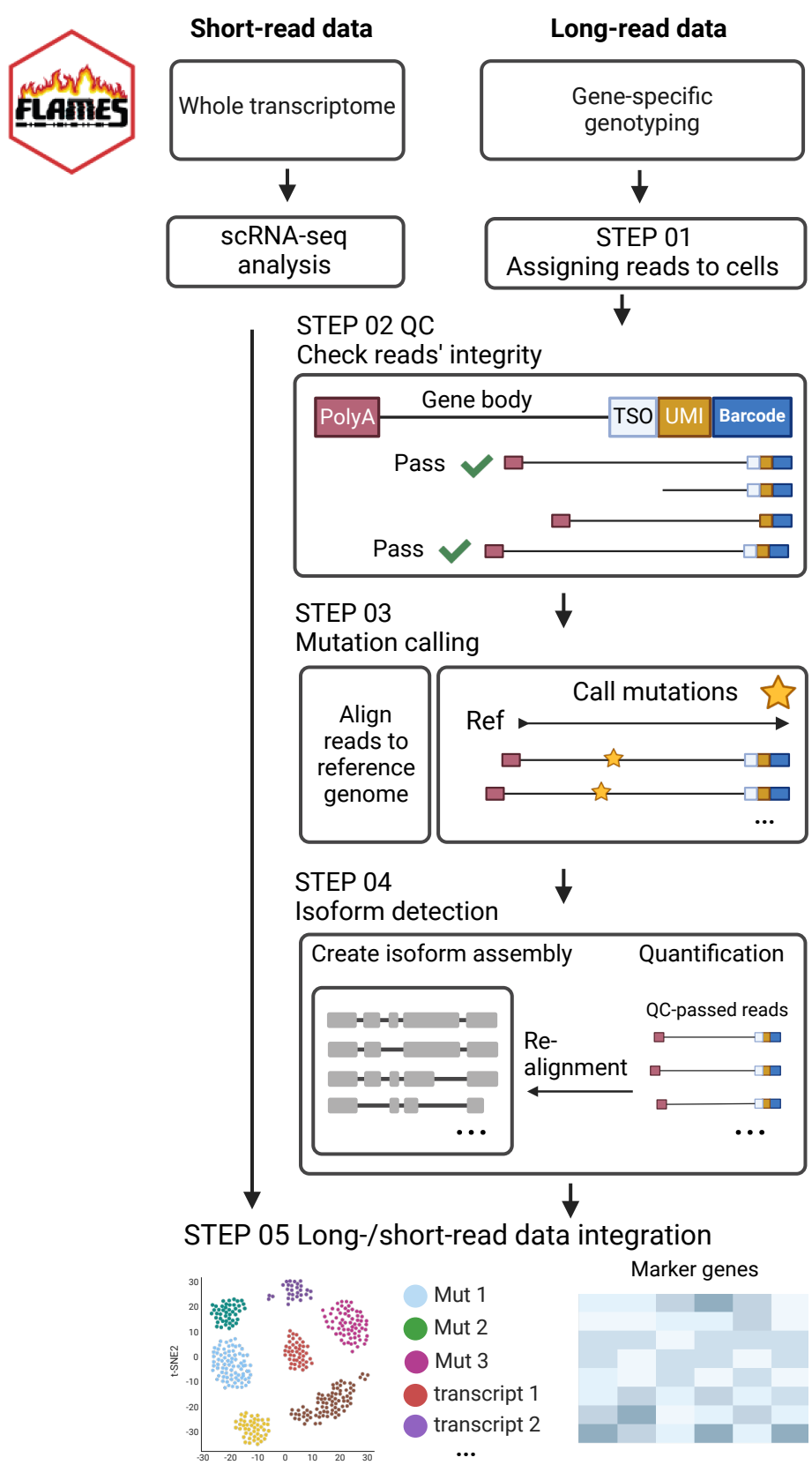
A) Bar plot showing the cells with *SF3B1* 6bp-deletion (Chr2:197402104:197402109) across all samples. A criterion of > 2 of *SF3B1* with 6bp-del transcripts (UMIs) per cell was employed.

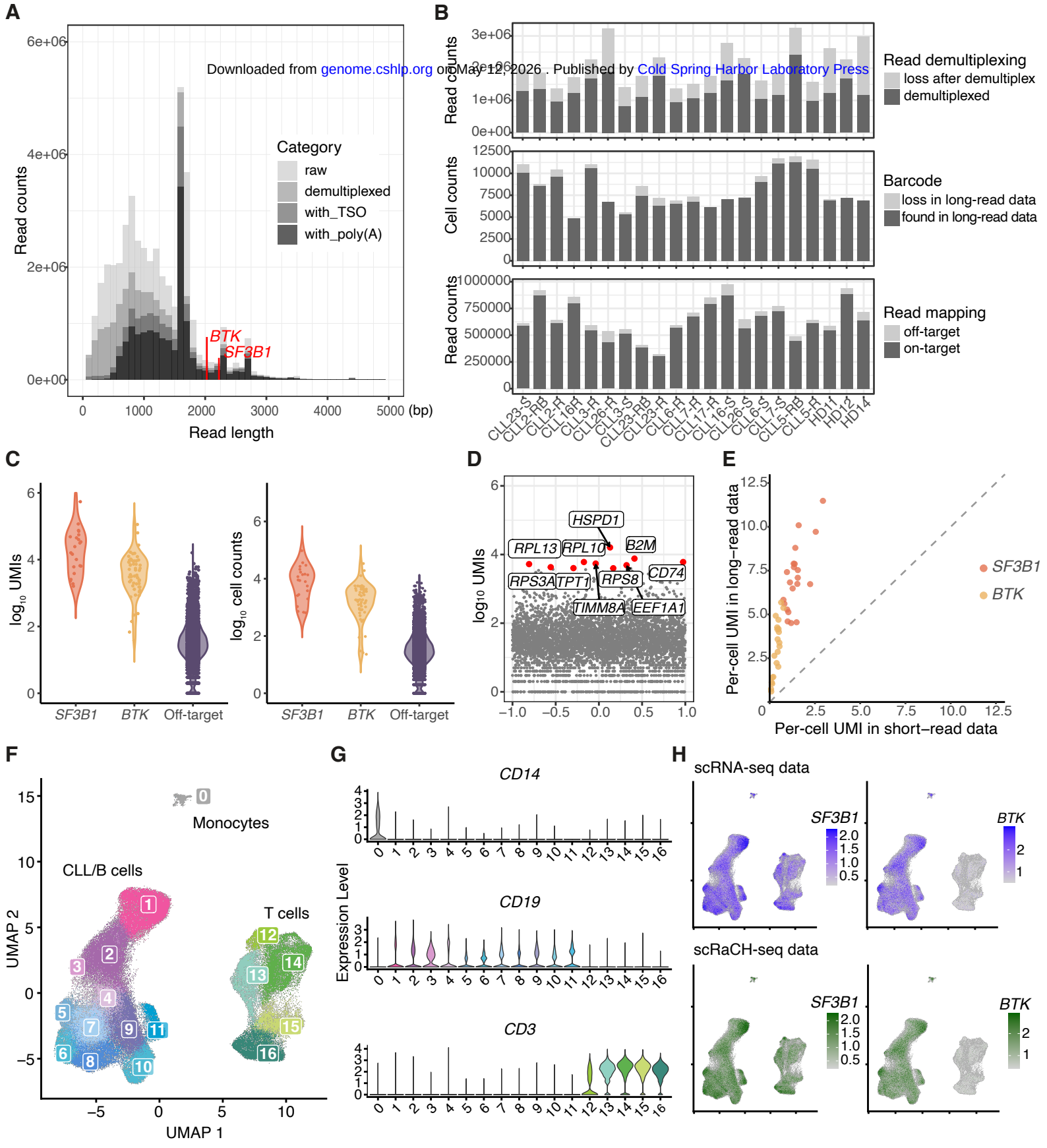
B) UMAP projection of cells carrying >2 SF3B1 KVR700**R altered transcripts (red).

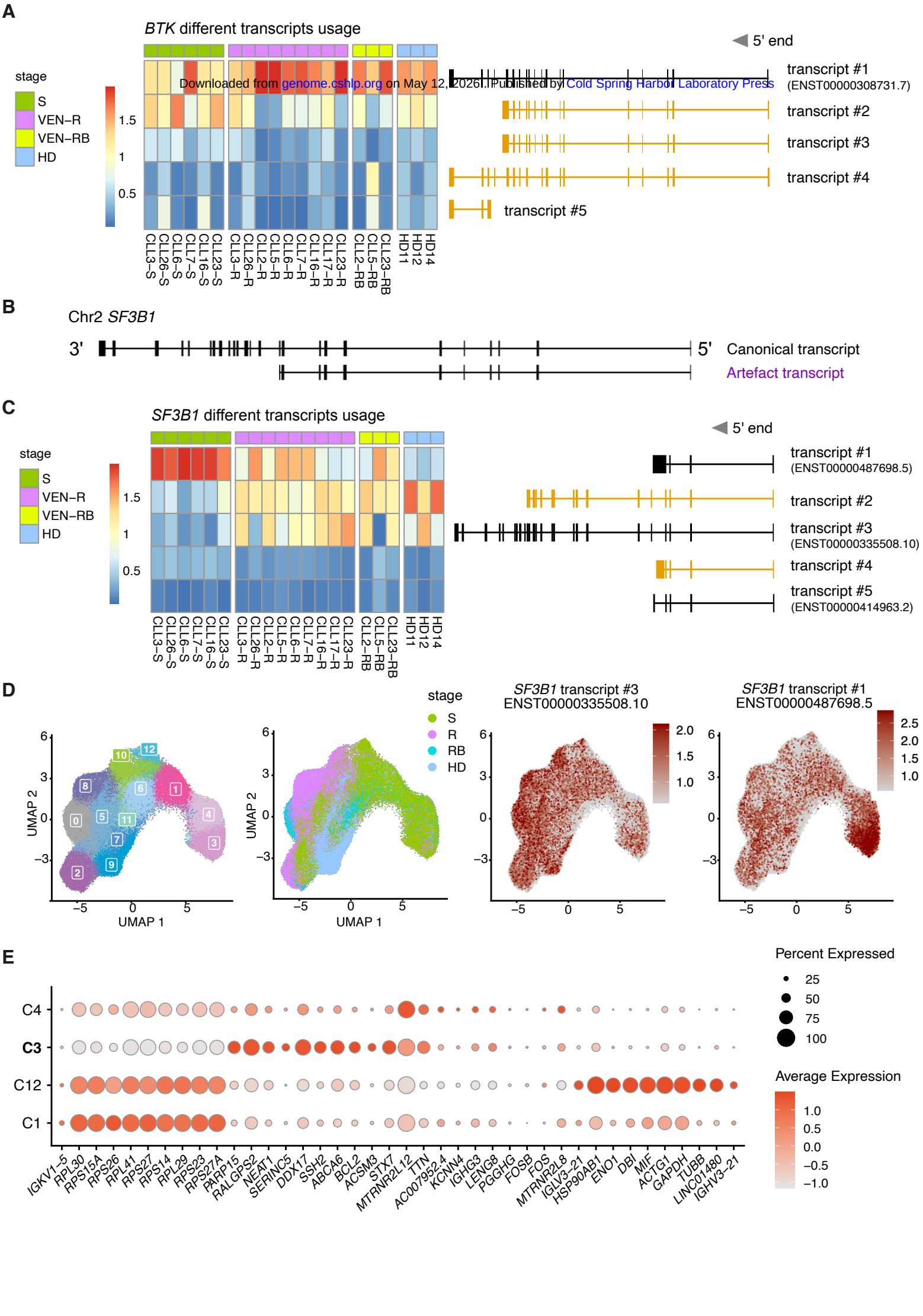
C) Volcano plot showing the differentially expressed genes (DEGs; FDR<0.05) between SF3B1 KVR700**R mutant and wild-type CLL cells from all venetoclax relapsed CLL samples.

D) Volcano plot showing the differentially expressed genes (DEGs; FDR<0.05) between SF3B1 KVR700**R mutant and wild-type CLL cells from CLL17.


Mutation
Probes
Exons
Barcode + UMI + TSO
Beads
Cells
Oil
Single cell droplet capture
Reverse transcription + cDNA amplification
Standard protocol
scRaCH-seq
Fragmentation
Whole transcriptome
Gene A
Gene B
Gene N
Short-read sequencing
Cell 1
Cell N
Enriched target gene
Long-read sequencing


Short-read data
Long-read data
Whole transcriptome
Gene-specific genotyping
scRNA-seq analysis
STEP 01 Assigning reads to cells
STEP 02 QC Check reads' integrity
PolyA
Gene body
TSO
UMI
Barcode
Pass
Pass
STEP 03 Mutation calling
Align reads to reference genome
Call mutations
Ref
STEP 04 Isoform detection
Create isoform assembly
Quantification
Re-alignment
QC-passed reads
STEP 05 Long-/short-read data integration
Marker genes
t-SNE2
Mut 1
Mut 2
Mut 3
transcript 1
transcript 2




A
BTK different transcripts usage
stage
S
VEN−R
VEN−RB
HD
1.5
1
0.5
CLL3−S
CLL26−S
CLL6−S
CLL7−S
CLL16−S
CLL23−S
CLL3−R
CLL26−R
CLL2−R
CLL5−R
CLL6−R
CLL7−R
CLL16−R
CLL17−R
CLL23−R
CLL2−RB
CLL5−RB
CLL23−RB
HD11
HD12
HD14
5' end
transcript #1
(ENST00000308731.7)
transcript #2
transcript #3
transcript #4
transcript #5
Downloaded from genome.cshlp.org on May 12, 2026 - Published by Cold Spring Harbor Laboratory Press
B
Chr2 SF3B1
3'
5'
Canonical transcript
Artefact transcript
C
SF3B1 different transcripts usage
stage
S
VEN−R
VEN−RB
HD
1.5
1
0.5
5' end
transcript #1
(ENST00000487698.5)
transcript #2
transcript #3
(ENST00000335508.10)
transcript #4
transcript #5
(ENST00000414963.2)
D
stage
S
R
RB
HD
UMAP 1
UMAP 2
SF3B1 transcript #3
ENST00000335508.10
SF3B1 transcript #1
ENST00000487698.5
2.0
1.5
1.0
2.5
2.0
1.5
1.0
E
Percent Expressed
25
50
75
100
Average Expression
1.0
0.5
0.0
−0.5
−1.0
C4
C3
C12
C1
IGKV1−5
RPL30
RPS15A
RPS26
RPL41
RPS27
RPS14
RPL29
RPS23
RPS27A
PARP15
RALGPS2
NEAT1
SERINC5
DDX17
SSH2
ABCA6
BCL2
ACSM3
STX7
MTRNR2L12
TTN
AC007952.4
KCNN4
IGHG3
LENG8
PGGHG
FOSB
FOS
MTRNR2L8
IGLV3−21
HSP90AB1
ENO1
DBI
MIF
ACTG1
GAPDH
TUBB
LINC01480
IGHV3−21