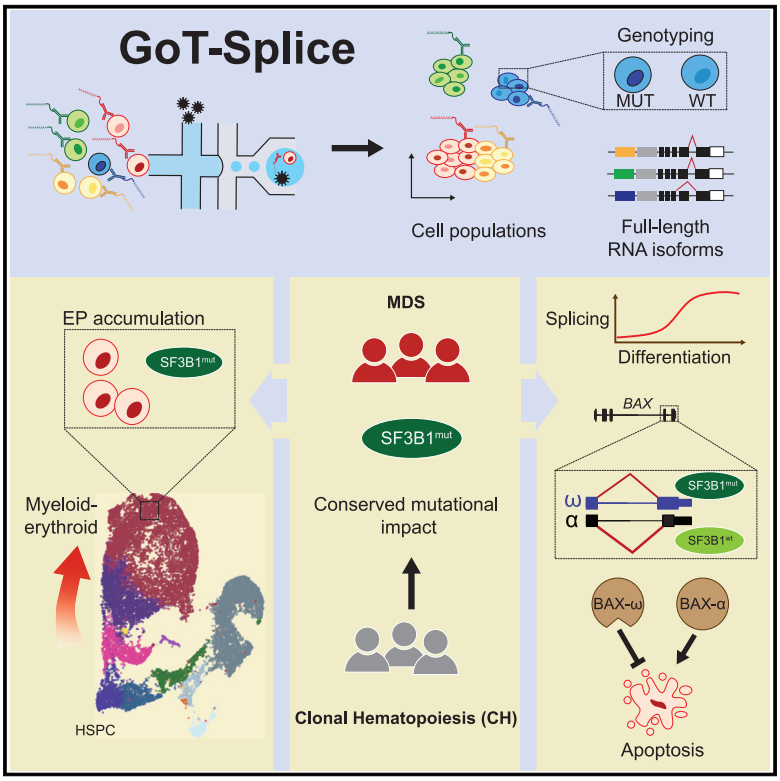
Cell Stem Cell

Resource

Single-cell multi-omics defines the cell-type-specific impact of splicing aberrations in human hematopoietic clonal outgrowths

Graphical abstract



Highlights

- GoT-Splice profiles single-cell genotype, expression, surface markers, and splicing
- *SF3B1*^{mut} cells show lineage skewing and stage-specific mis-splicing in MDS
- *SF3B1*^{mut}-specific isoform Bax-ω may provide MDS anti-apoptotic advantage
- In CH, *SF3B1*^{mut} cells display erythroid lineage bias and mirror MDS mis-splicing

Authors

Mariela Cortés-López, Paulina Chamely, Allegra G. Hawkins, ..., Omar Abdel-Wahab, Federico Gaiti, Dan A. Landau

Correspondence

federico.gaiti@uhn.ca (F.G.), dlandau@nygenome.org (D.A.L.)

In brief

Cortés-López and colleagues develop GoT-Splice for the concurrent profiling of gene expression, surface proteins, somatic mutations, and RNA splicing in individual cells. By utilizing this method, they investigate the effects of *SF3B1* mutations in patients with myelodysplastic syndrome and clonal hematopoiesis, unveiling splicing abnormalities that lead to lineage-specific clonal expansions.

Cortés-López et al., 2023, Cell Stem Cell *30*, 1262–1281
September 7, 2023
https://doi.org/10.1016/j.stem.2023.07.012

CellPress

CellPress

Cell Stem Cell

Resource

Single-cell multi-omics defines the cell-type-specific impact of splicing aberrations in human hematopoietic clonal outgrowths

Mariela Cortés-López,[1,2,19] Paulina Chamely,[1,2,19] Allegra G. Hawkins,[3,19] Robert F. Stanley,[4,19] Ariel D. Swett,[1,2] Saravanan Ganesan,[1,2] Tarek H. Mouhieddine,[5] Xiaoguang Dai,[6] Lloyd Kluegel,[1,2] Celine Chen,[1,2,7] Kiran Batta,[8] Nili Furer,[9] Rahul S. Vedula,[5] John Beaulaurier,[10] Alexander W. Drong,[6] Scott Hickey,[10] Neville Dusaj,[1,2,7] Gavriel Mullokandov,[1,2] Adam M. Stasiw,[1,2] Jiayu Su,[1,11] Ronan Chaligné,[12] Sissel Juul,[6] Eoghan Harrington,[6] David A. Knowles,[1,11,13] Catherine J. Potenski,[1,2] Daniel H. Wiseman,[8] Amos Tanay,[14] Liran Shlush,[9] Robert C. Lindsley,[5] Irene M. Ghobrial,[5] Justin Taylor,[15] Omar Abdel-Wahab,[4] Federico Gaiti,[16,17,20,*] and Dan A. Landau[1,2,18,20,21,*]

[1]New York Genome Center, New York, NY, USA
[2]Division of Hematology and Medical Oncology, Department of Medicine and Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA
[3]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA
[4]Molecular Pharmacology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[5]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[6]Oxford Nanopore Technologies Inc., New York, NY, USA
[7]Tri-Institutional MD-PhD Program, Weill Cornell Medicine, Rockefeller University, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[8]Division of Cancer Sciences, The University of Manchester, Manchester, UK
[9]Weizmann Institute of Science, Department of Molecular Cell Biology, Rehovot, Israel
[10]Oxford Nanopore Technologies Inc., San Francisco, CA, USA
[11]Department of Systems Biology, Columbia University, New York, NY, USA
[12]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[13]Department of Computer Science, Columbia University, New York, NY, USA
[14]Weizmann Institute of Science, Department of Computer Science and Applied Mathematics, Rehovot, Israel
[15]Sylvester Comprehensive Cancer Center, University of Miami, Miller School of Medicine, Miami, FL, USA
[16]University Health Network, Princess Margaret Cancer Centre, Toronto, ON, Canada
[17]University of Toronto, Medical Biophysics, Toronto, ON, Canada
[18]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
[19]These authors contributed equally
[20]Senior author
[21]Lead contact
*Correspondence: federico.gaiti@uhn.ca (F.G.), dlandau@nygenome.org (D.A.L.)
https://doi.org/10.1016/j.stem.2023.07.012

SUMMARY

RNA splicing factors are recurrently mutated in clonal blood disorders, but the impact of dysregulated splicing in hematopoiesis remains unclear. To overcome technical limitations, we integrated genotyping of transcriptomes (GoT) with long-read single-cell transcriptomics and proteogenomics for single-cell profiling of transcriptomes, surface proteins, somatic mutations, and RNA splicing (GoT-Splice). We applied GoT-Splice to hematopoietic progenitors from myelodysplastic syndrome (MDS) patients with mutations in the core splicing factor *SF3B1*. *SF3B1*^{mut} cells were enriched in the megakaryocytic-erythroid lineage, with expansion of *SF3B1*^{mut} erythroid progenitor cells. We uncovered distinct cryptic 3′ splice site usage in different progenitor populations and stage-specific aberrant splicing during erythroid differentiation. Profiling *SF3B1*-mutated clonal hematopoiesis samples revealed that erythroid bias and cell-type-specific cryptic 3′ splice site usage in *SF3B1*^{mut} cells precede overt MDS. Collectively, GoT-Splice defines the cell-type-specific impact of somatic mutations on RNA splicing, from early clonal outgrowths to overt neoplasia, directly in human samples.

INTRODUCTION

Genetic diversity in the form of clonal outgrowths has been ubiquitously observed across normal and malignant human tissues.[1–13] Likewise, single-cell RNA sequencing (scRNA-seq) has revealed phenotypic diversity as a hallmark of both normal and malignant human tissues.[14–20] These two axes of cellular diversity likely exhibit complex interplay, as cell state may affect

the phenotypic impact of somatic mutations.[21] Recent advances in single-cell multi-omics sequencing have allowed us to link genetic variation and transcriptional cell state diversity in somatic evolution of human tissues.[15,22,23] For example, using genotyping of transcriptomes (GoT)[15] technology, which enables genotyping of somatic mutations together with high-throughput droplet-based scRNA-seq, we demonstrated that the effects of somatic mutations on cellular fitness in myeloproliferative disorders vary as a function of progenitor cell identity.[15]

Mutations in genes encoding RNA splicing factors demonstrate the challenge of linking genotype to phenotype in complex human tissues. Somatic change-of-function mutations in RNA splicing factors are recurrent in hematologic malignancies,[24–26] highlighting the importance of dysregulated RNA splicing in human hematopoietic disorders. *SF3B1* (splicing factor 3b subunit 1), a core component of the spliceosome complex, is a commonly mutated splicing factor across hematologic malignancies and solid tumors, and is implicated in the pathogenesis of myelodysplastic syndromes (MDSs).[27,28] *SF3B1* mutations also increase the risk of myeloid neoplasms in individuals with clonal hematopoiesis (CH), compared with other CH driver mutations.[1,2] *SF3B1* mutations result in incorrect branch point recognition during RNA splicing, often leading to an increased usage of aberrant (or cryptic) intron-proximal 3′ splice sites in hundreds of genes.[29] Such aberrant 3′ splice site recognition typically results in the inclusion of short intronic fragments in spliced mRNA, commonly causing frameshifts that render the transcript a substrate for nonsense mediated mRNA decay (NMD).[30] Through mis-splicing, *SF3B1* mutations have been shown to affect cell metabolism[31] and ribosomal biogenesis,[32] leading to the aberrant hematopoietic differentiation typical of MDS. However, the mechanisms through which mis-splicing disrupts hematopoietic differentiation in humans remain elusive.

To date, cellular and murine models have been critical for elucidating the role of splicing factor mutations in disordered hematopoiesis. Nonetheless, these methods may not fully recapitulate MDS development in humans. For example, alternatively spliced genes from murine models of *SF3B1*^{mut} MDS, which phenotypically resemble human MDS, show limited overlap with those identified in humans.[33] Analysis of splice-altering mutations in humans has been further hampered by three main limitations. First, normal wild-type (WT) and aberrant mutated (MUT) cells are often admixed, limiting identification of signals specifically linked to the *SF3B1*^{mut} genotype. This challenge is amplified in the context of CH, where MUT cells are typically a minority of the hematopoietic progenitor population. Second, the hematopoietic differentiation process yields significant complexity of cell progenitor types that further hinders the ability to link mutated genotypes with distinct cellular phenotypes. *SF3B1*^{mut} MDS is indeed associated with a specific clinico-morphological phenotype of refractory anemia and accumulation of ringed sideroblasts,[28,34] strongly suggesting that the interplay between cell identity and *SF3B1* mutations is fundamental in driving disrupted hematopoietic differentiation. Finally, scRNA-seq by 3′ or 5′ biased short-read sequencing provides an incomplete picture of the consequences of splicing factor mutations on the transcriptome and their downstream effects.

To overcome these limitations and identify cell-identity-dependent mis-splicing mediated by *SF3B1* mutations, we developed GoT-Splice by integrating GoT[15] with long-read single-cell transcriptome profiling (with Oxford Nanopore Technologies [ONT]) as well as proteogenomics (with Cellular Indexing of Transcriptomes and Epitopes by Sequencing [CITE-seq]).[35] This enabled the simultaneous profiling of gene expression, cell surface protein markers, somatic mutation genotyping, and RNA splicing within the same single cell. The application of GoT-Splice to bone marrow samples from individuals with *SF3B1*^{mut} MDS and CH revealed that, while *SF3B1* mutations arise in uncommitted hematopoietic stem progenitor cells (HSPCs), their enrichment increases along the differentiation trajectory into committed erythroid progenitors (EPs), in line with the *SF3B1*^{mut}-driven dyserythropoiesis phenotype. Importantly, the integration of GoT with full-length isoform mapping via long-read sequencing demonstrated that *SF3B1* mutations exert cell-type-specific mis-splicing, already apparent in CH.

RESULTS

GoT integrated with proteogenomics reveals enrichment of *SF3B1*^{mut} cells in the erythroid lineage linked to overexpression of cell-cycle and mRNA translation genes

As the impact of somatic mutations on the transcriptome varies as a function of underlying cell identity in myeloproliferative neoplasms,[15] we hypothesized that an interplay between cell identity and *SF3B1* mutations may drive disrupted hematopoietic differentiation in MDS. To test this, we applied GoT[15] (Figure 1A) to CD34+ bone marrow progenitor cells from three untreated MDS patients with *SF3B1* K700E mutations (discovery cohort, MDS01–03) and a separate cohort of MDS patients undergoing treatment (validation cohort, MDS04–06) with erythropoietin (EPO) and/or granulocyte colony-stimulating factor (G-CSF; Figure 1B; Table S1). As normal hematopoietic development has been extensively studied using flow cytometry cell surface markers, we further integrated GoT with single-cell proteogenomics (CITE-seq[35,36]; Figure 1A). After sequencing and quality control filtering, we obtained 24,315 cells across the six MDS samples (Figures S1A and S1B; MDS02 was sequenced in two technical replicates). To chart the differentiation map of the CD34+ progenitor cells, we integrated the data across the six MDS samples and clustered based on transcriptomic data alone, agnostic to the genotyping and protein information (Figures 1C, S1C, and S1D). Using previously annotated RNA identity markers for human CD34+ progenitor cells,[37] validated via antibody-derived tag (ADT) markers in the CITE-seq panel (Tables S1 and S2), we identified the expected progenitor subtypes in the primary MDS cohort, along with a population of mature monocytic cells characterized by CD14 expression and lack of CD34 expression, which is often observed in CD34+ sorting of human bone marrow,[38,39] and was not completely removed with monocyte-specific blocking reagents[39] (Figures 1C and S1E–S1G). Cell clustering was further validated using RNA and ADT multimodal integration (Figure S1H). The expected progenitor subtypes were similarly identified in the MDS validation cohort (MDS04–06; Figures S1I–S1K).

Genotyping data were available for 15,650 MDS cells (64.4% across MDS01–06) through GoT (Figures 1B and S2A–S2D). The per-patient mutant cell fractions obtained through GoT

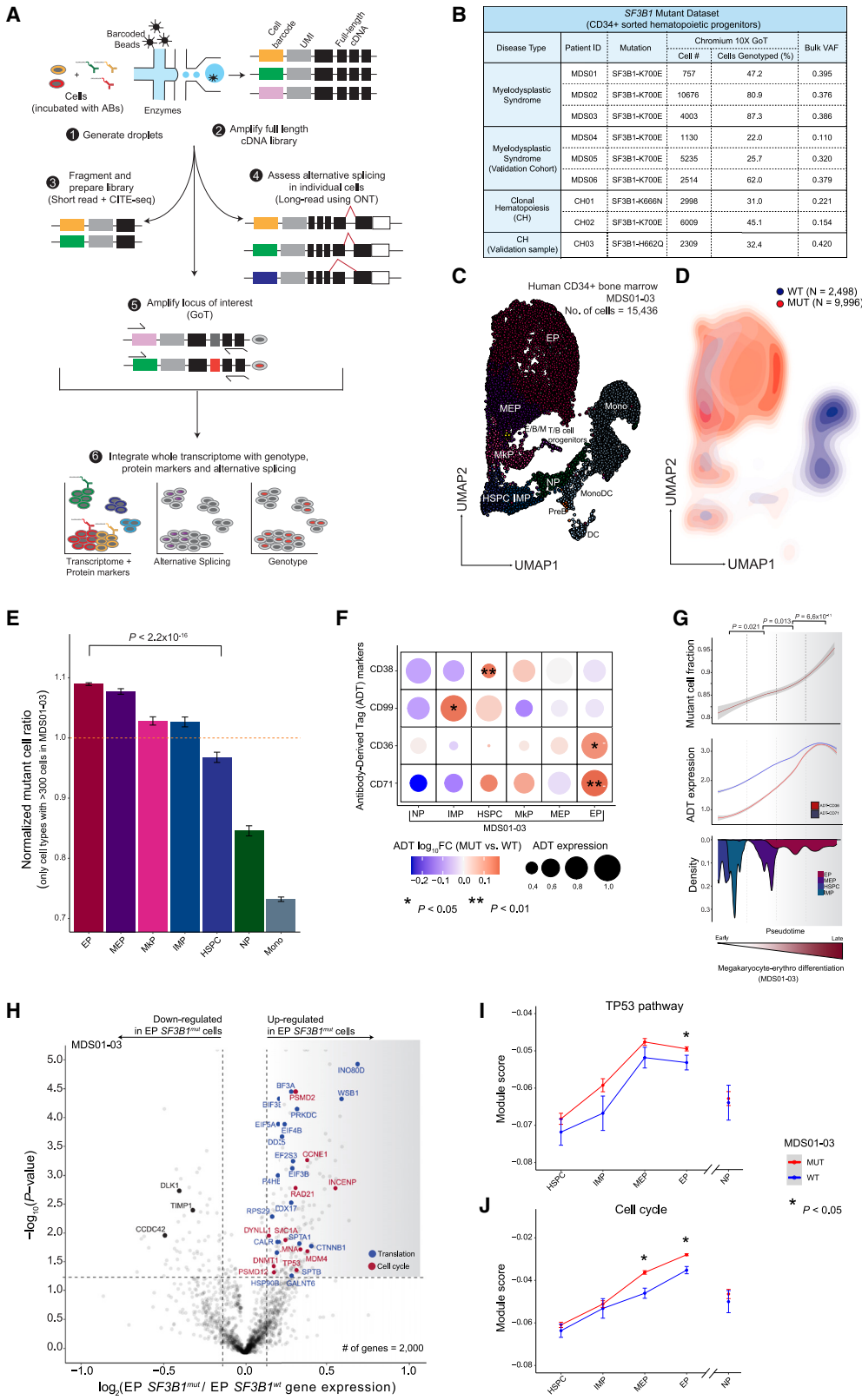


Figure 1. Enrichment of *SF3B1*^{mut} cells in the megakaryocytic-erythroid lineage

(A) GoT-Splice workflow combines GoT with CITE-seq and long-read full-length cDNA using ONT for the simultaneous single-cell profiling of protein and gene expression, somatic mutations, and alternative splicing.

(legend continued on next page)

were highly correlated with the variant allele frequencies (VAFs) obtained through bulk sequencing of matched unsorted peripheral blood mononuclear cells (Pearson's r = 0.81, p value = 0.008; Figure S2A). Projection of the genotyping information onto the differentiation map showed co-mingling of MUT and WT cells throughout the differentiation topology, highlighting the need for single-cell multi-omics to link genotypes with cellular phenotypes in *SF3B1*^{mut} MDS. Although MUT cells were found across CD34+ progenitor cells, we observed an accumulation of MUT cells along the erythroid trajectory (Figure 1D), suggesting that *SF3B1* mutant cell frequency (MCF) varies as a function of the progenitor subtype. To confirm this, we evaluated the MCF across the different prevalent progenitor cell types (limited to progenitor subsets with > 300 cells). Across samples, we observed a significant increase in MCF in the megakaryocyte-erythroid lineage, with the highest MCF in EPs compared with HSPCs (p value < 10^{-16}; Figures 1E and S2D), consistent with the erythroid lineage-specific impact of mutated *SF3B1*.[40,41]

The ability to layer protein measurements on top of GoT data further allowed us to identify differentially expressed proteins between MUT and WT cells within each progenitor subset. After quality control filtering for ADT markers with adequate expression in at least two major progenitor subtypes (see STAR Methods), protein expression was highest in the expected cell types and correlated with mRNA expression at the cell and cell-type level, comparable to previous data[35] (Figures S2E and S2F). We directly compared protein expression between MUT and WT cells, accounting for sample-to-sample variability in MUT cells through downsampling (see STAR Methods) and observed differential expression of CD38, CD99, CD36, and CD71 in at least one progenitor cell type (Figure 1F; Table S2). CD38 is a known marker for the transition of primitive CD34+ stem and progenitor cells into more committed precursor cells.[37,42–44] Its overexpression in *SF3B1*^{mut} is consistent with the observed higher MCF in committed progenitor subsets. CD99, overexpressed in MUT immature myeloid progenitor (IMP) cells, was previously noted to be overexpressed in both AML and MDS stem cells, serving as a potential therapeutic target of malignant stem cells.[45,46] Finally, CD36 and CD71, erythroid lineage markers, were found to be overexpressed in MUT EPs when compared with WT EPs, consistent with the *SF3B1*^{mut}-driven dyserythropoiesis phenotype. We further leveraged these erythroid maturation cell surface protein markers to validate pseudotemporal (pseudotime) ordering of the continuous process of erythroid maturation[47] (Figure S2G). This analysis revealed an increase in MCF along erythroid lineage maturation (Figure 1G), confirming enrichment of *SF3B1*^{mut} cells along the differentiation trajectory into committed EPs.

To further explore *SF3B1* driven transcriptional dysregulation in committed EPs, we performed differential gene expression analysis between *SF3B1*^{mut} and *SF3B1*^{wt} cells. Mutated EPs upregulated genes encoding important translation and ribosome biogenesis factors (Figure 1H; Table S3), including several eukaryotic initiation factors (e.g., *EIF3A* [false discovery rate (FDR)-adjusted p value = 0.007], *EIF5A* [FDR-adjusted p value = 0.011]), DEAD-box helicases (e.g., *DDX5* [FDR-adjusted p value = 0.016]), and ribosome subunits (e.g., *RPS29* [FDR-adjusted p value = 0.1]). While we did not directly assess translational defects, this dysregulation of translation factor expression is evocative of studies showing that translational regulation is critical during hematopoiesis,[48–51] and may lead to cell- and tissue-type-restricted activation of *TP53* signaling pathway in myeloid disease.[52–57] Specifically, cells that require high levels of protein synthesis, such as EPs, may be more sensitive to even subtle changes in translational regulation.[58] In line with this notion, *TP53* gene target upregulation in *SF3B1*^{mut} cells was more prominent in the megakaryocyte-erythroid lineage, with no increased expression of *TP53*-related genes in earlier progenitors (HSPCs) or in neutrophil progenitors (NPs) compared with WT cells (Figure 1I). Our results therefore establish a molecular phenotype for the *SF3B1* mutation in human bone marrow progenitors, implicating changes in translation pathway genes.

Mutated EPs also upregulated genes related to the cell cycle (FDR-adjusted p value = 0.08; Figures 1H and 1J; Table S3). For example, we observed an increase in the expression of *CCNE1*, a positive regulator of the G1/S transition of the cell cycle,[59] and *MDM4*, which works together with *TP53* during the G1/S checkpoint to determine the fate of cells by regulating pathways involved in DNA repair, apoptosis, and senescence.[60,61] Increased expression of *MDM4* can attenuate *TP53* activation induced by ribosomal stress,[62,63] thereby reducing

(B) Patient metadata and quality controlled GoT data for *SF3B1*-mutant MDS and CH samples.
(C) Uniform manifold approximation and projection (UMAP) of CD34+ cells (n = 15,436 cells) from MDS patients with *SF3B1* K700E mutations (n = 3 individuals), overlaid with cluster cell-type assignments. HSPCs, hematopoietic stem progenitor cells; IMPs, immature myeloid progenitors; MkPs, megakaryocytic progenitors; MEPs, megakaryocytic-erythroid progenitors; EPs, erythroid progenitors; NPs, neutrophil progenitors; E/B/M, eosinophil/basophil/mast progenitor cells; T/B cell progenitors; Mono, monocyte; DCs, dendritic cells; Pre-B, precursors B cells; Mono DC, monocyte/dendritic cell progenitors.
(D) Density plot of *SF3B1*^{mut} vs. *SF3B1*^{wt} cells, with genotypes (MDS01–03) for 12,494 cells (80.9% of all cells).
(E) Normalized frequency of *SF3B1*^{mut} cells in progenitor subsets with at least 300 genotyped cells. Bars show analysis of MDS01–03 with mean ± SEM of 100 downsampling iterations to 1 genotyping UMI per cell. Cell types with >300 cells were analyzed. p value from likelihood ratio test of linear mixed model with or without mutation status.
(F) Differential ADT marker expression between *SF3B1*^{mut} and *SF3B1*^{wt} cells. Red, higher expression in *SF3B1*^{mut} cells; blue, higher expression in *SF3B1*^{wt} cells. Dot size corresponds to the average ADT expression across cells in each cell type. p values determined through permutation testing.
(G) Mutant cell fraction and ADT expression levels of CD36 and CD71 as a function of pseudotime along the megakaryocyte-erythroid differentiation trajectory for *SF3B1*^{mut} and *SF3B1*^{wt} cells in MDS01–03. Shading denotes 95% confidence interval. Histogram shows cell density of analyzed clusters, ordered by pseudotime. p values were calculated by Wilcoxon rank-sum test by comparing mutant cell fraction between pseudotime trajectory quartiles.
(H) Differential gene expression between *SF3B1*^{mut} and *SF3B1*^{wt} EP cells in MDS samples. Genes with an absolute $\log_2$(fold change) > 0.1 and p value < 0.05 were defined as differentially expressed (DE). Cell cycle (red) and translation (blue) pathways (Reactome) are highlighted.
(I) Expression (mean ± SEM) of TP53 pathway related genes (Reactome) between *SF3B1*^{mut} and *SF3B1*^{wt} cells in progenitor cells from MDS01–03 samples. Red, module score in *SF3B1*^{mut} cells; blue, module score in *SF3B1*^{wt} cells. p values from likelihood ratio test of linear mixed model with or without mutation status.
(J) Same as (I) for expression of cell cycle related genes (Reactome) between *SF3B1*^{mut} and *SF3B1*^{wt} cells in progenitor cells from MDS01–03 samples.

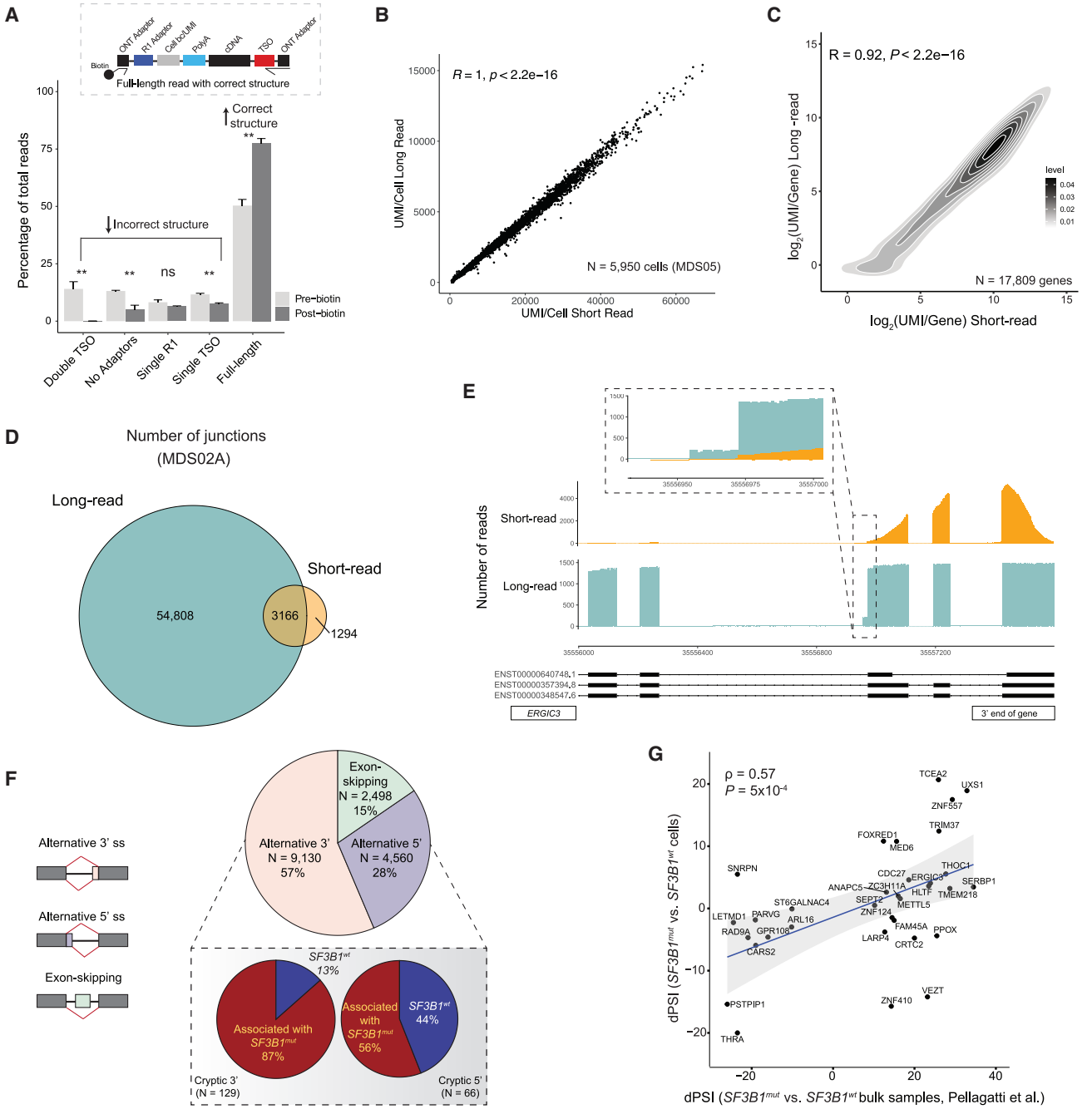


Figure 2. Simultaneous profiling of gene expression, cell surface protein markers, somatic mutation status, and alternative splicing at single-cell resolution

(A) Comparison of the percentage of ONT reads with either incorrect structure (double TSO, no adaptors, single R1 or single TSO) or correct structure (full-length reads) both before and after the inclusion of a biotin enrichment protocol step during preparation for sequencing. Bars show the aggregate analysis of n = 5 samples with mean ± SD of the percentage for each category.

(B) Scatterplot of the correlation between the number of UMIs/cell detected in long-read ONT vs. short-read Illumina data for cells sequenced across both platforms for sample MDS05.

(C) Density plot of the correlation between the number of UMIs/gene detected in long-read ONT vs. short-read Illumina data for sample MDS05.

(D) Number of splice junctions captured in the full-length long-read ONT data compared with short-read sequencing data (gene coverage ≥ 10 in both sequencing protocols, junction cluster coverage ≥ 600 and junction read support ≥ 1 read [see STAR Methods]), demonstrating increased junction capture with GoT-Splice across cells.

(E) Greater sequencing coverage uniformity of GoT-Splice compared with short-read sequencing over splice junctions, illustrated with the *ERGIC3* gene.

(legend continued on next page)

the functional impact of p53, thus promoting cell survival and accumulation. Taken together, these factors contribute to the enrichment of *SF3B1* mutations in the erythroid lineage.

GoT-Splice links somatic mutations, alternative splicing, and cellular phenotype at single-cell resolution

Given the pivotal role of *SF3B1* in mRNA splicing, we next explored how mis-splicing may link genotypes and cellular phenotypes. Indeed, *SF3B1* mutations promote recognition of alternative branch points, most often leading to increased usage of aberrant 3′ splice sites.[29] However, previous studies in primary human samples have been performed on bulk samples admixing MUT and WT cells as well as progenitor subtypes.[30,32,64,65] Conversely, short-read scRNA-seq does not adequately cover splice junctions. Recent advances suggest that long-read integration into scRNA-seq may overcome these limitations.[66–70] We therefore integrated GoT with full-length ONT long-read sequencing, allowing for high-throughput, single-cell integration of genotype, cell surface proteome, gene expression, and mRNA splicing information (GoT-Splice; Figure 1A). Single-cell cDNA sequencing with ONT presents unique challenges, as cDNA amplification artifacts are still productively sequenced when using standard ONT ligation chemistry, leading to a high fraction of uninformative reads in the highly amplified single-cell libraries. To enhance ONT efficiency, we incorporated a biotin enrichment step using on-bead PCR to selectively amplify full-length reads containing intact cell barcodes and unique molecular identifiers[67] (UMIs; Figure 2A), increasing the yield of full-length reads from 50.4% ± 2.7% to 77.6% ± 2.0% (mean ± SD) of sequenced reads. Thus, GoT-Splice delivers high-resolution single-cell full-length transcriptional profiles comparable to short-read sequencing (Figures 2B and 2C). However, full-length ONT sequencing alone is insufficient to support efficient genotyping of the $SF3B1^{mut}$ locus, as analysis of ONT-alone data revealed that only 3% of cells have at least 1 UMI covering the $SF3B1^{mut}$ locus vs. 56% of cells with GoT (average across MDS samples; Figure S2B), demonstrating the power of our integrated approach.

To accurately identify splice junctions using single-cell long-read sequencing, we developed an analytical pipeline that leverages the SiCeLoRe (Single Cell Long Read) pipeline[67] (Figure S3A). To reduce alignment noise, we generated a splice junction reference identified in single-cell Smart-Seq2 data from human CD34+ cells without *SF3B1* mutation (STAR Methods). Next, we performed intron-centric junction calling for the independent measurement of splicing at both the 5′ and 3′ ends of each intron. This allows for unbiased assessment of junctions and greater accuracy in measuring the degree of transcript mis-splicing compared with exon-centric quantification approaches,[71] which are typically used for cassette exon usage profiling and rely on potentially inaccurate or incomplete[72,73] predefined transcript models or splicing events. As anticipated, when comparing short-read and long-read sequencing, we found a 12.3-fold increase in the number of junctions detected using long-read sequencing, with the majority of junctions (~90%) unique to long-read data (Figures 2D and S3B). Notably, at the single-cell level, despite lower absolute number of UMI/cell, we observed a 5.5-fold increase in detected junctions with long-read sequencing (Figure S3C). Additionally, GoT-Splice afforded greater coverage uniformity across the entire transcript, compared with 3′-biased coverage in short-read sequencing, enabling the detection of splicing events further from the 3′ transcript end (Figure 2E). To further highlight the discovery power of long-read sequencing, we compared short- and long-read capture of cryptic 3′ splicing events in relation to distance from the 3′ transcript end (STAR Methods), showing that long-read sequencing identifies substantially greater numbers of cryptic events both along the length of the transcript and overall compared with short-read (Figure S3D), with the majority of cryptic 3′ splicing events detected in short-reads also captured in long reads (Figure S3D, pie chart inset).

The most common mis-splicing events (57%) observed in MDS $SF3B1^{mut}$ cells involved alternative 3′ splice sites (Figure 2F), consistent with prior reports.[29,74] Notably, such alternative 3′ splice site usage was not observed in a $SF3B1^{wt}$ CD34+ sample (Figure S3E). Among the differentially mis-spliced cryptic 3′ splice sites (0–100 bp from canonical splice site) between $SF3B1^{mut}$ and $SF3B1^{wt}$ cells, 87% were used more highly in $SF3B1^{mut}$ cells (Figure 2F, inset), aligning with known characteristics of *SF3B1* mutations. ONT long-read sequencing also allowed quantification of different splicing events within the same mRNA transcript. While only one aberrant 3′ splice site event was observed for most mRNA transcripts, we identified 428 genes (21.4% of total genes with at least one cryptic 3′ splice site) with more than one aberrant 3′ splice site event. These cryptic 3′ splicing events frequently appeared in different copies of the transcript (Figure S3F). Consistent with previous MDS bulk sequencing data,[29,75] we observed a relative enrichment of purines upstream of the aberrant 3′ splice site when compared with the canonical 3′ splice site (Figure S3G).

To assess how our method compares with available isoform detection tools, we compared the recovery and quantification of novel splicing junctions with full-length isoform quantification methods FLAMES[68] and IsoQuant.[76] While isoform junction annotation was largely comparable across all three methods (Figure S4A), detection of cryptic 3′ events was increased using our approach, compared with FLAMES and IsoQuant, which are designed to annotate and quantify full-length isoforms. For these cryptic 3′ events, we observed more variation on local splicing event assignment, with little agreement between the three methods (Figure S4A, right). Additionally, many 3′ cryptic events detected by our method were not identified by FLAMES and IsoQuant (Figure S4A, right). We also observed a high correlation between GoT-Splice delta percent spliced in (dPSI; percent spliced in) measurements obtained by comparing $SF3B1^{mut}$

(F) Pie chart summarizing the distribution of different alternative splicing events detected after junction annotation. Inset: differences in the usage of cryptic 3′ and 5′ splice site events between $SF3B1^{mut}$ and $SF3B1^{wt}$ cells measured with a dPSI ($SF3B1^{mut}$ PSI-$SF3B1^{wt}$ PSI). Associated with $SF3B1^{mut}$: +ve dPSI; associated with $SF3B1^{wt}$: −ve dPSI.

(G) Comparison of dPSI values for shared cryptic 3′ splicing events identified in the MUT vs. WT cell comparison from GoT-Splice of $SF3B1^{mut}$ MDS01–03 samples and in the $SF3B1^{mut}$ vs. $SF3B1^{wt}$ bulk comparison from bulk RNA sequencing of CD34+ cells of MDS samples in Pellagatti et al.[32] Correlation coefficient ρ calculated using Spearman's correlation and p value derived from two-tailed Student's t-test.

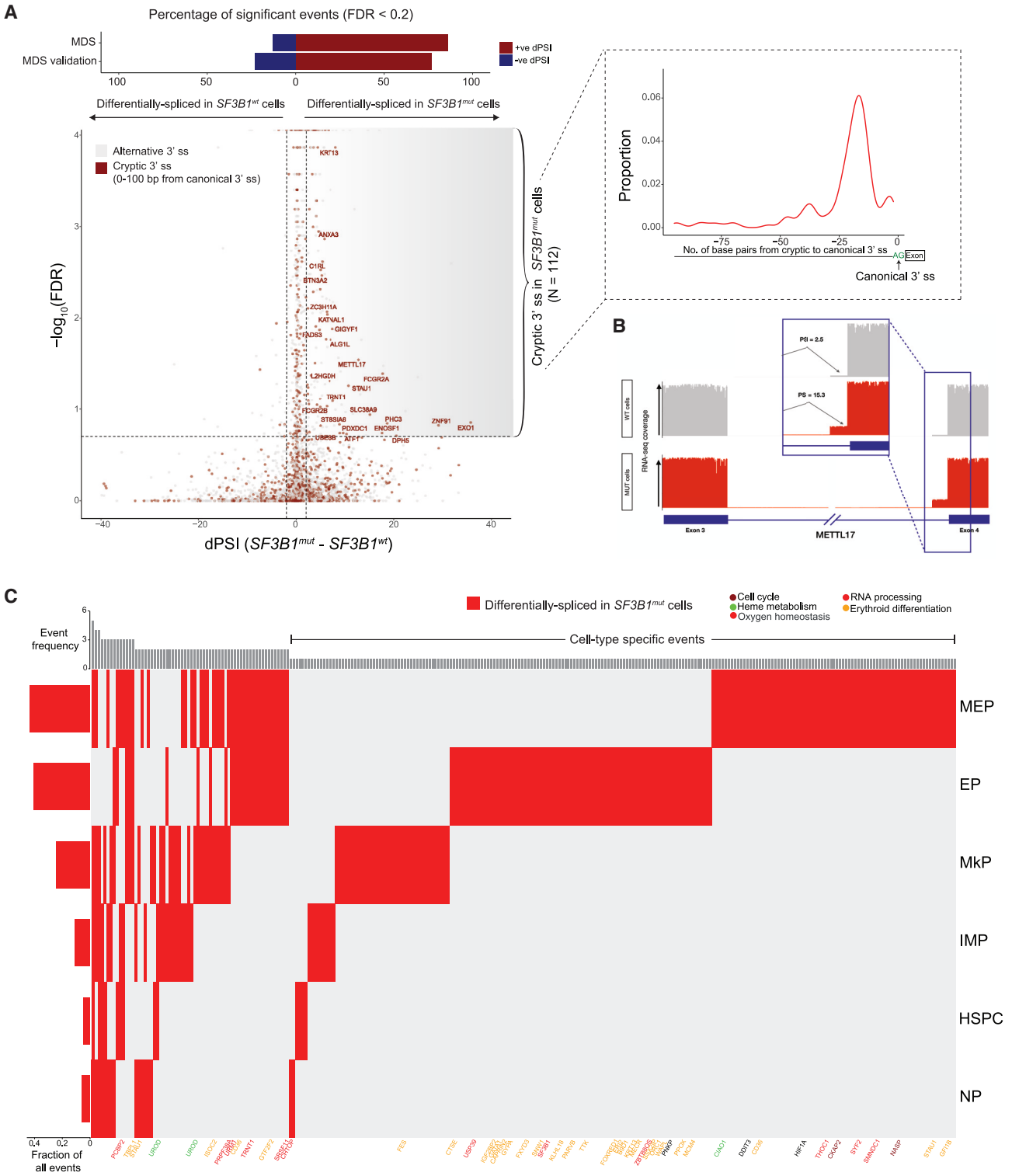


Figure 3. Progenitor cell-type-specific mis-splicing in *SF3B1*^{mut} MDS

(A) Differential splicing analysis between *SF3B1*^{mut} and *SF3B1*^{wt} cells across MDS samples. Junctions with absolute dPSI > 2 and FDR-adjusted p value < 0.2 were defined as differentially spliced. Top: bars showing percentage of genes differentially spliced in *SF3B1*^{mut} and *SF3B1*^{wt} cells in MDS and MDS validation cohorts. Inset: expected peak in the number of identified cryptic 3′ splice sites at 15–20 base pairs upstream of the canonical 3′ splice site in *SF3B1*^{mut} cells. (B) Sashimi plot of *METTL17* intron junction with an *SF3B1*^{mut} associated cryptic 3′ splice site showing RNA-seq coverage in *SF3B1*^{mut} vs. *SF3B1*^{wt} cells within MDS samples. Inset: expected increase in PSI value for the usage of this cryptic 3′ splice site in *SF3B1*^{mut} cells.

(legend continued on next page)

and *SF3B1*^{wt} cells and dPSI derived from bulk RNA-seq of CD34^{+} cells from *SF3B1*^{mut} vs. *SF3B1*^{wt} MDS samples[32] for shared cryptic 3′ splice sites (Figure 2G). This correlation with dPSI derived from bulk data of *SF3B1*^{mut} MDS samples[32] was not statistically significant in IsoQuant; in contrast, FLAMES preserved the positive correlation observed when using GoT-Splice (Figure 2G) although with a smaller Spearman's correlation coefficient value compared with GoT-Splice (Figure S4B). Furthermore, in line with previous work in MDS, the majority of the cryptic 3′ splice sites identified by GoT-Splice were ~15–20 bp upstream of the canonical 3′ site[29] (Figures 3A, 3B, and S4C–S4F). The dPSI for the differentially spliced cryptic 3′ splice site events obtained by comparing *SF3B1*^{mut} and *SF3B1*^{wt} cells were highly correlated across MDS samples (average Pearson's r of 0.55; p value < 0.001), highlighting the ability of our differential splicing analysis pipeline to identify statistically robust recurrent mis-splicing events (Figures 3A, 3B, and S4C–S4F; Table S4). Unlike GoT-Splice, which identified a far larger number of mis-splicing events validated with manual review, full-length isoform quantification methods did not demonstrate the expected increased usage of cryptic 3′ splice sites in MUT vs. WT cells (Figures 3A, 3B, and S4G), with equal number of statistically significant mis-splicing events, further suggesting lower performance in identifying 3′ cryptic mis-splicing. This is potentially due to the applied correction of splice sites in these full-length isoform algorithms, which may hinder cryptic mis-splicing detection, particularly for cryptic splice sites found within 10 bp of the canonical site (Figure S4H).

To assess how the *SF3B1*^{mut} MDS alternative splicing profiles compare to another hematologic malignancy, we compared cryptic 3′ events detected from S*F3B1*^{mut} MDS cells (positive dPSI, FDR-adjusted p value < 0.2) to those detected using previously published bulk RNA-seq data from *SF3B1*^{mut} chronic lymphocytic leukemia (CLL) samples.[77] We detected more cryptic 3′ events in our MDS dataset, with 10 events overlapping between MDS and CLL (Figure S4I; Table S4). This low overlap is likely driven by expression differences between MDS and CLL cells, as genes with shared events had higher expression levels in MDS cells compared with CLL-only events (Figure S4J).

To demonstrate GoT-Splice's generalizability for profiling somatic mutations, we analyzed a *DNMT3A*^{mut} CH sample (R882C; VAF of 0.09). Recent work has implicated DNMT3A in splicing regulation in hematopoiesis, independent of DNA methylation.[78] We quantified the distribution of alternative splicing patterns and found that exon skipping was the most common event (Figure S5A), as previously reported.[79] Comparison between *DNMT3A*^{mut} and *DNMT3A*^{wt} cells revealed genotype-specific events including the *SRSF3* exon 4 skipping event,[80] exclusive to *DNMT3A*^{mut} cells (Figures S5B and S5C). *SRSF3* exon 4 harbors a premature termination codon (PTC) that causes NMD. Thus, the lack of exon 4 usage in *DNMT3A*^{mut} cells can lead to the overexpression of *SRSF3*, a known oncogenic splicing factor.[81] We further applied GoT-Splice to CD34+ cells from an acute myeloid leukemia (AML) patient with a mutation in the splicing factor *U2AF1* (S34F; VAF of 0.16). Despite challenges associated with low expression of *U2AF1* in this sample, genotyping data were available for 1,662 AML cells (12.8% of all cells) through GoT-Splice (Figure S5D). Alternative splicing events between *U2AF1*^{mut} *and U2AF1*^{wt} cells were enriched for exon skipping events[82] (Figure S5E). Through our differential splicing analysis pipeline, we further identified 103 significant differentially spliced events between *U2AF1*^{mut} *and U2AF1*^{wt} cells (Figure S5F), two of which were exon skipping events in *KIN* and *DAP3* that were more prevalent in *U2AF1*^{mut} cells (Figure S5G), as previously reported.[82] These findings demonstrate the generalizability of our method beyond *SF3B1* mutation detection and reveal how *DNMT3A* and *U2AF1* mutations result in different splicing changes compared with *SF3B1* mutations (exon skipping vs. alternative 3′ splice sites, respectively).

Altogether, GoT-Splice enables to link somatic mutations to transcriptional and cell surface protein marker phenotypes, and single-cell splicing changes.

GoT-Splice shows progenitor-specific patterns in *SF3B1*^{mut}- mis-splicing

An important advantage of GoT-Splice is the ability to detect splicing changes at single-cell resolution, allowing the comparison of alternative splicing aberrations between MUT and WT cells in specific cell subsets (Figure 3C; Table S4). We identified both shared and unique *SF3B1*^{mut} cryptic 3′ splice site events across progenitor subtypes in MDS, with the highest usage of cryptic 3′ splice sites occurring in the megakaryocyte-erythroid lineage. *SF3B1*^{mut} MEPs and EPs accounted for most of the cell-type-specific cryptic 3′ splice site events, highlighting the specific impact of *SF3B1* mutations on the erythroid lineage. These progenitor-specific patterns in *SF3B1*^{mut} mis-splicing were confirmed in the validation cohort (MDS04–06; Figures S5H–S5J). In both MDS cohorts, progenitor-specific cryptic 3′ splice sites involved genes related to cell cycle (e.g., *CENPT*),[83] RNA processing (e.g., *CHTOP*, *SF3B1*,[84] *SRSF11*, and *PRPF38A*), oxygen homeostasis (e.g., *HIF1A*), erythroid differentiation (e.g., *CD36*, *FOXRED1*, and *GATA1*[34,85,86]), and heme metabolism (e.g., *UROD*, *PPOX*, and *CIAO1*) (Figures 3C, S5H, and S5I; Table S4). Notably, long-read sequencing was advantageous in detecting cryptic splicing events in functionally important genes (*PPOX* and *UROD*) with poor short-read coverage due to substantial drop-off in the 10× short-reads across the transcript (Figure S5J). Although some genes and pathways identified in the analysis of cryptic 3′ splice sites across cohorts have previously been reported to be disrupted by alternative splicing in bulk studies of *SF3B1*^{mut} MDS samples,[32] their cell-type specificity was unknown. For instance, while the alternative splicing event in *SF3B1* itself has been suggested before as being neoplasm-specific, here, we resolved its erythroid-specific pattern. This isoform—*SF3B1ins*—is predicted to affect splicing by impairing U2

(C) dPSI values between *SF3B1*^{mut} and *SF3B1*^{wt} cells for cryptic 3′ splicing events identified in main progenitor subsets across MDS samples. Columns- cryptic 3′ junctions differentially spliced in at least one cell type, with p value ≤ 0.05 and dPSI ≥ 2. Rows: cell type. Genes highlighted for cell cycle (purple), heme metabolism (green), oxygen homeostasis (black), RNA processing (red), and erythroid differentiation (yellow) pathways. Left bar plots show the fraction of differentially spliced cryptic 3′ splice sites per cell. Top bar plots quantify the total number of cell types where an event is differentially spliced, with the cell-type-specific events on the right.

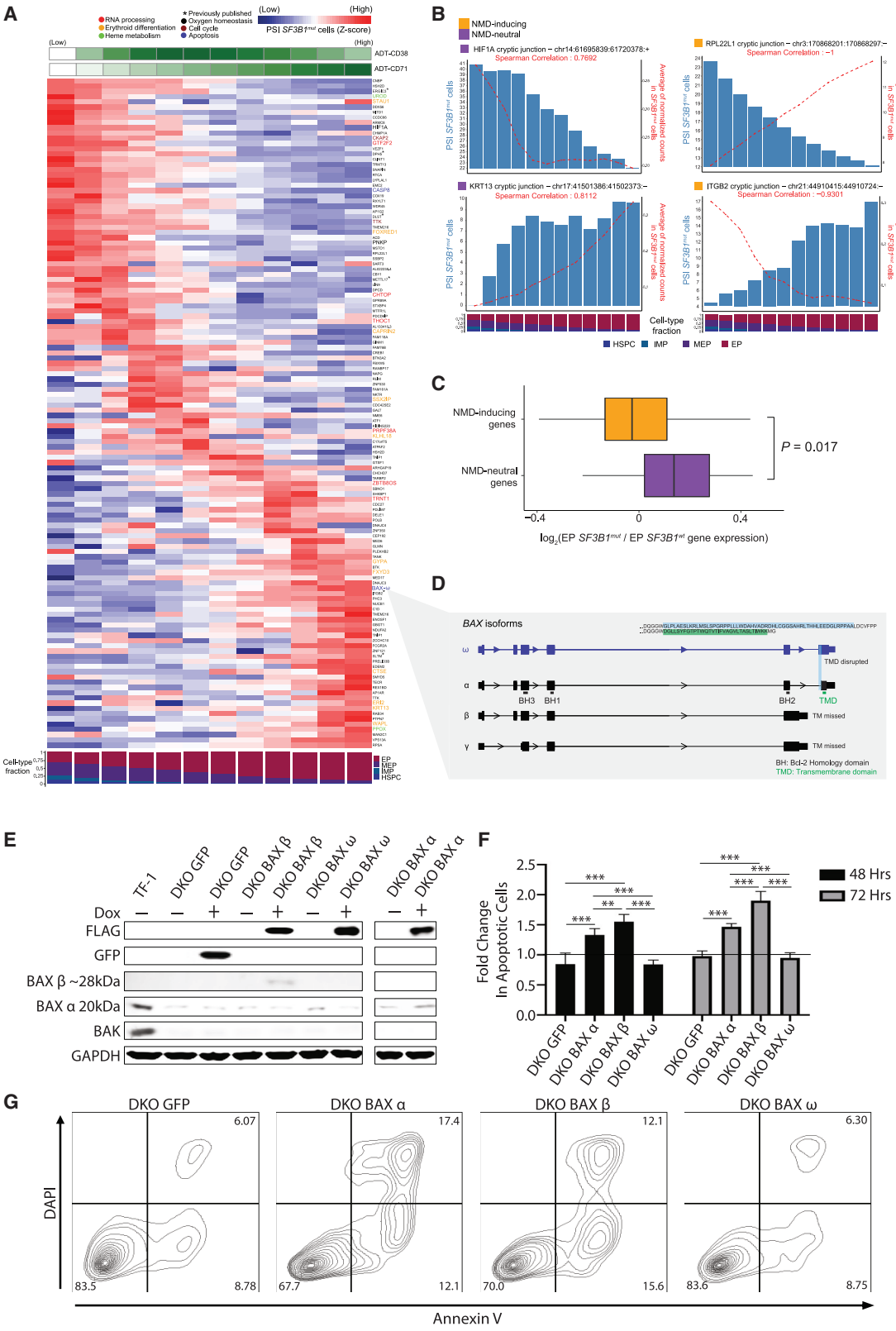


Figure 4. *SF3B1*^{mut}-associated mis-splicing changes along the continuum of erythropoiesis

(A) Percent spliced in (PSI) of junctions in *SF3B1*^{mut} cells along the hematopoietic differentiation trajectory (HSPCs, IMPs, MEPs, and EPs). Rows, (*Z* score normalized), cryptic 3′ splice sites; columns, PSI for usage of a given cryptic 3′ splice site in each window (size of 3,000 *SF3B1*^{mut} cells, sliding by 300 *SF3B1*^{mut}

(legend continued on next page)

small nuclear ribonucleoprotein particle (snRNP) assembly,[84] likely contributing to enhanced mis-splicing dysregulation in the megakaryocyte-erythroid lineage. In addition, cell cycle plays a critical role in terminal differentiation of hematopoietic stem cells[87] and RNA processing, erythroid differentiation, and heme metabolism pathways are directly linked to the regulation of erythropoiesis.[88–90] Furthermore, we observed significant overlap of megakaryocyte-erythroid lineage-specific aberrantly spliced genes between discovery and validation MDS cohorts (p value = 0.00029, Fisher's exact test, with 46.8% of the cryptically spliced genes in MDS also aberrantly spliced in the MDS validation cohort; Figure S5I), indicating cell-type and differentiation-stage dependency of *SF3B1*^{mut}-induced alternative splicing.[27,91–93]

Notably, erythropoiesis occupies a continuum of cell states and is dependent on transcriptional changes that occur along the differentiation trajectory.[47] Analyzing *SF3B1*^{mut} mis-splicing along this continuum (Figure 4A) revealed that some erythroid differentiation, oxygen homeostasis and heme metabolism genes can be mis-spliced more frequently at the earliest stages of EP maturation (e.g., *UROD*, *HIF1A*, and *FOXRED1*[94]), while others display increased mis-splicing in more differentiated EPs (e.g., *GYPA* and *PPOX*). *UROD* is part of the heme biosynthesis pathway; heme is an important structural component of erythroid cells and plays a regulatory role in the differentiation of erythroid precursors.[95] *PPOX* encodes an enzyme involved in mitochondrial heme biosynthesis and, as such, its degradation leads to ineffective erythropoiesis and mitochondrial iron accumulation typical of MDS with ring sideroblast clinical phenotype.[96] These results provide evidence that pathogenic *SF3B1*^{mut}-driven mis-splicing impacts key mediators of hemoglobin synthesis and erythroid differentiation at all stages of erythroid maturation.[97,98]

In some cases, the degree of mis-splicing of a particular transcript (measured by PSI) positively correlated with its expression across the erythroid differentiation trajectory. In others, mis-splicing was negatively correlated with gene expression, often in cryptic 3′ splice site events predicted to lead to transcript degradation by NMD (Figure 4B for examples). Cryptic 3′ splice sites result in the inclusion of short intronic fragments in mRNA and often introduce a PTC.[90,99,100] mRNAs harboring an NMD-inducing PTC located $\geq$ 50 bp upstream of the last exon-exon junction are predicted to undergo NMD, which prevents production of potentially aberrant proteins. In contrast, mRNAs harboring an NMD-neutral PTC, generally located $\leq$ 50 bp upstream of the last exon-exon junction or in the last exon, fail to trigger NMD and produce dysfunctional proteins.[101,102] We classified cryptic 3′ splice sites detected in the MDS samples into two major groups: (1) NMD-inducing events (due to the introduction of a PTC) and (2) NMD-neutral events (Table S4). In accordance with previous reports,[74] of the 421 cryptic 3′ splice sites significantly associated with *SF3B1*^{mut} cells, 228 (54%) were classified as NMD-inducing events while the remaining 193 (46%) harbored NMD-neutral events. Despite the somewhat equal identification of NMD-inducing and NMD-neutral events, we observed an overall decrease in the expression of genes with NMD-inducing event in the MUT cells that harbored these mis-spliced transcripts when compared with NMD-neutral events (p value = 0.017; Figure 4C).

NMD-inducing events affected key genes in erythroid development, such as *UROD*, *GYPA*, *FOXRED1*, and *PPOX*. Transcript loss via NMD[103,104] may thus contribute to disrupted terminal differentiation of EPs. Notable among NMD-neutral affected genes, we identified *BAX*, a member of the Bcl-2 gene family and transcriptional target of *TP53*. *BAX* plays a vital role in the apoptotic cascade, balancing survival, differentiation, and proliferation of EPs during later stages of erythropoiesis[105] (Figure 4A). The identified *BAX* cryptic 3′ splice site, though NMD-neutral, causes a frameshift in the last exon, disrupting the C terminus of the protein. This *BAX* isoform, known as *BAX*-ω (Figure 4D), has been shown to protect cells from apoptotic cell death.[106,107]

To functionally evaluate the significance of *SF3B1*^{mut}-specific alternatively spliced *BAX* isoforms, we generated *BAX* and *BAK1* double knockout (DKO) human TF-1 erythroid leukemia cells and re-expressed FLAG-tagged *BAX*-α, *BAX*-β, and *BAX*-ω isoforms under a doxycycline-inducible promoter (Figure 4E). *BAX* and *BAK1* have functionally redundant pro-apoptotic roles[108] and DKO cell lines have previously been utilized in similar functional validation experiments of *BAX* variants.[109,110] TF-1 cells are an established immature cell line of erythroid origin and are dependent on growth factors for proliferation and survival.[111] We tested the ability of different *BAX* isoforms to induce apoptosis under cytokine-depleted conditions. As expected, TF-1 DKO inducible GFP control cells had no increase in apoptosis while induction of *BAX*-α and *BAX*-β expression led to a significant

cells). Only junctions differentially spliced in at least one cell-type with a dPSI > 2 were analyzed. ADT expression of CD71 and cell-type fractions are shown. Rows ordered according to PSI peak. Genes highlighted for cell cycle (purple), heme metabolism (green), oxygen homeostasis (black), RNA processing (red), erythroid differentiation (yellow), and apoptosis (blue) pathways.

(B) Examples of mis-spliced genes at different erythroid maturation stages. Bars represent PSI in *SF3B1*^{mut} cells. Red lines represent junction ONT expression in *SF3B1*^{mut} cells.

(C) Fold change ($\log_2$) of gene expression between *SF3B1*^{mut} and *SF3B1*^{wt} EP cells in NMD-inducing vs. NMD-neutral genes. p value from Wilcoxon rank-sum test.

(D) *BAX* gene model and relevant isoforms. Characteristic domains are highlighted in main isoform *BAX*-α. The cryptic 3′ splicing event on the terminal exon defines the *BAX*-ω isoform, characterized by frameshift disruption of the transmembrane domain (TMD).

(E) Western blot of TF-1 BAX/BAK double knockout (DKO) cells with doxycycline-inducible expression of control GFP or FLAG-tagged BAX isoforms α, β, and ω after 24 h.

(F) Fold change in annexin V positive TF-1 DKO cells expressing different BAX isoforms under cytokine-depleted conditions + doxycycline (1 μg/mL) at 48 and 72 h normalized to apoptotic cells—doxycycline treatment (black line). n = 2 independent experiments performed in triplicate. Bars represent mean values. Error bars represent ± SD; **p value < 0.01, ***p value < 0.001. p values from two-tailed Student's t-test.

(G) Representative annexin V/DAPI flow cytometry plots of different BAX isoforms after 72 h under cytokine-depleted conditions + doxycycline. Percent frequencies noted in relevant quadrants.

increase in apoptotic cells at 48 and 72 h, consistent with their established roles as inducers of apoptosis.[112] Importantly, expression of *BAX-ω* showed no change in apoptosis under cytokine-depleted conditions, supporting an anti-apoptotic role for *BAX-ω* expression in immature erythroid cells (Figures 4F and 4G). Collectively, our data highlight *SF3B1*^{mut}-specific mis-splicing in the induction of NMD in erythroid differentiation genes and alternative splicing of apoptosis mediators as important events in the pathogenesis of *SF3B1*^{mut} MDS cells.

Accumulation of *SF3B1*^{mut} cells in the erythroid progenitor population and extensive mis-splicing in clonal hematopoiesis

While *SF3B1* mutations are the most common genetic alterations in MDS patients, they are also associated with a high-risk of malignant transformation in CH.[4–8,113,114] However, the study of *SF3B1* mutations directly in primary human samples has been largely limited to MDS, where confounding co-occurrence of other genetic alterations is common. Additionally, it remains unclear exactly how splicing mutations impact cellular phenotypes in CH. For example, while *SF3B1* mutations have been proposed to be drivers of CH,[115] enhancing likelihood of progression to myeloid neoplasia, these mutations often occur as early genetic events in CH cases, with gradually increasing VAF over time. In contrast, mutations in *U2AF1* and *SRSF2* appear later in life, with rapidly increasing VAF.[116] Thus, CH presents a unique setting to interrogate the molecular consequences of *SF3B1* mutations in non-malignant human hematopoiesis.

We therefore isolated viable CD34+ cells from two CH samples with *SF3B1* mutations (VAFs: 0.15 and 0.22, from CD34+ autologous grafts collected from patients with multiple myeloma in remission) and performed GoT-Splice. A total of 9,007 cells across both samples passed quality filters (Figure S6A) and were integrated and clustered based on transcriptome data alone, agnostic to genotyping information (Figures 5A and S6B). Consistent with clinical data indicating normal hematopoietic production, we identified the expected progenitor subtypes using previously annotated progenitor identity markers (Figures 5A, S6C, and S6D). Genotyping data were available for 3,642 cells (40.4%) through GoT (Figure S6E), and copy-number analysis with scRNA-seq data confirmed the expected absence of chromosomal gains or losses (Figure S6F).

Projection of the genotyping information onto the differentiation map (Figure 5B), showed no novel cell identities formed by the *SF3B1* mutations, consistent with the fact that patients with CH exhibit no overt peripheral blood count or morphological abnormalities. However, a differentiation pseudotime ordering analysis revealed enrichment of *SF3B1*^{mut} cells at later pseudotime points compared with *SF3B1*^{wt} cells (Figures 5C and S6G). As in MDS, MUT cells were enriched in more differentiated EPs compared with the earlier HSPCs (p value < 0.001, linear mixed model, Figures 5D and S6H), showing that *SF3B1*^{mut} CH cells already demonstrate an erythroid lineage bias.

To further identify transcriptional dysregulation in *SF3B1*^{mut} CH cells, we performed differential gene expression analysis between MUT and WT cells. We observed an upregulation of genes involved in mRNA translation in the *SF3B1*^{mut} HSPCs (where spliceosome mutations originate from the most primitive multipotent compartment[115,117]) in CH (FDR-adjusted p value = 0.005; Figures 5E and 5F; Table S5), a pathway also observed to be upregulated in the MDS analysis (Figure 1H). In CH, upregulation of mRNA translation pathway genes was observed across multiple cell subtypes along erythroid differentiation, while absent in NPs (Figure 5G). Thus, although no overt blood count abnormalities are observed with *SF3B1* mutation in CH individuals, both the erythroid differentiation bias and aberrant transcriptional profiles are already apparent at this early pre-disease stage.

The analysis of differentially used alternative 3′ splice sites between *SF3B1*^{mut} and *SF3B1*^{wt} CH cells revealed a marked increase in cryptic 3′ splice site usage in *SF3B1*^{mut} cells, as observed in MDS (Figure 6A). These mutant-specific cryptic 3′ splice sites affected genes including *UROD*, *OXA1L*, *SERBP1*, *MED6*, and *ERGIC3*, which were also detected to be cryptically spliced in the *SF3B1*^{mut} MDS cells. Importantly, the lower VAF associated with pre-malignant CH samples highlights the necessity for GoT-Splice to increase the detection of mis-splicing events occurring at low frequencies, and that may otherwise be missed in bulk sequencing studies (Figures 6B and S6I). To validate these results, we analyzed CD34+ cells from a CH sample collected from an individual without myeloma history (CH03, STAR Methods) and observed high correlation of shared splice junctions with our previous samples (Spearman's rho = 0.96, p value = 2.03×10^{-12}; Figure S6J). We also recapitulated 3′ cryptic splice site and exon inclusion events in key genes (e.g., *SERBP1* and *HNRNPA1*; Figure S6K).

To compare mis-spliced transcripts between CH and MDS, we compared cryptic 3′ splice sites with a p value < 0.05 and dPSI of $\geq$ 2 in at least one cell type along the erythroid differentiation trajectory (HSPC, IMP, MEP, or EP) in both CH and MDS cohorts (Table S6). While the overall number of significant cryptic 3′ splice sites in CH was lower than in MDS, we observed a significant overlap in shared cryptic events (p value < 10^{-16}, Fisher's exact test; Figure 6C). Similarly, to MDS, we identified stage-specific mis-splicing events in erythroid maturation, the majority of which overlapped with MDS cryptic 3′ splice sites (Figure 6D). Notably, CH and MDS showed similar mis-splicing dynamics (i.e., increased PSI of *BAX-ω* in *SF3B1*^{mut} cells) in the *BAX* transcript along the erythroid differentiation trajectory (Figure 6E; Table S6). Collectively, these data demonstrate that the aberrant splicing phenotype is already apparent in CH, impacting genes that are also observed in MDS *SF3B1*^{mut} induced mis-splicing.

DISCUSSION

Here, we present GoT-Splice, a single-cell multi-omics integration that enables joint profiling of genotype, gene expression, protein, and aberrant splicing all within the same cell. GoT[15] allows for the comparison between somatically MUT and WT cells within the same sample for genotype-to-phenotype inferences. By further optimization of long-read sequencing of scRNA-seq libraries,[67] we could simultaneously capture both short and long-read data within the same cell, making it possible to analyze the impact of somatic mutations on transcriptional and splicing phenotypes. This stands in contrast to other methods that provide single-cell genotyping capture but either lack mRNA capture[118] or have lower throughput without full-length isoform data.[22,119,120]

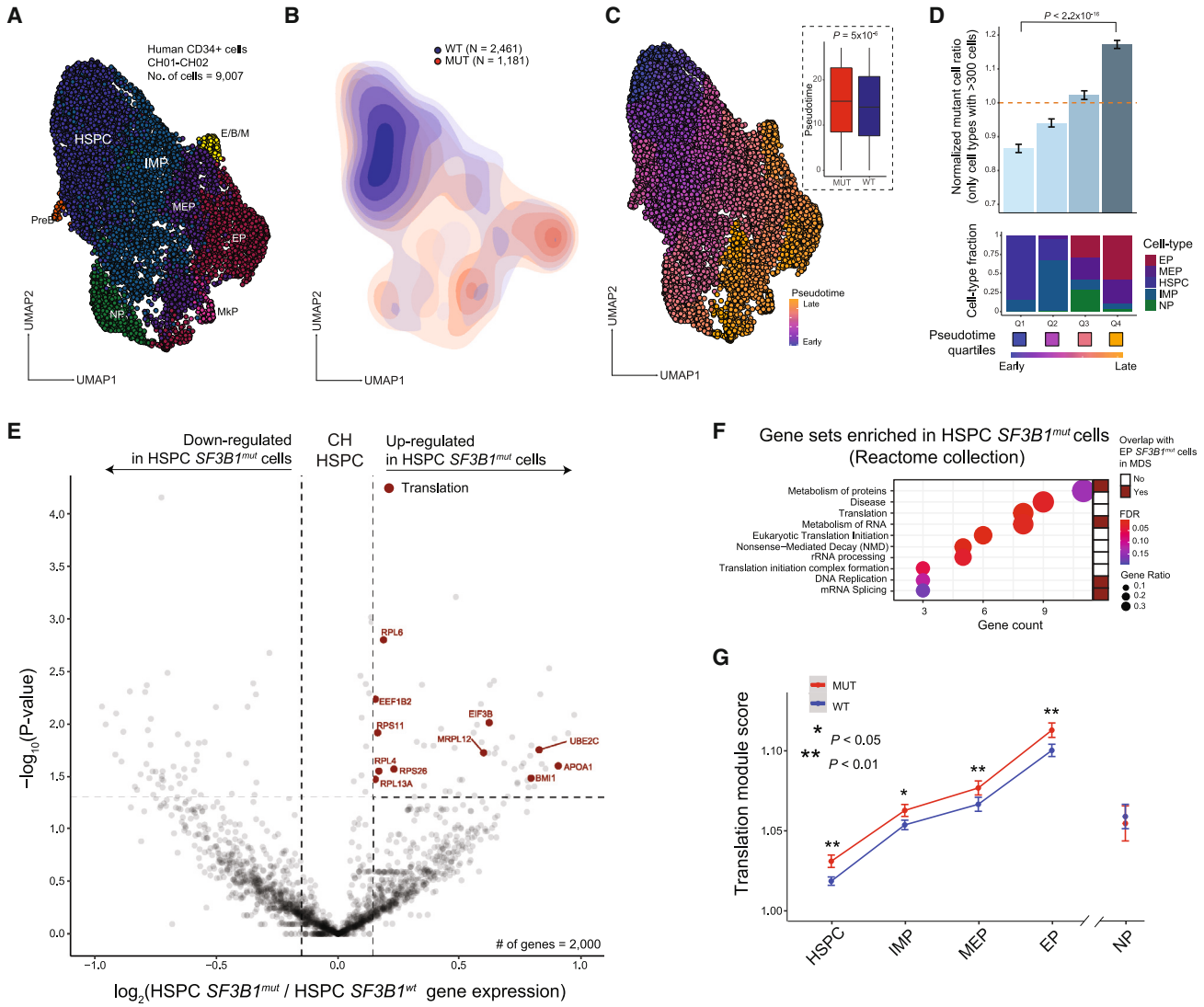


Figure 5. *SF3B1* mutations are enriched along the erythroid lineage in clonal hematopoiesis

(A) UMAP of CD34+ (n = 9,007) cells from clonal hematopoiesis (CH) samples with *SF3B1* K700E or *SF3B1* K666N mutation (n = 2 individuals), overlaid with cluster cell-type assignments. See Figure 1A for cell-type descriptions.

(B) Density plot of $SF3B1^{mut}$ vs. $SF3B1^{wt}$ cells.

(C) UMAP of CD34+ cells from CH samples overlaid with pseudotemporal ordering. Inset: pseudotime in $SF3B1^{mut}$ vs. $SF3B1^{wt}$ cells in the aggregate of CH01–02. p value for comparison of means from Wilcoxon rank-sum test.

(D) Normalized ratio of mutated cells along pseudotime quartiles. Bars show aggregate analysis of samples CH01–02 with mean ±SE of 100 downsampling iterations to 1 genotyping UMI per cell. Only cell types with >300 cells were analyzed. p value from likelihood ratio test of linear mixed model with or without mutation status. Bottom: fraction of cell types within each pseudotime quartile.

(E) Differential gene expression between $SF3B1^{mut}$ and $SF3B1^{wt}$ HSPC cells in CH samples. Genes with an absolute $\log_2$(fold change) > 0.1 and p value < 0.05 were defined as differentially expressed (DE). DE genes in the translation pathway (red, Reactome) are highlighted (see Table S5).

(F) Gene set enrichment analysis of DE genes in $SF3B1^{mut}$ HSPC cells across CH samples. Gene sets that overlap with $SF3B1^{mut}$ EP cells in MDS highlighted (red).

(G) Expression (mean ± SEM) of translation-related genes (Reactome) between $SF3B1^{mut}$ and $SF3B1^{wt}$ cells in progenitor cells from CH01–02 samples. p values from likelihood ratio test of linear mixed model with or without mutation status.

To date, few tools are available to process and analyze single-cell long-read data, especially for the purpose of alternative splicing. To address existing analytic gaps, we developed a long-read splicing analysis pipeline that detects and quantifies alternative splicing events within single cells and highlights differential junction usage across cell subpopulations. For processing the long-read data, we opted for an intron-centric approach followed by split 5′ and 3′ PSI measurements. Calculating the rate of splicing at the 5′ and 3′ ends of the intron improves the detection of the true splicing rate of each individual intron, compared with exon-centric approaches.[71] In addition, our pipeline detected differential splicing patterns between MUT and WT cells, both across entire samples and within individual cell types, with sample-aware permutation testing to integrate across samples.

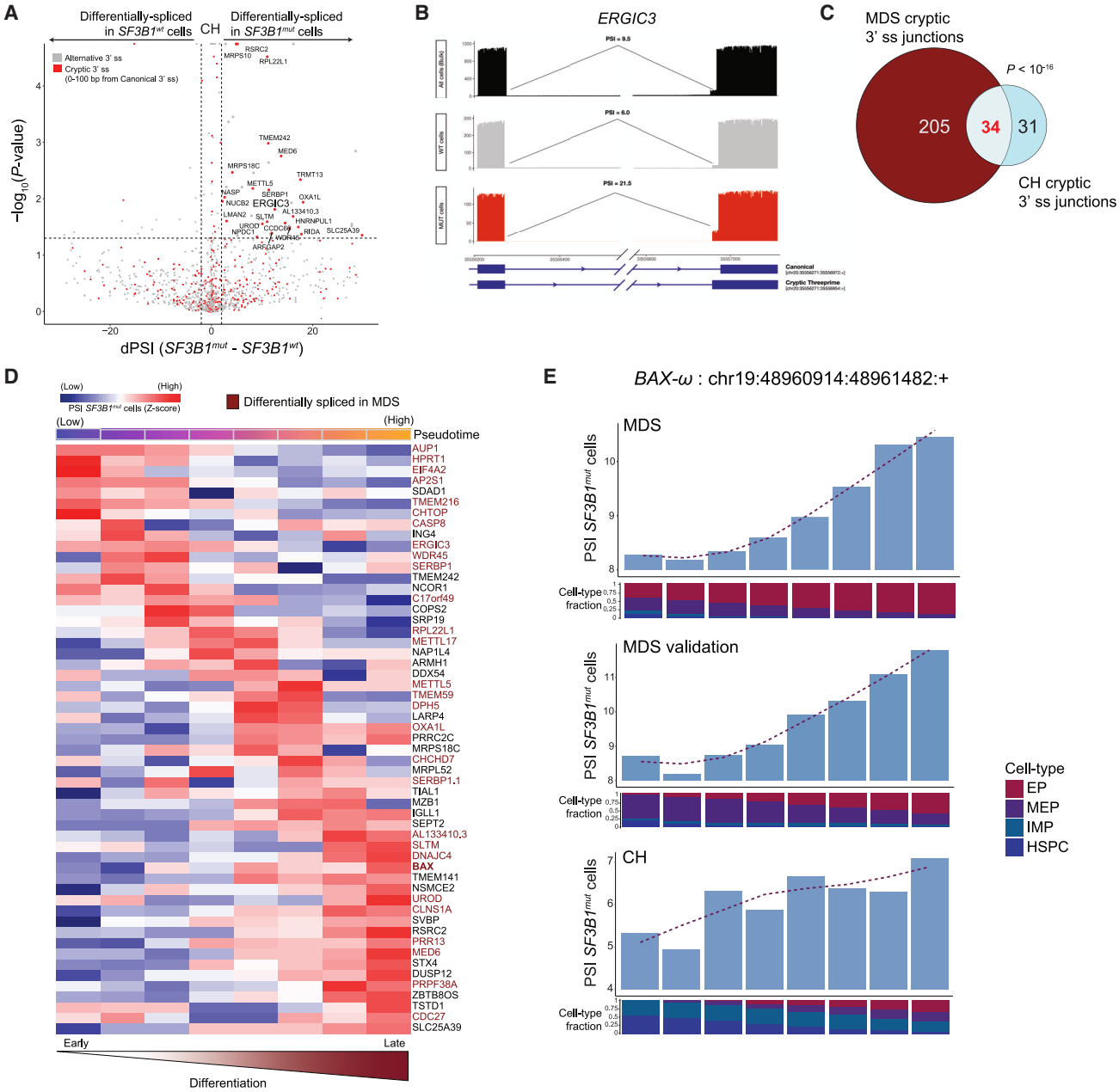


Figure 6. *SF3B1*^{mut} clonal hematopoiesis progenitor cells display cell-type-specific cryptic 3′ splice site usage

(A) Differential splicing analysis between *SF3B1*^{mut} and *SF3B1*^{wt} cells across CH samples. Junctions with an absolute delta percent spliced in (dPSI) > 2 and FDR-adjusted p value < 0.2 were defined as differentially spliced.

(B) Sashimi plot of *ERGIC* intron junction with an *SF3B1*^{mut} associated cryptic 3′ splice site showing RNA-seq coverage in *SF3B1*^{mut} vs. *SF3B1*^{wt} cells within CH samples, as well as compared with the CH samples when treated as bulk (pseudobulk of all cells regardless of genotype). PSI values showing the expected increase in usage of this cryptic 3′ splice site in *SF3B1*^{mut} cells alone when compared with both *SF3B1*^{wt} cells as well as all cells (pseudobulk of sample).

(C) Venn diagram of overlapping genes with cryptic junctions significantly differentially spliced in at least one erythroid lineage cell type (HSPCs, IMPs, MEPs, EPs) with a dPSI > 2 between MDS01–03 and CH samples. p value for the overlap from Fisher's exact test.

(D) Percent spliced in (PSI) of junctions in *SF3B1*^{mut} cells along the hematopoietic differentiation trajectory of erythroid lineage cells. Rows (*Z* score normalized), cryptic 3′ splice sites; columns, PSI for usage of a given cryptic 3′ splice site in each window (size of 600 *SF3B1*^{mut} cells, sliding by 60 *SF3B1*^{mut} cells). Only junctions differentially spliced in at least one cell type with a dPSI > 2 were analyzed. Pseudotime across each window shown. Rows are ordered according to the peak in PSI. Cryptic events also differentially spiced in MDS highlighted (red).

(E) Bar plots of PSI values for usage of the *BAX-ω* isoform across each window of *SF3B1*^{mut} cells in the MDS, MDS validation and CH cohorts along the hematopoietic differentiation trajectory of erythroid lineage cells. Fraction of cell types in each window shown per cohort (MDS: *SF3B1*^{mut} cells [n = 6,376] ordered by CD71 expression, MDS validation: *SF3B1*^{mut} cells [n = 987] ordered by pseudotime, CH: MUT cells [n = 1,021] ordered by pseudotime).

Finally, we provided information regarding the translational consequences of the alternatively spliced junctions. Altogether, our pipeline offers a comprehensive toolkit to process and analyze differential splicing events in scRNA-seq long-read data.

By applying GoT-Splice to the most common splice-altering mutation (*SF3B1*), we interrogated differentiation biases, differential gene expression, protein expression, and splicing patterns, comparing *SF3B1*^{mut} vs. *SF3B1*^{wt} cells co-existing within the same bone marrow. Importantly, while GoT revealed that *SF3B1*^{mut} cells arise early on in uncommitted HSPCs, we observed a differentiation bias of *SF3B1*^{mut} cells toward the EP fate. This finding is of particular interest given the clinical association between *SF3B1* mutations and dysplastic erythropoiesis. Notably, an increase in cell cycle and checkpoint gene expression (*MDM4* and *CCNE1*) as well as the overexpression of erythroid lineage markers, CD36 and CD71, specifically in *SF3B1*^{mut} EPs, support the enrichment of *SF3B1* mutations along the erythroid lineage.

CH samples likewise showed erythroid biased differentiation with higher MUT cell frequency in committed EPs compared with HSPCs. *SF3B1*^{mut} CH cells showed upregulation of genes in pathways involved in translation and mRNA processing, similar to *SF3B1*^{mut} cells in MDS. This finding suggests that the pervasive mis-splicing observed with *SF3B1* mutations may disrupt translation, which may also contribute to dyserythropoiesis.[121,122] Thus, in addition to the shared erythroid differentiation bias in MDS and CH, aberrant transcriptional profiles linked to the dyserythropoiesis phenotype are also already apparent at the pre-disease CH stage.

Leveraging the single-cell resolution of GoT-Splice and differential splicing analysis between *SF3B1*^{mut} and *SF3B1*^{wt} cells revealed cell-type-specific effects of *SF3B1* mutations on mis-splicing. Key genes involved in pathways important for terminal differentiation of hematopoietic stem cells as well as the regulation of erythropoiesis (namely RNA processing, erythroid differentiation, cell cycle and heme metabolism) were cryptically spliced across distinct *SF3B1*^{mut} progenitor cell types, many of which were previously reported to be affected in bulk studies of *SF3B1*^{mut} MDS.[29,56,75] While some cryptic events were neutral, many key genes important for erythroid differentiation were NMD-inducing (e.g., *UROD*, *GYPA*, and *PPOX*) or caused a frameshift event that may affect protein structure and function in both the primary and validation MDS cohorts, such as key apoptosis mediators (e.g., *BAX*). Indeed, our functional data support an anti-apoptotic role for *BAX*-ω in *SF3B1*^{mut} cells. These data are consistent with the recent discovery of C-terminal *BAX* mutations in myeloid clones that arise in CLL patients upon prolonged exposure to venetoclax, demonstrating a role for *BAX* C-terminal alterations in conferring a survival advantage to myeloid cells with this pro-apoptotic treatment.[109] Of note, early clinical observations reported lower response to venetoclax in *SF3B1*^{mut} AML,[123,124] consistent with a potential anti-apoptotic effect of *BAX*-ω. Together, these findings suggest a potential mechanism underlying the observed erythroid-dysplasia phenotype in *SF3B1*^{mut} MDS. Despite the injury to translational machinery (Figures 1H and 1I), *SF3B1*^{mut} EPs may gain some degree of protection against cell death due to the disruption of protein function of pro-apoptotic genes in the *TP53* pathway[125,126] resulting from aberrant splicing, as exemplified by the *BAX*-ω isoform. Overall, mis-splicing of genes involved in erythroid differentiation and apoptosis regulation may therefore lead to the accumulation of *SF3B1*^{mut} EP cells that fail to reach terminal differentiation,[105] leading to the dyserythropoiesis clinical phenotype.

Collectively, this work advances our ability to connect somatic genotypes with complex phenotypes in human samples. Splicing changes have a critical role in cancer biology,[127,128] also evidenced by the prevalence of splice factor mutations across blood and solid tumor malignancies.[129,130] The ability to layer genotyping together with rich splicing annotation can thus enable the investigation of aberrant splicing in cancer evolution. Notably, somatic evolution not only affects cancer, but has been recently shown to ubiquitously impact non-malignant human tissues in the form of somatic mosaicism.[131,132] However, in somatic mosaicism, the challenge of connecting genotypes to cellular phenotypes is magnified given the admixture of MUT and WT cells, and thus studies to date in humans have been largely limited to genotyping. Overcoming this challenge requires advances in single-cell multi-omics for genotype-phenotype mapping in human somatic mosaicism.[21,133] This context highlights the importance of this work as one of the first phenotypic studies of clonal mosaicism in human samples, leading to the observation that the somatic mutation-related phenotype aligns with the more advanced MDS cellular phenotype. We speculate that this observation and other recent data[133] suggest that clonal mosaicisms and neoplastic disorders may, at times, lie across a continuum, whereby clinical phenotypes appear as a result of increasing frequency of MUT cells rather than a qualitative phenotypic change.

Limitations of the study

Although GoT-Splice offers a powerful approach to simultaneously assess multiple single-cell modalities, including genotype and splicing isoforms, we note some limitations. The percentage of cells successfully genotyped can vary due to factors like gene expression levels or the efficiency of the 10x Genomics capture. For instance, in the *SF3B1* analysis conducted in this study, the percentage of genotyped cells ranged from approximately 20% to nearly 90% across individuals. Additionally, due to potential incomplete capture of the heterozygously mutated allele, MUT cells may be mis-classified as WT when the WT allele is captured but the mutant allele is missed. We note that this confounding factor is expected to diminish true signals in MUT vs. WT comparisons, rather than leading to erroneous signals. To address this, we have implemented mitigating strategies such as genotyping WT cells with two or more genotyping amplicon UMIs and downsampling to determine MCF (see STAR Methods for more details). Finally, larger cohort sizes and further functional characterizations are needed to validate our biological findings and examine the impact of isoform *BAX*-ω on *SF3B1*^{mut} cells in MDS and CH.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stem.2023.07.012.

ACKNOWLEDGMENTS

The work was enabled by the Weill Cornell Flow Cytometry Core. R.F.S. is supported by an American Society of Hematology Research Training Award for Fellows. N.D. is supported by a F30 Predoctoral Fellowship from the NHLBI of the National Institutes of Health (F30HL156496) and by a Medical Scientist Training Program grant from the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM007739 to the Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program. R.C. is supported by Lymphoma Research Foundation and Marie Skłodowska-Curie fellowships. D.A.K. is supported by NSF CAREER award DBI2146398. D.H.W. is supported by the Oglesby Charitable Trust. O.A.-W. is supported by a Break Through Cancer award. F.G. is supported by an American Society of Hematology Scholar Award (200264-02), Princess Margaret Cancer Foundation, Ontario Institute for Cancer Research Investigator Award, and J. P. Bickell Foundation. D.A.L. is supported by the Burroughs Wellcome Fund Career Award for Medical Scientists, Valle Scholar Award, Leukemia Lymphoma Scholar Award, and the Mark Foundation Emerging Leader Award. This work was supported by the National Heart Lung and Blood Institute (R01HL157387-01A1 and R01HL128239), the National Cancer Institute (R01s CA242020, R01 CA251138, P50 254838, and R33 CA267219), the Edward P. Evans MDS Foundation, a Tri-Institutional Stem Cell Initiative award, the National Institutes of Health Common Fund Somatic Mosaicism Across Human Tissues (1UG3NS132139-01), and the National Human Genome Research Institute, Center of Excellence in Genomic Science (RM1HG011014). This work received research support from Oxford Nanopore Technologies.

AUTHOR CONTRIBUTIONS

M.C.-L., P.C., A.G.H., R.F.S., J.T., O.A.-W., F.G., and D.A.L. conceived the project and designed the study. K.B., N.S.F., R.S.V., D.H.W., A.T., L.S., R.C.L., I.M.G., and O.A.-W. provided clinical samples. A.G.H., R.F.S., A.D.S., S.G., X.D., S.H., and R.C. performed the experiments. M.C.-L., P.C., A.G.H., L.K., C.C., J.B., A.W.D., N.D., G.M., A.M.S., J.S, and F.G. performed the analyses. M.C.-L., P.C., A.G.H., R.F.S., T.H.M., R.C., S.J., E.H., D.A.K., C.J.P., I.M.G., J.T., O.A.-W., F.G., and D.A.L. helped interpret results. M.C.-L., P.C., A.G.H., R.F.S., O.A.-W., F.G., and D.A.L. wrote the manuscript. All authors reviewed and approved the final manuscript.

DECLARATION OF INTERESTS

F.G. serves as a consultant for S2 Genomics Inc. X.D., J.B., A.W.D., S.H., S.J., and E.H. are employees of Oxford Nanopore Technologies Inc. and are shareholders and/or share option holders. I.M.G. serves on the advisory or consulting board of Bristol Myers Squibb, Takeda, Janssen, Sanofi, Novartis, Amgen, Celgene, Cellectar, Pfizer, Menarini Silicon Biosystems, Oncopeptides, The Binding Site, GlazoSmithKlein, AbbVie, Adaptive, and 10x Genomics. O.A.-W. has served as a consultant for H3B Biomedicine, Foundation Medicine Inc., Merck, Pfizer, and Janssen, and O.A.-W. is on the Scientific Advisory Board of Envisagenics Inc. and AIChemy. O.A.-W. has received prior research funding from H3B Biomedicine and LOXO Oncology unrelated to the current manuscript. D.A.L. has served as a consultant for AbbVie, AstraZeneca, and Illumina and is on the Scientific Advisory Board of Mission Bio, Pangea, Alethiomics, and C2i Genomics; D.A.L. has received prior research funding from BMS, 10x Genomics, Ultima Genomics, and Illumina unrelated to the current manuscript.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community. One or more of the authors of this paper self-identifies as living with a disability. One or more of the authors of this paper received support from a program designed to increase minority representation in their field of research.

Received: October 18, 2022
Revised: May 28, 2023
Accepted: July 18, 2023
Published: August 14, 2023

REFERENCES

1. Abelson, S., Collord, G., Ng, S.W.K., Weissbrod, O., Mendelson Cohen, N., Niemeyer, E., Barda, N., Zuzarte, P.C., Heisler, L., Sundaravadanam, Y., et al. (2018). Prediction of acute myeloid leukaemia risk in healthy individuals. Nature *559*, 400–404. https://doi.org/10.1038/s41586-018-0317-6.
2. Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M.S., Cheang, G., Lee, S., Samuel, M., Ritchie, E.K., Guzman, M.L., Ballman, K.V., et al. (2018). Somatic mutations precede acute myeloid leukemia years before diagnosis. Nat. Med. *24*, 1015–1023. https://doi.org/10.1038/s41591-018-0081-z.
3. Teixeira, V.H., Pipinikas, C.P., Pennycuick, A., Lee-Six, H., Chandrasekharan, D., Beane, J., Morris, T.J., Karpathakis, A., Feber, A., Breeze, C.E., et al. (2019). Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. Nat. Med. *25*, 517–525. https://doi.org/10.1038/s41591-018-0323-0.

4. Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L., Druley, T.E., Fisher, D.S., and Blundell, J.R. (2020). The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science *367*, 1449–1454. https://doi.org/10.1126/science.aay9333.

5. Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478. https://doi.org/10.1038/nm.3733.

6. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487. https://doi.org/10.1056/NEJMoa1409405.

7. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498. https://doi.org/10.1056/NEJMoa1408617.

8. Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., Kennedy, J.A., Schimmer, A.D., Schuh, A.C., Yee, K.W., et al. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328–333. https://doi.org/10.1038/nature13038.

9. Mustjoki, S., and Young, N.S. (2021). Somatic mutations in "benign" disease. N. Engl. J. Med. *384*, 2039–2052. https://doi.org/10.1056/NEJMra2101920.

10. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P.V., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880–886. https://doi.org/10.1126/science.aaa6806.

11. Yokoyama, A., Kakiuchi, N., Yoshizato, T., Nannya, Y., Suzuki, H., Takeuchi, Y., Shiozawa, Y., Sato, Y., Aoki, K., Kim, S.K., et al. (2019). Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature *565*, 312–317. https://doi.org/10.1038/s41586-018-0811-x.

12. Yizhak, K., Aguet, F., Kim, J., Hess, J.M., Kübler, K., Grimsby, J., Frazer, R., Zhang, H., Haradhvala, N.J., Rosebrock, D., et al. (2019). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science *364*. https://doi.org/10.1126/science.aaw0726.

13. Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R., et al. (2018). Somatic mutant clones colonize the human esophagus with age. Science *362*, 911–917. https://doi.org/10.1126/science.aau3879.

14. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196. https://doi.org/10.1126/science.aad0501.

15. Nam, A.S., Kim, K.T., Chaligne, R., Izzo, F., Ang, C., Taylor, J., Myers, R.M., Abu-Zeinah, G., Brand, R., Omans, N.D., et al. (2019). Somatic mutations and cell identity linked by genotyping of transcriptomes. Nature *571*, 355–360. https://doi.org/10.1038/s41586-019-1367-0.

16. Chaligne, R., Gaiti, F., Silverbush, D., Schiffman, J.S., Weisman, H.R., Kluegel, L., Gritsch, S., Deochand, S.D., Gonzalez Castro, L.N., Richman, A.R., et al. (2021). Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. Nat. Genet. *53*, 1469–1479. https://doi.org/10.1038/s41588-021-00927-7.

17. Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell *171*, 1611–1624.e24. https://doi.org/10.1016/j.cell.2017.10.044.

18. Baryawno, N., Przybylski, D., Kowalczyk, M.S., Kfoury, Y., Severe, N., Gustafsson, K., Kokkaliaris, K.D., Mercier, F., Tabaka, M., Hofree, M., et al. (2019). A cellular taxonomy of the bone marrow stroma in homeostasis and leukemia. Cell *177*, 1915–1932.e16. https://doi.org/10.1016/j.cell.2019.04.040.

19. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell *178*, 835–849.e21. https://doi.org/10.1016/j.cell.2019.06.024.

20. Oren, Y., Tsabar, M., Cuoco, M.S., Amir-Zilberstein, L., Cabanos, H.F., Hütter, J.C., Hu, B., Thakore, P.I., Tabaka, M., Fulco, C.P., et al. (2021). Cycling cancer persister cells arise from lineages with distinct programs. Nature *596*, 576–582. https://doi.org/10.1038/s41586-021-03796-6.

21. Nam, A.S., Chaligne, R., and Landau, D.A. (2021). Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat. Rev. Genet. *22*, 3–18. https://doi.org/10.1038/s41576-020-0265-5.

22. Rodriguez-Meira, A., Buck, G., Clark, S.A., Povinelli, B.J., Alcolea, V., Louka, E., McGowan, S., Hamblin, A., Sousos, N., Barkas, N., et al. (2019). Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol. Cell *73*, 1292–1305.e8. https://doi.org/10.1016/j.molcel.2019.01.009.

23. Gaiti, F., Chaligne, R., Gu, H., Brand, R.M., Kothen-Hill, S., Schulman, R.C., Grigorev, K., Risso, D., Kim, K.T., Pastore, A., et al. (2019). Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature *569*, 576–580. https://doi.org/10.1038/s41586-019-1198-z.

24. Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., et al. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia *28*, 241–247. https://doi.org/10.1038/leu.2013.336.

25. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69. https://doi.org/10.1038/nature10496.

26. Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C.J., Ellis, P., Wedge, D.C., Pellagatti, A., et al. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood *122*, 3616–3627. https://doi.org/10.1182/blood-2013-08-518886.

27. Inoue, D., Bradley, R.K., and Abdel-Wahab, O. (2016). Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis. Genes Dev. *30*, 989–1001. https://doi.org/10.1101/gad.278424.116.

28. Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J.S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., et al. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N. Engl. J. Med. *365*, 1384–1395. https://doi.org/10.1056/NEJMoa1103283.

29. Darman, R.B., Seiler, M., Agrawal, A.A., Lim, K.H., Peng, S., Aird, D., Bailey, S.L., Bhavsar, E.B., Chan, B., Colla, S., et al. (2015). Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. Cell Rep. *13*, 1033–1045. https://doi.org/10.1016/j.celrep.2015.09.053.

30. Obeng, E.A., Chappell, R.J., Seiler, M., Chen, M.C., Campagna, D.R., Schmidt, P.J., Schneider, R.K., Lord, A.M., Wang, L., Gambe, R.G., et al. (2016). Physiologic expression of Sf3b1(K700E) causes impaired erythropoiesis, aberrant splicing, and sensitivity to therapeutic spliceosome modulation. Cancer Cell *30*, 404–417. https://doi.org/10.1016/j.ccell.2016.08.006.

31. Dalton, W.B., Helmenstine, E., Walsh, N., Gondek, L.P., Kelkar, D.S., Read, A., Natrajan, R., Christenson, E.S., Roman, B., Das, S., et al. (2019). Hotspot SF3B1 mutations induce metabolic reprogramming and vulnerability to serine deprivation. J. Clin. Invest. *129*, 4708–4723. https://doi.org/10.1172/JCI125022.

32. Pellagatti, A., Armstrong, R.N., Steeples, V., Sharma, E., Repapi, E., Singh, S., Sanchi, A., Radujkovic, A., Horn, P., Dolatshad, H., et al. (2018). Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood *132*, 1225–1240. https://doi.org/10.1182/blood-2018-04-843771.

33. Lee, S.C.-W., North, K., Kim, E., Jang, E., Obeng, E., Lu, S.X., Liu, B., Inoue, D., Yoshimi, A., Ki, M., et al. (2018). Synthetic lethal and convergent biological effects of cancer-associated spliceosomal gene mutations. Cancer Cell *34*, 225–241.e8. https://doi.org/10.1016/j.ccell.2018.07.003.

34. Lieu, Y.K., Liu, Z., Ali, A.M., Wei, X., Penson, A., Zhang, J., An, X., Rabadan, R., Raza, A., Manley, J.L., et al. (2022). SF3B1 mutant-induced missplicing of MAP3K7 causes anemia in myelodysplastic syndromes. Proc. Natl. Acad. Sci. USA *119*. https://doi.org/10.1073/pnas.2111703119.

35. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868. https://doi.org/10.1038/nmeth.4380.

36. Mimitou, E.P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat. Methods *16*, 409–412. https://doi.org/10.1038/s41592-019-0392-0.

37. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281. https://doi.org/10.1038/ncb3493.

38. Ong, S.M., Teng, K., Newell, E., Chen, H., Chen, J., Loy, T., Yeo, T.W., Fink, K., and Wong, S.C. (2019). A novel, five-marker alternative to CD16–CD14 gating to identify the three human monocyte subsets. Front. Immunol. *10*, 1761. https://doi.org/10.3389/fimmu.2019.01761.

39. Buus, T.B., Herrera, A., Ivanova, E., Mimitou, E., Cheng, A., Herati, R.S., Papagiannakopoulos, T., Smibert, P., Odum, N., and Koralov, S.B. (2021). Improving oligo-conjugated antibody signal in multimodal single-cell analysis. eLife *10*, e61973. https://doi.org/10.7554/eLife.61973.

40. De La Garza, A., Cameron, R.C., Gupta, V., Fraint, E., Nik, S., and Bowman, T.V. (2019). The splicing factor Sf3b1 regulates erythroid maturation and proliferation via TGFβ signaling in zebrafish. Blood Adv. *3*, 2093–2104. https://doi.org/10.1182/bloodadvances.2018027714.

41. Huang, Y., Hale, J., Wang, Y., Li, W., Zhang, S., Zhang, J., Zhao, H., Guo, X., Liu, J., Yan, H., et al. (2018). SF3B1 deficiency impairs human erythropoiesis via activation of p53 pathway: implications for understanding of ineffective erythropoiesis in MDS. J. Hematol. Oncol. *11*, 19. https://doi.org/10.1186/s13045-018-0558-8.

42. Shubinsky, G., and Schlesinger, M. (1997). The CD38 lymphocyte differentiation marker: new insight into its ectoenzymatic activity and its role as a signal transducer. Immunity *7*, 315–324. https://doi.org/10.1016/S1074-7613(00)80353-2.

43. Triana, S., Vonficht, D., Jopp-Saile, L., Raffel, S., Lutz, R., Leonce, D., Antes, M., Hernández-Malmierca, P., Ordoñez-Rueda, D., Ramasz, B., et al. (2021). Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. Nat. Immunol. *22*, 1577–1589. https://doi.org/10.1038/s41590-021-01059-0.

44. McKenzie, J.L., Gan, O.I., Doedens, M., and Dick, J.E. (2007). Reversible cell surface expression of CD38 on CD34-positive human hematopoietic repopulating cells. Exp. Hematol. *35*, 1429–1436. https://doi.org/10.1016/j.exphem.2007.05.017.

45. Chung, S.S., Eng, W.S., Hu, W., Khalaj, M., Garrett-Bakelman, F.E., Tavakkoli, M., Levine, R.L., Carroll, M., Klimek, V.M., Melnick, A.M., et al. (2017). CD99 is a therapeutic target on disease stem cells in myeloid malignancies. Sci. Transl. Med. *9*, eaaj2025. https://doi.org/10.1126/scitranslmed.aaj2025.

46. Kingwell, K. (2017). Cancer: CD99 marks malignant myeloid stem cells. Nat. Rev. Drug Discov. *16*, 166. https://doi.org/10.1038/nrd.2017.31.

47. Tusi, B.K., Wolock, S.L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J.R., Klein, A.M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. Nature *555*, 54–60. https://doi.org/10.1038/nature25741.

48. Signer, R.A.J., Magee, J.A., Salic, A., and Morrison, S.J. (2014). Haematopoietic stem cells require a highly regulated protein synthesis rate. Nature *509*, 49–54. https://doi.org/10.1038/nature13035.

49. Khajuria, R.K., Munschauer, M., Ulirsch, J.C., Fiorini, C., Ludwig, L.S., McFarland, S.K., Abdulhay, N.J., Specht, H., Keshishian, H., Mani, D.R., et al. (2018). Ribosome levels selectively regulate translation and lineage commitment in human hematopoiesis. Cell *173*, 90–103.e19. https://doi.org/10.1016/j.cell.2018.02.036.

50. Stevens, B.M., Khan, N., D'Alessandro, A., Nemkov, T., Winters, A., Jones, C.L., Zhang, W., Pollyea, D.A., and Jordan, C.T. (2018). Characterization and targeting of malignant stem cells in patients with advanced myelodysplastic syndromes. Nat. Commun. *9*, 3694. https://doi.org/10.1038/s41467-018-05984-x.

51. van Galen, P., Mbong, N., Kreso, A., Schoof, E.M., Wagenblast, E., Ng, S.W.K., Krivdova, G., Jin, L., Nakauchi, H., and Dick, J.E. (2018). Integrated stress response activity marks stem cells in normal hematopoiesis and leukemia. Cell Rep. *25*, 1109–1117.e5. https://doi.org/10.1016/j.celrep.2018.10.021.

52. Ajore, R., Raiser, D., McConkey, M., Jöud, M., Boidol, B., Mar, B., Saksena, G., Weinstock, D.M., Armstrong, S., Ellis, S.R., et al. (2017). Deletion of ribosomal protein genes is a common vulnerability in human cancer, especially in concert with TP53 mutations. EMBO Mol. Med. *9*, 498–507. https://doi.org/10.15252/emmm.201606660.

53. Dutt, S., Narla, A., Lin, K., Mullally, A., Abayasekara, N., Megerdichian, C., Wilson, F.H., Currie, T., Khanna-Gupta, A., Berliner, N., et al. (2011). Haploinsufficiency for ribosomal protein genes causes selective activation of p53 in human erythroid progenitor cells. Blood *117*, 2567–2576. https://doi.org/10.1182/blood-2010-07-295238.

54. Fumagalli, S., Ivanenkov, V.V., Teng, T., and Thomas, G. (2012). Suprainduction of p53 by disruption of 40S and 60S ribosome biogenesis leads to the activation of a novel G2/M checkpoint. Genes Dev. *26*, 1028–1040. https://doi.org/10.1101/gad.189951.112.

55. Bursac, S., Brdovcak, M.C., Donati, G., and Volarevic, S. (2014). Activation of the tumor suppressor p53 upon impairment of ribosome biogenesis. Biochim. Biophys. Acta *1842*, 817–830. https://doi.org/10.1016/j.bbadis.2013.08.014.

56. Pellagatti, A., Marafioti, T., Paterson, J.C., Barlow, J.L., Drynan, L.F., Giagounidis, A., Pileri, S.A., Cazzola, M., McKenzie, A.N.J., Wainscoat, J.S., et al. (2010). Induction of p53 and up-regulation of the p53 pathway in the human 5q− syndrome. Blood *115*, 2721–2723. https://doi.org/10.1182/blood-2009-12-259705.

57. Ebert, B.L., Pretz, J., Bosco, J., Chang, C.Y., Tamayo, P., Galili, N., Raza, A., Root, D.E., Attar, E., Ellis, S.R., et al. (2008). Identification of RPS14 as a 5Q- syndrome gene by RNA interference screen. Nature *451*, 335–339. https://doi.org/10.1038/nature06494.

58. Mills, E.W., and Green, R. (2017). Ribosomopathies: there's strength in numbers. Science *358*, eaan2755. https://doi.org/10.1126/science.aan2755.

59. Caldon, C.E., and Musgrove, E.A. (2010). Distinct and redundant functions of cyclin E1 and cyclin E2 in development and cancer. Cell Div. *5*, 2. https://doi.org/10.1186/1747-1028-5-2.

60. Toledo, F., and Wahl, G.M. (2007). MDM2 and MDM4: p53 regulators as targets in anticancer therapy. Int. J. Biochem. Cell Biol. *39*, 1476–1482. https://doi.org/10.1016/j.biocel.2007.03.022.

61. Perry, M.E. (2010). The regulation of the p53-mediated stress response by MDM2 and MDM4. Cold Spring Harb. Perspect. Biol. *2*, a000968. https://doi.org/10.1101/cshperspect.a000968.

62. Gilkes, D.M., Chen, L., and Chen, J. (2006). MDMX regulation of p53 response to ribosomal stress. EMBO J. *25*, 5614–5625. https://doi.org/10.1038/sj.emboj.7601424.

63. Le Goff, S., Boussaid, I., Floquet, C., Raimbault, A., Hatin, I., Andrieu-Soler, C., Salma, M., Leduc, M., Gautier, E.F., Guyot, B., et al. (2021). p53 activation during ribosome biogenesis regulates normal erythroid differentiation. Blood *137*, 89–102. https://doi.org/10.1182/blood.2019003439.

64. Dolatshad, H., Pellagatti, A., Fernandez-Mercado, M., Yip, B.H., Malcovati, L., Attwood, M., Przychodzen, B., Sahgal, N., Kanapin, A.A., Lockstone, H., et al. (2015). Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia *29*, 1092–1103. https://doi.org/10.1038/leu.2014.331.

65. Liu, Z., Yoshimi, A., Wang, J., Cho, H., Chun-Wei Lee, S., Ki, M., Bitner, L., Chu, T., Shah, H., Liu, B., et al. (2020). Mutations in the RNA splicing factor SF3B1 promote tumorigenesis through MYC stabilization. Cancer Discov. *10*, 806–821. https://doi.org/10.1158/2159-8290.CD-19-1330.

66. Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., et al. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. Preprint at bioRxiv. https://doi.org/10.1101/672931.

67. Lebrigand, K., Magnone, V., Barbry, P., and Waldmann, R. (2020). High throughput error corrected nanopore single cell transcriptome sequencing. Nat. Commun. *11*, 4025. https://doi.org/10.1038/s41467-020-17800-6.

68. Tian, L., Jabbari, J.S., Thijssen, R., Gouil, Q., Amarasinghe, S.L., Voogd, O., Kariyawasam, H., Du, M.R.M., 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. https://doi.org/10.1186/s13059-021-02525-6.

69. Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., and Brooks, A.N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat. Commun. *11*, 1438. https://doi.org/10.1038/s41467-020-15171-6.

70. Hardwick, S.A., Hu, W., Joglekar, A., Fan, L., Collier, P.G., Foord, C., Balacco, J., Lanjewar, S., Sampson, M.M., Koopmans, F., et al. (2022). Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. Nat. Biotechnol. *40*, 1082–1092. https://doi.org/10.1038/s41587-022-01231-3.

71. Pervouchine, D.D., Knowles, D.G., and Guigó, R. (2013). Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics *29*, 273–274. https://doi.org/10.1093/bioinformatics/bts678.

72. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158. https://doi.org/10.1038/s41588-017-0004-9.

73. Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife *5*, e11752. https://doi.org/10.7554/eLife.11752.

74. Shiozawa, Y., Malcovati, L., Gallì, A., Sato-Otsubo, A., Kataoka, K., Sato, Y., Watatani, Y., Suzuki, H., Yoshizato, T., Yoshida, K., et al. (2018). Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. Nat. Commun. *9*, 3649. https://doi.org/10.1038/s41467-018-06063-x.

75. Mupo, A., Seiler, M., Sathiaseelan, V., Pance, A., Yang, Y., Agrawal, A.A., Iorio, F., Bautista, R., Pacharne, S., Tzelepis, K., et al. (2017). Hemopoietic-specific Sf3b1-K700E knock-in mice display the splicing defect seen in human MDS but develop anemia without ring sideroblasts. Leukemia *31*, 720–727. https://doi.org/10.1038/leu.2016.251.

76. Prjibelski, A.D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A.L., and Tilgner, H.U. (2023). Accurate isoform discovery with IsoQuant using long reads. Nat. Biotechnol. *41*, 915–918. https://doi.org/10.1038/s41587-022-01565-y.

77. Wang, L., Brooks, A.N., Fan, J., Wan, Y., Gambe, R., Li, S., Hergert, S., Yin, S., Freeman, S.S., Levin, J.Z., et al. (2016). Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. Cancer Cell *30*, 750–763. https://doi.org/10.1016/j.ccell.2016.10.005.

78. Ramabadran, R., Wang, J.H., Reyes, J.M., Guzman, A.G., Gupta, S., Rosas, C., Brunetti, L., Gundry, M.C., Tovy, A., Long, H., et al. (2023). DNMT3A-coordinated splicing governs the stem state switch towards differentiation in embryonic and haematopoietic stem cells. Nat. Cell Biol. *25*, 528–539. https://doi.org/10.1038/s41556-023-01109-9.

79. Banaszak, L.G., Giudice, V., Zhao, X., Wu, Z., Gao, S., Hosokawa, K., Keyvanfar, K., Townsley, D.M., Gutierrez-Rodrigues, F., Fernandez Ibanez, M.D.P., et al. (2018). Abnormal RNA splicing and genomic instability after induction of DNMT3A mutations by CRISPR/Cas9 gene editing. Blood Cells Mol. Dis. *69*, 10–22. https://doi.org/10.1016/j.bcmd.2017.12.002.

80. Thomas, J.D., Polaski, J.T., Feng, Q., De Neef, E.J., Hoppe, E.R., McSharry, M.V., Pangallo, J., Gabel, A.M., Belleville, A.E., Watson, J., et al. (2020). RNA isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. Nat. Genet. *52*, 84–94. https://doi.org/10.1038/s41588-019-0555-z.

81. Che, Y., and Fu, L. (2020). Aberrant expression and regulatory network of splicing factor-SRSF3 in tumors. J. Cancer *11*, 3502–3511. https://doi.org/10.7150/jca.42645.

82. Biancon, G., Joshi, P., Zimmer, J.T., Hunck, T., Gao, Y., Lessard, M.D., Courchaine, E., Barentine, A.E.S., Machyna, M., Botti, V., et al. (2022). Precision analysis of mutant U2AF1 activity reveals deployment of stress granules in myeloid malignancies. Mol. Cell *82*, 1107–1122.e7. https://doi.org/10.1016/j.molcel.2022.02.025.

83. Jones, M., Osawa, G., Regal, J.A., Weinberg, D.N., Taggart, J., Kocak, H., Friedman, A., Ferguson, D.O., Keegan, C.E., and Maillard, I. (2014). Hematopoietic stem cells are acutely sensitive to Acd shelterin gene inactivation. J. Clin. Invest. *124*, 353–366. https://doi.org/10.1172/JCI67871.

84. Bergot, T., Lippert, E., Douet-Guilbert, N., Commet, S., Corcos, L., and Bernard, D.G. (2020). Human cancer-associated mutations of SF3B1 lead to a splicing modification of its own RNA. Cancers *12*, 652. https://doi.org/10.3390/cancers12030652.

85. Abdulhay, N.J., Fiorini, C., Verboon, J.M., Ludwig, L.S., Ulirsch, J.C., Zieger, B., Lareau, C.A., Mi, X., Roy, A., Obeng, E.A., et al. (2019). Impaired human hematopoiesis due to a cryptic intronic GATA1 splicing mutation. J. Exp. Med. *216*, 1050–1060. https://doi.org/10.1084/jem.20181625.

86. Ling, T., Crispino, J.D., Zingariello, M., Martelli, F., and Migliaccio, A.R. (2018). GATA1 insufficiencies in primary myelofibrosis and other hematopoietic disorders: consequences for therapy. Expert Rev. Hematol. *11*, 169–184. https://doi.org/10.1080/17474086.2018.1436965.

87. Pietras, E.M., Warr, M.R., and Passegué, E. (2011). Cell cycle regulation in hematopoietic stem cells. J. Cell Biol. *195*, 709–720. https://doi.org/10.1083/jcb.201102131.

88. Chung, J., Chen, C., and Paw, B.H. (2012). Heme metabolism and erythropoiesis. Curr. Opin. Hematol. *19*, 156–162. https://doi.org/10.1097/MOH.0b013e328351c48b.

89. Dzierzak, E., and Philipsen, S. (2013). Erythropoiesis: development and differentiation. Cold Spring Harb. Perspect. Med. *3*, a011601. https://doi.org/10.1101/cshperspect.a011601.

90. Moore, K.S., and von Lindern, M. (2018). RNA binding proteins and regulation of mRNA translation in erythropoiesis. Front. Physiol. *9*, 910. https://doi.org/10.3389/fphys.2018.00910.

91. Wong, J.J.-L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013).

Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583–595. https://doi.org/10.1016/j.cell.2013.06.052.

92. Edwards, C.R., Ritchie, W., Wong, J.J.-L., Schmitz, U., Middleton, R., An, X., Mohandas, N., Rasko, J.E.J., and Blobel, G.A. (2016). A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood *127*, e24–e34. https://doi.org/10.1182/blood-2016-01-692764.
93. Chen, L., Kostadima, M., Martens, J.H.A., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033. https://doi.org/10.1126/science.1251033.
94. Asimomitis, G., Deslauriers, A.G., Kotini, A.G., Bernard, E., Esposito, D., Olszewska, M., Spyrou, N., Arango Ossa, J.E., Mortera-Blanco, T., Koche, R.P., et al. (2022). Patient-specific MDS-RS iPSCs define the mis-spliced transcript repertoire and chromatin landscape of SF3B1-mutant HSPCs. Blood Adv. *6*, 2992–3005. https://doi.org/10.1182/bloodadvances.2021006325.
95. Chiabrando, D., Mercurio, S., and Tolosano, E. (2014). Heme and erythropoieis: more than a structural role. Haematologica *99*, 973–983. https://doi.org/10.3324/haematol.2013.091991.
96. Malcovati, L., Stevenson, K., Papaemmanuil, E., Neuberg, D., Bejar, R., Boultwood, J., Bowen, D.T., Campbell, P.J., Ebert, B.L., Fenaux, P., et al. (2020). SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS. Blood *136*, 157–170. https://doi.org/10.1182/blood.2020004850.
97. Clough, C.A., Pangallo, J., Sarchi, M., Ilagan, J.O., North, K., Bergantinos, R., Stolla, M.C., Naru, J., Nugent, P., Kim, E., et al. (2022). Coordinated missplicing of TMEM14C and ABCB7 causes ring sideroblast formation in SF3B1-mutant myelodysplastic syndrome. Blood *139*, 2038–2049. https://doi.org/10.1182/blood.2021012652.
98. Conte, S., Katayama, S., Vesterlund, L., Karimi, M., Dimitriou, M., Jansson, M., Mortera-Blanco, T., Unneberg, P., Papaemmanuil, E., Sander, B., et al. (2015). Aberrant splicing of genes involved in haemoglobin synthesis and impaired terminal erythroid maturation in SF3B1 mutated refractory anaemia with ring sideroblasts. Br. J. Haematol. *171*, 478–490. https://doi.org/10.1111/bjh.13610.
99. Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nat. Struct. Mol. Biol. *16*, 107–113. https://doi.org/10.1038/nsmb.1550.
100. Hug, N., Longman, D., and Cáceres, J.F. (2016). Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res. *44*, 1483–1495. https://doi.org/10.1093/nar/gkw010.
101. Kurosaki, T., Popp, M.W., and Maquat, L.E. (2019). Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nat. Rev. Mol. Cell Biol. *20*, 406–420. https://doi.org/10.1038/s41580-019-0126-2.
102. Supek, F., Lehner, B., and Lindeboom, R.G.H. (2021). To NMD or not to NMD: nonsense-mediated mRNA decay in cancer and other genetic diseases. Trends Genet. *37*, 657–668. https://doi.org/10.1016/j.tig.2020.11.002.
103. Nickless, A., Bailis, J.M., and You, Z. (2017). Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci. *7*, 26. https://doi.org/10.1186/s13578-017-0153-7.
104. Leeksma, A.C., Derks, I.A.M., Kasem, M.H., Kilic, E., de Klein, A., Jager, M.J., van de Loosdrecht, A.A., Jansen, J.H., Navrkalova, V., Faber, L.M., et al. (2020). The effect of SF3B1 mutation on the DNA damage response and nonsense-mediated mRNA decay in cancer. Front. Oncol. *10*, 609409. https://doi.org/10.3389/fonc.2020.609409.
105. Testa, U. (2004). Apoptotic mechanisms in the control of erythropoiesis. Leukemia *18*, 1176–1199. https://doi.org/10.1038/sj.leu.2403383.
106. Zhou, M., Demo, S.D., McClure, T.N., Crea, R., and Bitler, C.M. (1998). A novel splice variant of the cell death-promoting protein BAX. J. Biol. Chem. *273*, 11930–11936. https://doi.org/10.1074/jbc.273.19.11930.
107. Van de Casteele, M., Kefas, B.A., Ling, Z., Heimberg, H., and Pipeleers, D.G. (2002). Specific expression of Bax-ω in pancreatic β-cells is down-regulated by cytokines before the onset of apoptosis. Endocrinology *143*, 320–326. https://doi.org/10.1210/endo.143.1.8574.
108. Lindsten, T., Ross, A.J., King, A., Zong, W.X., Rathmell, J.C., Shiels, H.A., Ulrich, E., Waymire, K.G., Mahar, P., Frauwirth, K., et al. (2000). The combined functions of proapoptotic Bcl-2 family members bak and bax are essential for normal development of multiple tissues. Mol. Cell *6*, 1389–1399. https://doi.org/10.1016/s1097-2765(00)00136-2.
109. Blombery, P., Lew, T.E., Dengler, M.A., Thompson, E.R., Lin, V.S., Chen, X., Nguyen, T., Panigrahi, A., Handunnetti, S.M., Carney, D.A., et al. (2022). Clonal hematopoiesis, myeloid disorders and BAX-mutated myelopoiesis in patients receiving venetoclax for CLL. Blood *139*, 1198–1207. https://doi.org/10.1182/blood.2021012775.
110. Moujalled, D.M., Brown, F.C., Chua, C.C., Dengler, M.A., Pomilio, G., Anstee, N.S., Litalien, V., Thompson, E., Morley, T., MacRaild, S., et al. (2023). Acquired mutations in BAX confer resistance to BH3-mimetic therapy in acute myeloid leukemia. Blood *141*, 634–644. https://doi.org/10.1182/blood.2022016090.
111. Kitamura, T., Tange, T., Terasawa, T., Chiba, S., Kuwaki, T., Miyagawa, K., Piao, Y.F., Miyazono, K., Urabe, A., and Takaku, F. (1989). Establishment and characterization of a unique human cell line that proliferates dependently on GM-CSF, IL-3, or erythropoietin. J. Cell. Physiol. *140*, 323–334. https://doi.org/10.1002/jcp.1041400219.
112. Fu, N.Y., Sukumaran, S.K., Kerk, S.Y., and Yu, V.C. (2009). Baxbeta: a constitutively active human Bax isoform that is under tight regulatory control by the proteasomal degradation mechanism. Mol. Cell *33*, 15–29. https://doi.org/10.1016/j.molcel.2008.11.025.
113. Steensma, D.P., Bejar, R., Jaiswal, S., Lindsley, R.C., Sekeres, M.A., Hasserjian, R.P., and Ebert, B.L. (2015). Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood *126*, 9–16. https://doi.org/10.1182/blood-2015-03-631747.
114. Malcovati, L., Gallì, A., Travaglino, E., Ambaglio, I., Rizzo, E., Molteni, E., Elena, C., Ferretti, V.V., Catricalà, S., Bono, E., et al. (2017). Clinical significance of somatic mutation in unexplained blood cytopenia. Blood *129*, 3371–3378. https://doi.org/10.1182/blood-2017-01-763425.
115. Mian, S.A., Rouault-Pierre, K., Smith, A.E., Seidl, T., Pizzitola, I., Kizilors, A., Kulasekararaj, A.G., Bonnet, D., and Mufti, G.J. (2015). SF3B1 mutant MDS-initiating cells may arise from the haematopoietic stem cell compartment. Nat. Commun. *6*, 10004. https://doi.org/10.1038/ncomms10004.
116. Fabre, M.A., de Almeida, J.G., Fiorillo, E., Mitchell, E., Damaskou, A., Rak, J., Orrù, V., Marongiu, M., Chapman, M.S., Vijayabaskar, M.S., et al. (2022). The longitudinal dynamics and natural history of clonal haematopoiesis. Nature *606*, 335–342. https://doi.org/10.1038/s41586-022-04785-z.
117. Mortera-Blanco, T., Dimitriou, M., Woll, P.S., Karimi, M., Elvarsdottir, E., Conte, S., Tobiasson, M., Jansson, M., Douagi, I., Moarii, M., et al. (2017). SF3B1-initiating mutations in MDS-RSs target lymphomyeloid hematopoietic stem cells. Blood *130*, 881–890. https://doi.org/10.1182/blood-2017-03-776070.
118. Miles, L.A., Bowman, R.L., Merlinsky, T.R., Csete, I.S., Ooi, A.T., Durruthy-Durruthy, R., Bowman, M., Famulare, C., Patel, M.A., Mendez, P., et al. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature *587*, 477–482. https://doi.org/10.1038/s41586-020-2864-x.
119. Rodriguez-Meira, A., O'Sullivan, J., Rahman, H., and Mead, A.J. (2020). TARGET-seq: a protocol for high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Star Protoc. *1*, 100125. https://doi.org/10.1016/j.xpro.2020.100125.
120. Rodriguez-Meira, A., Norfo, R., Wen, W.X., Chédeville, A.L., Rahman, H., O'Sullivan, J., Wang, G., Louka, E., Kretzschmar, W.W., Paterson, A., et al. (2022). Deciphering TP53 mutant cancer evolution with single-cell multi-omics. Preprint at bioRxiv. https://doi.org/10.1101/2022.03.28.485984.

121. Magee, J.A., and Signer, R.A.J. (2021). Developmental stage-specific changes in protein synthesis differentially sensitize hematopoietic stem cells and erythroid progenitors to impaired ribosome biogenesis. Stem Cell Rep. *16*, 20–28. https://doi.org/10.1016/j.stemcr.2020.11.017.

122. Iskander, D., Wang, G., Heuston, E.F., Christodoulidou, C., Psaila, B., Ponnusamy, K., Ren, H., Mokhtari, Z., Robinson, M., Chaidos, A., et al. (2021). Single-cell profiling of human bone marrow progenitors reveals mechanisms of failing erythropoiesis in Diamond-Blackfan anemia. Sci. Transl. Med. *13*, eabf0113. https://doi.org/10.1126/scitranslmed.abf0113.

123. Stahl, M., Menghrajani, K., Derkach, A., Chan, A., Xiao, W., Glass, J., King, A.C., Daniyan, A.F., Famulare, C., Cuello, B.M., et al. (2021). Clinical and molecular predictors of response and survival following venetoclax therapy in relapsed/refractory AML. Blood Adv. *5*, 1552–1564. https://doi.org/10.1182/bloodadvances.2020003734.

124. Zhang, H., Nakauchi, Y., Köhnke, T., Stafford, M., Bottomly, D., Thomas, R., Wilmot, B., McWeeney, S.K., Majeti, R., and Tyner, J.W. (2020). Integrated analysis of patient samples identifies biomarkers for venetoclax efficacy and combination strategies in acute myeloid leukemia. Nat. Cancer *1*, 826–839. https://doi.org/10.1038/s43018-020-0103-x.

125. Wang, H., Guo, M., Wei, H., and Chen, Y. (2023). Targeting p53 pathways: mechanisms, structures, and advances in therapy. Signal Transduct. Target. Ther. *8*, 92. https://doi.org/10.1038/s41392-023-01347-1.

126. Moura, P.L., Mortera-Blanco, T., Hofman, I.J.F., Todisco, G., Kretzschmar, W.W., Björklund, A.-C., Creignou, M., Hagemann-Jensen, M., Ziegenhain, C., Granados, D.C., et al. (2023). Erythroid differentiation intensifies RNA mis-splicing in SF3B1-mutant myelodysplastic syndromes with ring sideroblasts. Preprint at bioRxiv. https://doi.org/10.1101/2023.04.11.536355.

127. Marasco, L.E., and Kornblihtt, A.R. (2022). The physiology of alternative splicing. Nat. Rev. Mol. Cell Biol. *24*, 242–254. https://doi.org/10.1038/s41580-022-00545-z.

128. Bonnal, S.C., López-Oreja, I., and Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer—implications for care. Nat. Rev. Clin. Oncol. *17*, 457–474. https://doi.org/10.1038/s41571-020-0350-x.

129. Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Cancer Genome Atlas Research Network, Buonamici, S., and Yu, L. (2018). Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell Rep. *23*, 282–296.e4. https://doi.org/10.1016/j.celrep.2018.01.088.

130. Yoshida, K., and Ogawa, S. (2014). Splicing factor mutations and cancer. Wiley Interdiscip. Rev. RNA *5*, 445–459. https://doi.org/10.1002/wrna.1222.

131. Ogawa, H., Horitani, K., Izumiya, Y., and Sano, S. (2022). Somatic mosaicism in biology and disease. Annu. Rev. Physiol. *84*, 113–133. https://doi.org/10.1146/annurev-physiol-061121-040048.

132. Youssoufian, H., and Pyeritz, R.E. (2002). Mechanisms and consequences of somatic mosaicism in humans. Nat. Rev. Genet. *3*, 748–758. https://doi.org/10.1038/nrg906.

133. Nam, A.S., Dusaj, N., Izzo, F., Murali, R., Myers, R.M., Mouhieddine, T.H., Sotelo, J., Benbarche, S., Waarts, M., Gaiti, F., et al. (2022). Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. Nat. Genet. *54*, 1514–1526. https://doi.org/10.1038/s41588-022-01179-9.

134. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

135. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. https://doi.org/10.1038/s41586-019-0969-x.

136. Mulè, M.P., Martins, A.J., and Tsang, J.S. (2022). Normalizing and denoising protein expression data from droplet-based single cell profiling. Nat. Commun. *13*, 2099. https://doi.org/10.1038/s41467-022-29356-8.

137. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

138. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10* (2), giab008.

139. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res *27*, 491–499.

140. Tickle, T., Tirosh, I., Georgescu, C., Brown, M., and Haas, B. (2022). infercnv: infer copy number variation from single-cell RNA-seq. data. (Bioconductor). https://doi.org/10.18129/B9.bioc.infercnv.

141. Müller, S., Cho, A., Liu, S.J., Lim, D.A., and Diaz, A. (2018). CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. Bioinformatics *34*, 3217–3219. https://doi.org/10.1093/bioinformatics/bty316.

142. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. *28*, 396–411. https://doi.org/10.1101/gr.222976.117.

143. Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. *19*, 40. https://doi.org/10.1186/s13059-018-1417-1.

144. Nowicka, M., and Robinson, M.D. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Res *5*, 1356. https://doi.org/10.12688/f1000research.8900.2.

145. Mouhieddine, T.H., Sperling, A.S., Redd, R., Park, J., Leventhal, M., Gibson, C.J., Manier, S., Nassar, A.H., Capelletti, M., Huynh, D., et al. (2020). Clonal hematopoiesis is associated with adverse outcomes in multiple myeloma patients undergoing transplant. Nat. Commun. *11*, 2996. https://doi.org/10.1038/s41467-020-16805-5.

146. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296. https://doi.org/10.1186/s13059-019-1874-1.

147. Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., and White, J.-S.S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. *24*, 127–135. https://doi.org/10.1016/j.tree.2008.10.008.

148. Vaquero-Garcia, J., Norton, S., and Barash, Y. (2018). LeafCutter vs. MAJIQ and comparing software in the fast moving field of genomics. Preprint at bioRxiv. https://doi.org/10.1101/463927.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CITE-seq ADTs (see Table S2)	BioLegend TotalSeq-A	N/A
CD34-PE-Vio770	Miltenyi Biotec	clone AC136; RRID: AB_2660374
BAX	Santa Cruz Biotechnology	sc-7480; RRID: AB_626729
BAK	Thermo Fisher	MA5-36225; RRID: AB_2884059
GFP	Cell Signaling Technology	2555S; RRID: AB_10692764
GAPDH	Cell Signaling Technology	5174S; RRID: AB_10622025
FLAG	Sigma Aldrich	F1804; RRID: AB_262044
Bacterial and virus strains		
Lentiviral expression vector	Addgene	RRID: Addgene_128055
Neomycin lentiviral expression vector	Addgene	RRID: Addgene_139449
Tetracycline inducible lentiviral expression vector	Addgene	RRID: Addgene_162823
Chemicals, peptides, and recombinant proteins		
FSC	TaKaRa	#631106
Human GM-CSF	R&D Systems	215-GM
DAPI	Sigma-Aldrich	D9542
Annexin V	BioLegend	640920
RIPA buffer	Cell Signaling Technology	N/A
Intercept Blocking Buffer	LI-COR	N/A
Critical commercial assays		
Ligation Sequencing Kit	Oxford Nanopore Technologies	SQK-LSK110 SQK-LSK114
LongAmp Taq 2X Master Mix	New England BioLaboratories	M0287S
Chromium 3' (v.3.1 chemistry)	10x Genomics	N/A
SPRI beads	Beckman Coulter Life Science	B23317
MycoAlert Mycoplasma Detection Kit	Lonza	LT07-701
MycoAlert Assay Control Set	Lonza	LT07-518
True-Stain Monocyte Blocker	BioLegend	Cat#422302; RRID: AB_2818986
Deposited data		
scRNAseq CH01-02, CH04, MDS01-06 and AML01A-B (See Table S1), raw (FASTQ) and processed (gene matrix counts, barcodes, features, and isoform junction counts) samples.	This paper	GEO: GSE204845 EGA: EGAS00001007402
Human reference GRCh38 (GENCODE v32/ Ensembl 98)	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build#GRCh38_2020A
MSigDB C2 curated gene sets	GSEA	RRID: SCR_016863; https://www.gsea-msigdb.org/gsea/msigdb/human/collection_details.jsp#C2
Human transcript reference GRCh38.p12 (v31)	GENCODE	RRID: SCR_014966; https://www.gencodegenes.org/human/release_31.html
Experimental models: Cell lines		
TF-1	ATCC	CRL-2003; RRID: CVCL_0559
TF1-Cas9 BAX and BAK DKO cells	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oligonucleotides		
Used in GoT and cDNA ONT sequencing; KO experiments (see Table S1)	This paper	N/A
Software and algorithms		
Cell Ranger (v.3.1.0)	10x Genomics	RRID: SCR_017344; https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/3-1
Seurat (v.3.1)	Stuart et al.[134]	RRID: SCR_016341; https://github.com/satijalab/seurat/releases/tag/v3.1.0
Monocle3 (v.1.0)	Cao et al.[135]	RRID: SCR_018685; https://github.com/cole-trapnell-lab/monocle3
IronThrone (v.2.1)	Nam et al.[133]	https://github.com/landau-lab/GoT-IronThrone
lme4 (v.1.2-1)	Bates et al.	RRID: SCR_015654; https://github.com/lme4/lme4
dsb (v.0.1.0)	Mulè et al.[136]	https://github.com/niaid/dsb
Guppy (v.3.0.6 - 4.0.11)	NanoporeTech	RRID: SCR_023196; https://github.com/nanoporetech/pyguppyclient
SiCeLoRe (v.1.0)	Lebrigand et al.[67]	RRID: SCR_018550; https://github.com/ucagenomix/sicelore
minimap2 (v.2.17)	Heng Li[137]	RRID: SCR_018550; https://github.com/lh3/minimap2/releases/tag/v2.17
SAMtools (v.1.9)	Danecek et al.[138]	RRID: SCR_002105; https://github.com/samtools/samtools/releases/tag/1.9
umi-tools (v.1.1.2)	Smith et al.[139]	RRID: SCR_017048; https://github.com/CGATOxford/UMI-tools/releases/tag/1.1.2
InferCNV (v.1.4.0)	Tickle et al.[140]	RRID: SCR_021140; https://github.com/broadinstitute/infercnv
CONICSmat (v.0.0.0.1)	Müller et al.[141]	https://github.com/diazlab/CONICS
FLAMES (v.1.3.4)	Tian et al.[68]	https://github.com/OliverVoogd/FLAMES
IsoQuant (v.3.1.1)	Prjibelski et al.[76]	https://github.com/ablab/IsoQuant/releases/tag/v3.1.1
SQANTI3 (v.5.1.1)	Tardaguila et al.[142]	https://github.com/ConesaLab/SQANTI3
SUPPA2 (v.2.3)	Trincado et al.[143]	https://github.com/comprna/SUPPA
DRIMSeq (v.1.22.0)	Nowicka and Robinson[144]	https://bioconductor.org/packages/release/bioc/html/DRIMSeq.html
ONT-Splice (v.1.0.0)	This paper	https://doi.org/10.5281/zenodo.8084364;https://github.com/landau-lab/ONT-sc-splice

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dan A. Landau (dlandau@nygenome.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The processed single cell RNA-Seq data are available through the NCBI Gene Expression Omnibus (GEO) and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. De-identified patient FASTQ files have been deposited at the European Genome–phenome Archive (EGA) and accession numbers are listed in the key resources table. They are available upon request if access is granted.
- All original code has been deposited at GitHub and Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

The study was approved by the local ethics committee and by the Institutional Review Board (IRB) of Weill Cornell Medicine, University of Manchester, and Dana-Farber Cancer Institute, conducted in accordance with the Declaration of Helsinki protocol. Cryopreserved mononuclear cells isolated from bone marrow biopsies from myelodysplastic syndrome patients with *SF3B1* mutations were retrieved from Memorial Sloan Kettering and University of Manchester. Additionally, cryopreserved G-CSF mobilized stem cell grafts (without additional mobilizing agents such as plerixafor or cyclophosphamide) from CH patients with *SF3B1* mutations were retrieved from the Dana-Farber Cancer Institute and the Weizmann Institute of Science (Table S1). To confirm the absence of additional genetic mutations, CH samples were sequenced using a previously described panel[145] that includes myeloma driver mutations as well as CH-specific mutations. All samples underwent ultra-low pass whole genome sequencing, rejecting the presence of tumor contamination. Cryopreserved mononuclear cells and grafts were thawed and stained using standard procedures. Cells were first incubated with Human FcX blocking solution (BioLegend, #422302) and then incubated with the surface antibody CD34-PE-Vio770 (clone AC136, lot #5180718070, dilution 1:50, Miltenyi Biotec) and DAPI (Sigma-Aldrich) for 10 minutes at 4°C. Cells were then sorted for DAPI-negative, CD34+ cells using BD Influx at the Weill Cornell Medicine flow cytometry core.

Cell lines and tissue culture

TF-1 human erythroleukemia cell line was purchased from ATCC. All TF-1 generated cell lines were maintained in RPMI + 10% FCS or RPMI + tetracycline-free FCS (TaKaRa #631106) with 2ng/mL recombinant human GM-CSF (R&D Systems; 215-GM) unless otherwise noted. All cell lines were cultured at 37°C and 5% CO_2 in the presence of penicillin (100 U/mL) and streptomycin (100 μg/mL). All cell lines were *Mycoplasma*-free and routinely tested by Antibody and Bioresource Core at MSKCC (MycoAlert Mycoplasma Detection Kit, Lonza, LT07-701; MycoAlert Assay Control Set, Lonza, LT07-518).

Generation of cell lines, virus packaging, and transduction

TF-1 Cas9 stably expressing cells were generated utilizing a lentiviral expression vector from Addgene (#108100) after puromycin mammalian antibiotic selection marker was exchanged for blasticidin. 3 sgRNAs targeting human *BAX* and 3 sgRNAs targeting human *BAK1* were cloned into a dsRED lentiviral expression vector (Addgene #128055) or neomycin lentiviral expression vector (#139449), respectively. Each sgRNA was tested individually and in combination to identify TF-1 Cas9 cells with the best knockout of human *BAX* and *BAK1*. TF-1 Cas9 *BAX* and *BAK* DKO cells were generated by transducing TF-1 Cas9 cells with dsRED lentiviral expression vector with sgRNA CGAGTGTCTCAAGCGCATCG targeting human *BAX* and neomycin lentiviral expression vector with sgRNA CATGAAGTCGACCACGAAGC targeting human *BAK1*. Cells were selected by flow sorting for dsRED or by mammalian antibiotic selection with neomycin (2mg/mL). 3xFLAG tagged BAX isoforms α (NM_138761.4) β (NM_004324.4) and ω (also known as transcript 1, NM_001291428.2), were cloned into a tetracycline inducible lentiviral expression vector (Addgene #162823) and selected with puromycin (10μg/mL). For cloning, vectors were PCR amplified and inserts were generated by gBlock (IDT) followed by Gibson Assembly. Lentiviral supernatants were produced by transfecting HEK293T cells with lentiviral constructs and packaging plasmids pVSVG and psPAX2 using PEI. Virus supernatants were collected and used for transduction in the presence of polybrene (4μg/mL).

Cytokine depletion assay

TF-1 Cas9 DKO BAX isoform cells were washed twice with RPMI 10% FCS without GM-CSF. Cells were then plated in triplicate in a 24 well plate in the presence or absence of doxycycline (1ug/ml). Cells were stained with annexin V (BioLegend 640920)/DAPI (Sigma Aldrich D9542) and assessed for apoptosis by flow cytometry.

Western blots

For western blot analysis, cells were lysed with RIPA buffer (Cell Signaling Technology) containing protease/phosphatase inhibitor cocktail (Sigma Aldrich). Protein concentration was measured using the BCA Protein Assay Kit (Pierce). Equivalent amounts of each sample were loaded on 4-12% Bis-Tris gels (Invitrogen), transferred to 0.2μm PVDF membrane, and blotted with Intercept Blocking Buffer (Li-Cor). The following antibodies were used for western blot analysis: BAX (Santa Cruz Biotechnology sc-7480), BAK1 (ThermoFisher, MA5-36225), GFP (Cell Signaling Technology, 2555S), GAPDH 1:5000 (Cell Signaling Technology 5174S), FLAG 1:100 (Sigma Aldrich, F1804). All primary antibodies were diluted to final concentration of 1:1,000 in Intercept Blocking Buffer (LI-COR) unless otherwise noted.

METHOD DETAILS

GoT-Splice with CITE-seq

GoT-Splice with CITE-seq integrates Genotyping of Transcriptomes (GoT) with both long-read single-cell transcriptome profiling (with Oxford Nanopore Technologies [ONT]) and proteogenomics (with CITE-seq). GoT was performed as previously described.[15] For samples without CITE-seq, CD34+ cells were sorted, and RNA was prepared for sequencing following the standard 10x Genomics Chromium 3' (v.3.1 chemistry) protocol and according to manufacturer's recommendations for the generation of scRNA-seq

libraries (Figure 1A). For GoT-Splice samples that were processed with CITE-seq, prior to sorting, cells were blocked with FcX block for 15 minutes prior to being stained with Total-SeqA antibodies for 30 minutes on ice (see Table S2 for list of antibodies used). The standard 10x Genomics Chromium 3' (v.3.1 chemistry) and CITE-seq protocols[35,36] were carried out according to manufacturer's recommendations for the generation of scRNA-seq and ADT libraries (Figure 1A). At the cDNA amplification step in the 10x Genomics protocol, 1 μL of 1 μM spike-in primer (5'-GATCCTCGTCCTCATTGAACCGC-3') was added to increase the yield of *SF3B1* cDNA and 1 μL of 0.2 μM ADT PCR additive primer (5' – CCTTGGCACCCGAGAATTCC – 3') was added to amplify ADT. After cDNA amplification and a double-sided cleanup with SPRI beads to separate cDNA and ADT fractions, the ADT fraction was amplified for 10 cycles with SI-PCR oligo (10x Genomics) and TruSeq Small RNA RPI-x (Illumina) primers to index the samples. SPRI was used to clean up the ADT final products. In both samples in which CITE-seq was conducted and not conducted, cDNA was allocated for gene expression library creation (standard 10x protocol; 25% of cDNA), targeted genotyping (10% of cDNA), and ONT sequencing with biotin enrichment (10 ng of cDNA). Any remaining cDNA was stored. For locus-specific amplification (GoT), two serial PCRs were performed with nested reverse primers, based on the *SF3B1* mutation of interest. For mutations upstream of K700E, (5'-GATCCTCGTGGTCATT GAACCGC-3' and 5'-CACCCGAGAATTCCAGGCTACTATGATCTCTACCATGAGACCTG-3') and, for K700E mutations, (5'-GTGC AAAAGCAAGAAGTCCT-3' and 5'-CACCCGAGAATTCCATGAACATGGTCTTGTGGATGAG-3') were used as reverse primers. These reverse primers and the generic forward SI-PCR amplify the site of interest from the cDNA template (10 PCR cycles each). The second locus-specific reverse primers contain a partial Illumina TruSeq Small RNA read 2 handle and a locus-specific region to allow *SF3B1* specific priming. The SI-PCR oligo (10x Genomics) anneals to the partial Illumina TruSeq read 1 sequence, preserving the cell barcode (CB) and unique molecule identifier (UMI). After these rounds of amplification and SPRI purification to remove unincorporated primers, a third PCR was performed with a generic forward PCR primer (P5_generic, 5' – AATGATACGGCGACCACCGA GATCTACAC – 3') to retain the CB and UMI together with an RPI-x primer (Illumina) to complete the P7 end of the library and add a sample index (6 PCR cycles). Gene expression, ADT, and *SF3B1* amplicon libraries were pooled to receive 25,000, 5,000, and 5,000 reads per cell, respectively, during Illumina sequencing. The cycle settings were as follows: 28 cycles for read 1, 90 cycles for read 2, 10 cycles for i7, and 10 cycles for i5 sample index. To examine splicing patterns broadly in the whole transcriptome, full-length cDNA was sequenced using the Oxford Nanopore Technologies sequencing on PromethION and GridION flow cells. To enrich for transcripts that contain CBs and UMIs and decrease the presence of PCR artifacts, on-bead PCR with a biotinylated primer selecting for an adapter upstream of the CB was completed[67] (Figure 2A). In brief, 10 ng of full-length cDNA was amplified with LongAmp master mix (NEB) and TSO (5'-NNNAAGCAGTGGTATCAACGCAGAG-3') and biotinylated read 1 (5'-/5Biosg/AAAAACTA CACGACGCTCTTCCGATCT-3') primers for 5 cycles. M270 streptavidin beads (ThermoFisher) were washed with 1X SSPE buffer, resuspended in 5X SSPE buffer and incubated with PCR amplicon after clean-up with 0.8X SPRI beads. After a 15-minute incubation, the beads were washed with 1X SSPE and 10 mM Tris-HCl (pH 8) resuspended in PCR master mix, and further amplified with LongAmp master mix, TSO and read 1 (5' – NNNCTACACGACGCTCTTCCGATCT – 3') primers for 5 cycles. After cleanup with SPRI, 100-300 ng of each full-length cDNA library was sequenced on one PromethION or GridION flow cell with SQK-LSK110.

GoT in U2AF1 was performed with a similar protocol to the described above, targeting the U2AF1 S34F mutation using a cDNA spiked primer to enrich for the transcript (5' - GCCTCCATCTTCGGCACCGAGA - 3') and 2 PRC rounds (PCR0 - Forward: 5'-GCCTCCATCTTCGGCACCGAGA-3' PCR0 – Reverse: 5'- CTACACGACGCTCTTCCGATCT -3' and PCR1 - Forward: 5'- CACCC GAGAATTCCAGCATGTCGTCATGGAGACAGGTGC-3' PCR1 - Reverse: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTC -3') with the same specifications as SF3B1. The full-length GoT library was sequenced on a MinION flow cell with SQK-LSK114.

ScRNA-seq Illumina data processing, alignment, clustering, and cell-type classification

10x Illumina data were processed using Cell Ranger (v.3.1.0) with default parameters and reads were aligned to the human reference sequence GRCh38. For all samples, the Seurat package (v.3.1) was used to perform QC filtering, and unbiased clustering of CD34+ sorted cells.[146] As an overview, for each sample dataset, cells with number of UMIs (nCount_RNA) < 1,500 or nCount_RNA > 3 median absolute deviation above the median nCount_RNA value, number of unique genes (nFeature_RNA) > 3 median absolute deviation above the median nFeature_RNA value and mitochondrial gene percentage (perc.mito) > 20% were filtered. Using the SCTransform function, each dataset was log normalized using the default scale factor of 10,000, scaled and potential confounders (such as nCount_RNA, perc.mito and S phase and G2M phase gene expression scores) were regressed out of the data. SCTransform also identified the top 3,000 variable genes found in each dataset that are used for integration. Before clustering, the individual datasets were integrated based on disease status (i.e. primary MDS samples, MDS01-03, were integrated together, MDS validation samples, MDS04-06, from patient treated with growth factors at the time of biopsy were integrated together and then the CH samples, CH01-02, were integrated together) and underwent batch correction within Seurat which implements canonical correlation analysis (CCA) and the principles of mutual nearest neighbors (MMN).[134] For integration, 30 canonical vectors were used for the CCA in the FindIntegrationAnchors function, and 30 principal components were used for the anchor weightings step in the IntegrateData function (as recommended in Seurat). Next, a principal component analysis (PCA) was performed using the variable genes of the integrated dataset and the JackStraw method was used to determine statistically significant principal components (PCs) to be used as inputs into the UMAP algorithm for cluster visualization. Clustering was performed with the FindNeighbors (using only significant PCs) and the FindClusters (resolution = 2) functions which rely on the k-nearest neighbors (KNN) algorithm to identify cell clusters. Unique clusters were manually assigned based on differentially expressed genes identified with the FindAllMarkers function which looked only at genes found in at least 25% of cells in either of the two input comparison groups and only returned results for genes

with at least a 0.25 log transformed fold change between groups. More specifically, cluster annotations were made according to the differential expression of canonical lineage marker genes identified in previous single-cell RNA-seq data of normal hematopoietic progenitor cells[37] (Table S1). Clusters with similar increased expression of these canonical markers were merged to form the main progenitor subsets: HSPCs, IMPs, NPs, MkPs, MEP, EPs, Pre-Bs and E/B/Ms in the primary MDS, MDS validation and CH cohort as well as Mono, MonoDCs, DCs, B cells and T cells in MDS and MDS validation. Finally, pseudotime analysis was performed using the Monocle3 R package with recommended parameters (v.0.2.1).[135]

IronThrone GoT for processing targeted amplicon sequences and performing mutation calling

Genotyping of single cells was carried out with the IronThrone (v.2.1) pipeline as previously described.[15,133] In brief, individual amplicon reads were assessed for the appropriate structure (*i.e.,* presence of the primer sequence and the expected sequence between the primer and given mutation site) and all reads were assessed for a matching cell barcode to the list generated from the 10x paired GEX dataset. A Levenshtein distance of 0.1 was allowed for all sequence matching and collapsing steps and only UMIs with a minimum of 2 supporting reads were retained for final genotyping. Following UMI collapse, genotype assignment of individual UMIs was conducted as described previously with majority rule of supporting reads for wildtype or mutant status (using a 0.7 PCR read ratio, above which the majority of PCR reads must be for a UMI to be called definitively). Rare UMIs that did not pass this threshold were removed as ambiguous. Additionally, to remove reads that result from PCR recombination, UMIs in the amplicon library that match UMIs of non-*SF3B1* genes in the gene expression library were discarded (as described in the IronThrone GoT pipeline).[15,133] Finally, given the heterozygous nature of these *SF3B1* mutations, each single cell was assigned as either mutant (MUT) or wildtype (WT) as follows: cells with at least 1 mutant UMI were assigned as MUT cells and cells with 0 mutant UMIs and at least 1 wildtype UMI were assigned as WT. As benchmarking, the *SF3B1* genomic regions of interest that were used for GoT were examined in each matching GEX library to determine how many UMIs were able to successfully capture the targeted sequence in conventional 10x data and, in all cases, less UMIs were captured in the GEX library (Figures S2B and S6E). While the genotyping information is derived from transcribed molecules alone and may be affected by whether transcripts from wildtype versus mutant alleles were expressed and/or captured, the fraction of MUT cells as determined by GoT using all cells with at least 1 UMI yielded similar values to those determined by bulk DNA exon sequencing (Figure S2A). Despite this, we systematically applied specific approaches to exclude the effect of this confounder (that is, the expression level of the target gene) on the conclusions of other downstream analyses. First, to rule out the possibility that higher *SF3B1* expression results in a greater ability to detect mutant alleles, and thereby in a higher mutant-cell frequency, we downsampled all cells to a single amplicon UMI before mutation calling when conducting the mutant-cell frequency analyses. Then, for the remaining of our downstream analyses between *SF3B1* mutant and wildtype cells (except for the differential gene expression and gene set enrichment analyses in CH due to low fraction of mutated cells, which decreases the likelihood of misclassifying mutant as wildtype), we took the more conservative approach considering only genotyped cells with two or more genotyping amplicon UMIs.

Mutant cell frequency

The frequency of mutant cells, as determined by GoT, was assessed as previously performed in Nam et al.[133] Firstly, we used only cells with at least 1 UMI and only considered cell types with at least 300 genotyped cells. To account for the potential confounding effect of a heterozygous mutation as well as variable *SF3B1* expression, we performed amplicon UMI downsampling to 1 UMI per genotyped cell prior to mutation calling for calculating MUT cell frequencies. An equal number of cells from each sample within the MDS cohort, were subsampled randomly for the integrated data to ensure equal representation from each patient. Genotyping amplicon UMIs were downsampled (x100 iterations) to 1 UMI per cell and MUT cell frequency was determined for each progenitor cluster for either the integrated dataset or individual samples. This frequency was then divided by the total mutant cell frequency across all progenitor subsets for each of the iterations. Linear mixed effects analysis was performed using the lme4 package (v.1.2-1). Progenitor identity was defined as the fixed effect, and for random effects, we used intercepts for individual patients (subjects) and iterative downsampling. p values were obtained by likelihood ratio tests of the full model with the fixed effect against the model without the fixed effect.[147]

Differential gene expression and gene set enrichment

The differential gene expression analysis (DGEA) comparing WT and MUT cells and gene set enrichment analysis (GSEA) were performed as done in Nam et al.[133] In brief, for each cohort we used a within-sample permutation test for the analysis of each progenitor cell subtype. To ensure equal representation from each patient, we downsampled the total number of mutated and wildtype cells to the same number across all patients. The observed $\log_2$ fold change values were calculated comparing the MUT versus WT cells for the tested genes. The tested genes included the top 2,000 most variable genes (excluding mitochondrial genes) which were filtered for those expressed in at least 10% of either group (MUT versus WT), for each progenitor subtype. Next, the WT and MUT labels were shuffled over 100,000 iterations, within each patient, and fold change values were re-calculated to create a background distribution. *P*-values were calculated per gene as a percent of permutations whose absolute fold change values were more extreme than the absolute value of the observed fold change (Tables S3 and S5). Hypergeometric test for GSEA of the integrated differentially expressed genes (p value < 0.05, $\log_2$(fold change) > 0.1) was performed using the Cluster Profile package (v.0.1.9). FDR multiple hypothesis testing correction was performed. MSigDB C2 curated gene sets were included in the analyses (Tables S3 and S5).

ADT processing

CITE-seq was performed on the primary MDS cohort (for samples MDS02-03) and as mentioned above, the 10x Illumina ADT data was processed using Cell Ranger (v.3.1.0) with default parameters and counts were generated for each marker in the CITE-seq panel (Table S2). After using the Seurat package (v.3.1) for QC filtering, and unbiased clustering of the CD34+ sorted cells based on RNA data, ADT data was also normalized using centered log-ratio (CLR) normalization, scaled and the expression of various ADT markers was used in confirming the cell-type assignment of different progenitor subsets. For benchmarking purposes, Seurat's Weighted Nearest Neighbor (WNN) Analysis was also performed, which is a multi-modal analysis that integrates both RNA and ADT data when performing cell clustering. This was used to compare to the clustering output when using the KNN algorithm that relies on RNA data alone (Figure S1H). For the WNN analysis, cells were filtered and integrated using SCTransform (as described above). The RNA data was logNormalized and the ADT data was run through CLR normalization and the RunPCA function for dimensionality reduction was also run independently on each modality. Next, the FindMultiModalNeighbors function was used which for each cell, calculates its closest neighbors in the dataset based on a weighted combination of RNA and ADT similarities. This constructs a WNN graph that was visualized with the RunUMAP function. The cell-type assignments generated from the initial clustering (with RNA data alone) were then projected onto this new UMAP for comparison (Figure S1H).

Denoised scaled by background normalization (DSB) filtering and differential protein expression

We used the dsb package[136] (v.0.1.0) as an alternative form of normalization for the ADT protein expression values. Normalized values were applied for selection filtering of ADT markers for which the true signal was above the background noise levels, within the captured cell-contained droplets. dsb discriminates between background noise by differentiating between empty droplets (containing ambient mRNA and antibody but no cell) and true cell-containing droplets. The background matrix was defined from the comparison of the raw feature barcode matrices from the 10x sequencing output versus the processed filtered feature barcode matrix results generated from running Cell Ranger (see STAR Methods above). The final output filters out empty droplets and retains only true cell-containing droplets based on the 10x cell calling algorithm. As such, the matrix of background noise is generated by subtracting out the positive cell containing droplets found in the filtered matrices from the negative empty droplets in the raw matrices. Furthermore, with an additional filter requiring the removal of drops with protein library size > 1.5 and number of genes < 80 was applied to refine the background noise signal. Normalization was performed using the DSBNormalizeProtein function omitting isotype controls and denoised counts. The dsb normalized values were defined as the number of standard deviations above the background noise and antibodies were then filtered, keeping only those with a dsb normalized expression value of > 2 in at least 1 cell type (Table S2). When performing the differential protein expression analysis across our patient samples, we used an iterative downsampling (x1,000) approach that, at each iteration, randomly samples an equal number of *SF3B1*^{mut} and *SF3B1*^{wt} cells from each patient sample before calculating the median $\log_{10}$FC of protein expression between *SF3B1*^{mut} / *SF3B1*^{wt} cells. This was done to ensure equal representation of genotyped cells from each patient. To calculate the median $\log_{10}$FC of *SF3B1*^{mut} / *SF3B1*^{wt} cells, we first modified the Seurat's FindMarker function to calculate the median instead of the mean expression, a measure that is more robust to outlier values. Then, for each downsampled object we obtained a table containing the $\log_{10}$FC of each antibody per cell type. $\log_{10}$FC matrices are combined by taking the median across the downsampled iterations, resulting in the median $\log_{10}$FC values. Statistical significance was assessed by performing permutation tests (x10,000) within each patient sample matrix. (Figure 1F) shows normalized ADT expression across cell types, using the maximum expression.

ScRNA-seq ONT long-read sequencing data processing, alignment, junction calling and annotation

Guppy (v.3.0.6 - 4.0.11) was used for base calling FAST5 files output from ONT sequencing. We then filtered for only reads containing a polyA tail within 100 base pairs of either 5' or 3' end using the 'NanoporeReadScanner-0.5.jar' within the SiCeLoRe-1.0 workflow. Due to the low number of genotyped cells (139) with two or more genotyping amplicon UMIs, MDS01 was excluded from all downstream splicing analyses. Filtered reads are aligned to the primary human genome, assembly GRCh38.p12 using minimap2 (v.2.17). Minimap2 was used with the '-ax splice' flag to prioritize annotated splice junctions. Additionally, we made use of the '–junc-bed' option, to increase alignment scores for those splice junctions found in the reference junction bed file. For our reference junctions, we used splice junctions from single-cell Smart-Seq2 data from human CD34+ cells obtained from a CH sample with no *SF3B1* mutation. Additionally, we used '–secondary=no' to suppress multi-mappings. In preparation to identify the cell barcodes and UMIs present in the long-read sequencing, we used the 'IlluminaParser-1.0.jar' in SiCeLoRe to parse the cell barcodes and UMIs present in the complementary short-read sequencing library. We continued to use SiCeLoRe to tag the aligned BAM files with cell barcodes and UMIs identified in the short-read library and generate consensus sequences for each unique cell barcode and UMI combination. Consensus sequences were used to create a gene by cell count matrix. For all other steps, we used the default parameters set by SiCeLoRe-1.0, following the workflow found at https://github.com/ucagenomix/sicelore. Intron-junction calling is then performed on consensus sequence BAM files, adapted from the method used in the LeafCutter pipeline for short-read RNA-seq data.[72,148] In brief, the intron-junction calling pipeline utilizes the pysam.fetch() function and iterates through each transcript in the BAM file, noting its cell barcode (CB) tag as well as the coordinates of each intron-junction for that transcript. On iterating through the BAM file, counts for the usage of each unique intron-junction and the corresponding CB are recorded. This ultimately generates an Intron-Junction x Cell Barcode count matrix for the given BAM file. Each intron-junction is then identified using annotations available in the GENCODE GRCh38.p12 v31 basic annotation reference file as canonical 3', canonical 5', alternative 3', alternative 5'. This outputs a metadata file with annotations for each junction corresponding to the junctions of the Intron-Junction x Cell Barcode count matrix. The

metadata included the 3' and 5' sites defining each junction, the distance from the canonical 3' or 5' site end for each start and end site, and the classification of each site. Additionally, junctions that share the same 3' or 5' splice site are classified into ''junction clusters'', providing a cluster coverage which is used in subsequent analyses, such as for calculating the percent spliced in values of different splicing events. Alternative 3' and 5' junctions were further broken down into alternative and cryptic based on the distance the junction was from the canonical splice site. If the alternative splice site was within 100 base pairs of the canonical splice site, it was classified as a cryptic splice site. Given the intron-centric approach of the pipeline, each event could be classified as either annotated, alternative 3', alternative 5', a cryptic 3' splice site, a cryptic 5' splice site or an exon-skipping event (see STAR Methods below for exon-skipping annotation; Tables S4 and S6).

Junction calling and annotation of short-read sequencing data and comparison to long-read sequencing

Using SAMtools (v.1.9), the post-aligned short-read sequencing BAM file (see STAR Methods above) was filtered to include only reads from the filtered cell barcodes (CB) selected through Cell Ranger, to remove all non-primary alignments and reads mapping to multiple genes. Next, using umi-tools (v.1.1.2) we ran umi-tools group with the '–per-cell flag', cell barcode and UMI pairs with only one supporting read were filtered out and umi-tools dedup was used to remove all duplicate reads. Intron junction calling and annotation was performed as described above. To compare the junction recovery across the transcript region, we used the Ensembl annotation database (v.104) to generate a transcript reference and filtered the database to include only protein coding transcripts as well as those with a transcript support level = 1 (i.e., those representing the most well supported transcript for that gene). From here, we calculated the distance of a junction from the end of its transcript by calculating the distance between the 3' end of that junction to the furthermost 3' junction end (which is at the 3' end of the transcript). This was done to avoid any measurement biases due to long UTR annotations.

Copy number variation analysis

The InferCNV package (v.1.4.0)[140] was used to analyze the single cell dataset for any duplications or deletions of entire chromosomes or large chromosome fragments. Briefly, by comparing expression levels of genes annotated by chromosomal position (using the CONICSmat package v.0.0.0.1[141]) to a set of reference cells (in this case, a one-versus-rest comparison of cells by patient of origin), a heatmap of relative expression can be generated and used to identify regions with significantly increased or decreased expression. We removed the few genes for which alternative positions have been reported (<2% of genes). We ran the InferCNV workflow with recommended parameters, using the i6 6-state Hidden Markov model (Figure S6F).

Differential transcript usage

All alternative 3' junctions were filtered to only include those that contained at least 5 total reads. To identify differentially used transcripts between *SF3B1*^{mut} and *SF3B1*^{wt} cells, junction reads were then pseudobulked based on mutation status across all MDS patients or all CH patients. We then computed the $\log_{10}$(odds ratio) of the likelihood of each junction being observed in the MUT cells over the WT cells. The genotype labels of each of the cells was permuted 100,000 times and we then repeated pseudobulking and computation of the $\log_{10}$(odds ratio) of each junction. Permutations of the genotype label were patient aware, so the mutant cell frequency across patients was unchanged for each permutation. The p value was determined based on the likelihood of seeing the observed odds ratio in comparison to the null distribution of the permuted odds ratios for each junction. The same testing was done within each cell type to identify the differentially used junctions between *SF3B1*^{mut} and *SF3B1*^{wt} cells within a specific cell type. We classified junctions as differentially spliced events if they had *P*-value < 0.05 and delta percent spliced in (dPSI) of >= 2 (a positive dPSI here represents a splicing event more highly used in the *SF3B1*^{mut} population of cells). To observe the usage of these differentially spliced cryptic 3' events (p value < 0.05 and dPSI >= 2) across the continuum of erythroid maturation as opposed to within discrete cellular states, erythroid lineage MUT cells (HSPCs, IMPs, MEPs, and EPs) were ordered from least to most differentiated, grouped into bins and the MUT cell PSI for each cryptic event was calculated per bin (Figures 4A and 6D). Specifically, in the primary MDS cohort (Figure 4A), 6301 *SF3B1*^{mut} cells were ordered by the expression of the erythroid marker CD71 (obtained from CITE-seq) and a bin size of 3000 *SF3B1*^{mut} cells, sliding by 300 *SF3B1*^{mut} cells at each step, was used to capture the continuous change in the usage of the different cryptic 3' junctions via the MUT cell PSI measurements per bin. The variance in the usage of each cryptic 3' event was measured by calculating the range of PSIs across all the bins along the continuum and only cryptic junctions that had a PSI range of at least 2 and average coverage across all bins of 10 reads were considered. This approach was taken to focus on cryptic events that had a variable signal and that were also well supported. In CH (Figure 6D), 1,020 MUT cells were ordered by pseudotime and a bin size of 600 *SF3B1*^{mut} cells, sliding by 60 *SF3B1*^{mut} cells at each step, was used to capture the MUT cell PSI per bin. Similarly, only cryptic junctions that had a PSI range of at least 2 and average coverage of 10 reads were considered.

For the *BAX* cryptic event (Figure 6E), to directly compare the per bin PSI values across all 3 cohorts (MDS, MDS validation and CH) we adjusted the bin and window sizes across the cohorts to ensure the same number of final bins for each cohort. To achieve this, we took the following approach: MDS - MUT cells ordered by CD71 expression, window size of 3750 *SF3B1*^{mut} cells, sliding by 375 *SF3B1*^{mut} cells, MDS validation: cells ordered by pseudotime, window size of 580 *SF3B1*^{mut} cells, sliding by 58 *SF3B1*^{mut} cells, CH: cells ordered by pseudotime, window size of 600 *SF3B1*^{mut} cells, sliding by 60 *SF3B1*^{mut} cells. To note, for each of the sliding window analyses, only MUT cells with at least 2 genotyping amplicon UMIs were considered.

Exon skipping and nonsense-mediated decay (NMD) annotations

To identify exon skipping events, for each gene in the GENCODE GRCh38.p12 v31 basic annotation reference file we determined its main functional isoform (as those that belong to the APPRIS database and carry the ''appris_princial'' tags) to compare to the transcript isoforms generated in our data. With this, for a given gene, each identified intron junction within our data was compared to the reference and labeled as an ''exon_skip'' if it excluded any of the exons present in the reference. The number of exons skipped was also recorded. To identify NMD inducing alternative splicing events, we also developed a pipeline that inspects each intron junction in the Intron-Junction x Cell Barcode count matrix and detects the presence of premature termination codons (PTCs) and frameshift events induced because of alternative splicing. In brief, this is done by grabbing the entire nucleotide sequence of a particular isoform noting the position of the last exon-exon junction, finding the position of the first start codon and from there, phasing along the triplets of nucleotides of that given sequence string. By following the known rules of NMD, each intron junction was further annotated as being (i) NMD-inducing (which would lead to NMD of its associated transcript) or (ii) NMD-neutral. Specifically, the 50-nucleotide rule was followed such that an event is labeled NMD-inducing if a PTC is introduced greater than 50 nucleotides away from the last exon-exon junction or NMD-neutral if a PTC is introduced within 50 nucleotides of the transcript's last exon-exon junction. Finally, Intron-junctions were labeled to cause frameshifts if the total number of nucleotides involved in an alternative 3' or 5' splicing event was not divisible by 3.

Motif enrichment analysis

High quality cryptic 3' junctions (MUT read coverage > 3, PSI >= 2, junction cluster read coverage > 20 across at least 2 junction clusters) were obtained from the junction quantification matrix from samples MDS05-06. Each of these cryptic 3' splice sites were then paired to a corresponding canonical junction, requiring both, canonical and cryptic junctions, to be part of the same splicing cluster (as described above). Flanking sequences, 50 nucleotides upstream and 10 nucleotides downstream of the 3' splice site were obtained from the two junction sets and used to calculate position weight matrices (PWM). For each position, a log odds ratio enrichment for each nucleotide was calculated using Fisher's exact test, comparing the cryptic 3' splice site nucleotide composition against the canonical. Reported positions were filtered according to their enrichment significance (p value < 0.05).

Isoform tool comparison

Given that full-length isoform tools provide transcript-level descriptions rather than local events like GoT-Splice (e.g., exon skipping or cryptic splice sites), the comparison to FLAMES[68] and IsoQuant[76] was performed at the local level, overlapping junctions responsible to define the local splicing changes. For this analysis, samples of the discovery cohort (MDS02-03) were used. First, FASTQ files were processed in the FLAMES (v.1.3.4) multi-sample scRNA-seq pipeline with standard parameters and using the same genome references than in GoT-Splice (described above). In parallel, collapsed BAM files resulting from the SiCeLoRe pipeline (see description above), were used as input for IsoQuant (v.3.1.1) in two steps: first all samples were run together, using the 'nanopore' data type input, generating an unified isoform annotation (in GTF format) and subsequently, each sample was processed individually to generate isoform per cell barcode matrices, using the previously created GTF as reference, which includes new isoforms detected across all the samples. To extract the junction annotations and assess quality of the recovered transcript annotations, SQANTI3[142] (v.5.1.1) QC script was applied, with default parameters, to the GTF isoform annotations produced by FLAMES and IsoQuant. Local splicing events (cryptic splice sites and exon skipping) in the filtered SQANTI3 GTF annotations were then produced with the SUPPA2[143] (v.2.3) 'generateEvents' option. PSI values were estimated following a similar approach to the one described in SUPPA2, briefly, PSI values reflect a ratio between the read counts corresponding to isoforms that include a particular event divided by the total isoforms over a particular region (isoforms that do not include the event + isoforms including the events). To identify MUT versus WT isoform proportion changes we merged the sample transcript per cell count matrices obtained in FLAMES or IsoQuant and using DRIMSeq[144] (v.1.22.0) we filtered for a minimal gene expression of 5, minimal transcript expression of 2 and gene expressed in a minimum of 2 cells and performed a likelihood ratio test (via the dmTest function). Isoforms with FDR adjusted p value < 0.05 and change in proportion > 0.25 were considered as significantly changed. Estimation of a splicing aberration bias in the significantly changed isoforms was determined by overlapping the isoforms with cryptic and alternative 3' ss annotated with GoT-Splice.

QUANTIFICATION AND STATISTICAL ANALYSIS

Categorical variables were compared using the hypergeometric test or Fisher's Exact test. Continuous variables were compared using the Wilcoxon rank-sum test, Student's t-test, non-parametric permutation test or Kolmogorov–Smirnov test, as appropriate. p values were adjusted for multiple comparisons by Benjamini-Hochberg FDR adjustment procedure. All *P*-values are two-sided and considered significant at the 0.05 level unless otherwise noted. To add further stringency and confidence to the results, we have independently analyzed a distinct cohort of samples (validation cohort) and specifically focused on reporting results that passed a statistical cutoff of 0.2 for FDR adjusted p values in both cohorts. We report genes with FDR adjusted p value < 0.05 in either cohort in Tables S4 and S6.