Figure S1

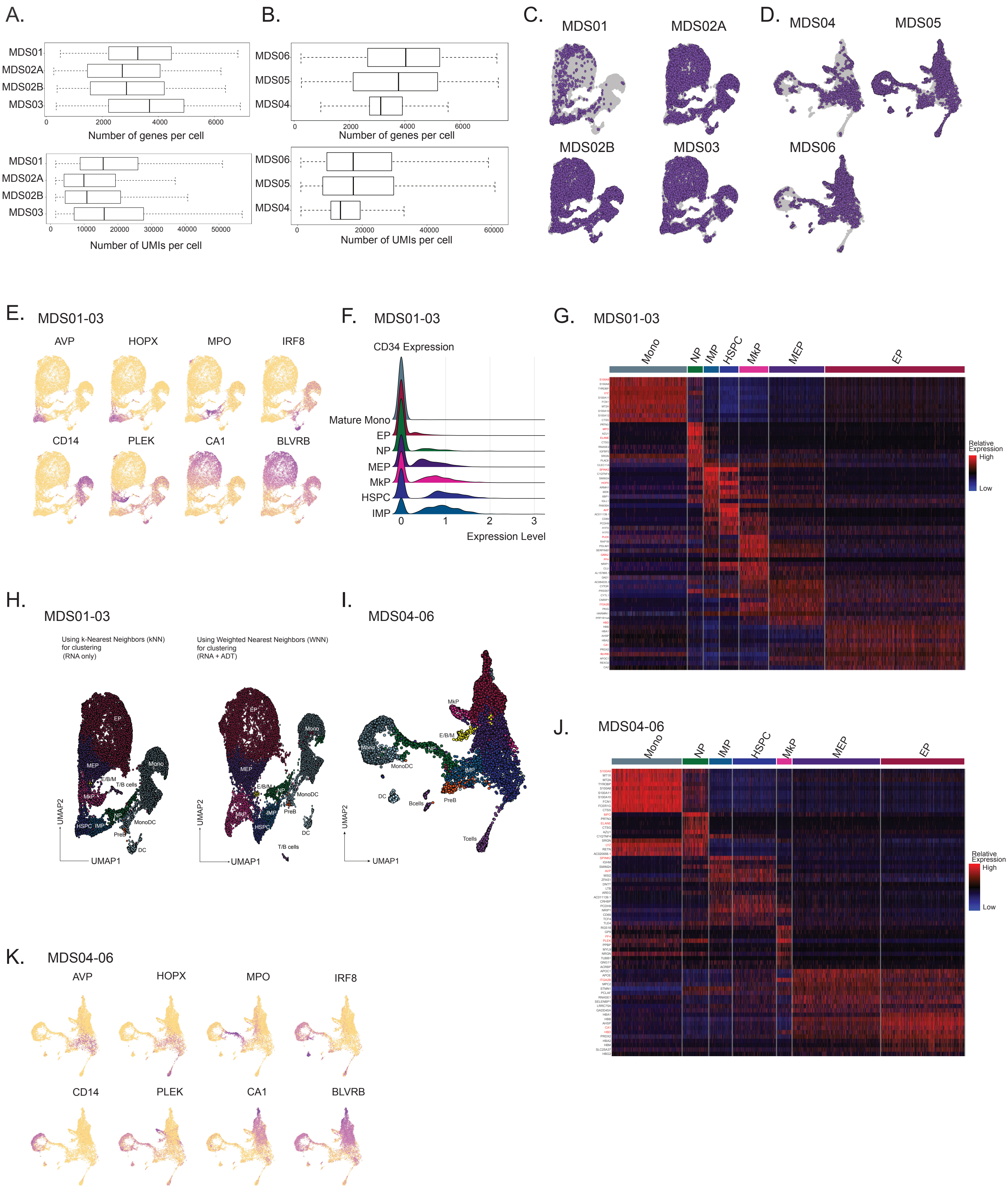
A.
MDS01
MDS02A
MDS02B
MDS03
0
2000
4000
6000
Number of genes per cell
MDS01
MDS02A
MDS02B
MDS03
0
10000
20000
30000
40000
50000
Number of UMIs per cell
B.
MDS06
MDS05
MDS04
0
2000
4000
6000
Number of genes per cell
MDS06
MDS05
MDS04
0
20000
40000
60000
Number of UMIs per cell
C.
MDS01
MDS02A
MDS02B
MDS03
D.
MDS04
MDS05
MDS06
E.
MDS01-03
AVP
HOPX
MPO
IRF8
CD14
PLEK
CA1
BLVRB
F.
MDS01-03
CD34 Expression
Mature Mono
EP
NP
MEP
MkP
HSPC
IMP
0
1
2
3
Expression Level
G.
MDS01-03
Mono
NP
IMP
HSPC
MkP
MEP
EP
Relative Expression
High
Low
H.
MDS01-03
Using k-Nearest Neighbors (kNN) for clustering (RNA only)
Using Weighted Nearest Neighbors (WNN) for clustering (RNA + ADT)
EP
MEP
E/B/M
T/B cells
MkP
Mono
NP
MonoDC
HSPC
IMP
PreB
DC
UMAP2
UMAP1
EP
Mono
MEP
E/B/M
NP
MonoDC
MkP
IMP
PreB
HSPC
DC
T/B cells
UMAP2
UMAP1
I.
MDS04-06
MkP
E/B/M
Mono
NP
MonoDC
IMP
DC
Bcells
PreB
Tcells
UMAP2
UMAP1
J.
MDS04-06
Mono
NP
IMP
HSPC
MkP
MEP
EP
Relative Expression
High
Low
K.
MDS04-06
AVP
HOPX
MPO
IRF8
CD14
PLEK
CA1
BLVRB

Figure S1. MDS samples QC, integration, clustering, and cell-type assignment, related to Figure 1.

(A) Number of genes per cell (top) and number of UMIs per cell (bottom) in CD34+ sorted hematopoietic progenitors from samples MDS01-03 after QC filters, shown by each patient sample. **(B)** Number of genes per cell (top) and number of UMIs per cell (bottom) in CD34+ sorted hematopoietic progenitors from samples MDS04-06 after QC filters, shown by each patient sample. **(C)** UMAP of CD34+ sorted progenitor cells for each individual sample of MDS01-03 after integration using the Seurat package. **(D)** UMAP of CD34+ sorted progenitor cells for each individual sample of MDS04-06 after integration using the Seurat package. **(E)** Expression of lineage-specific genes from Velten *et al.*[S1] scored and projected onto the UMAP representation of cells from MDS01-03. **(F)** CD34 expression per progenitor cell-type of CD34- monocytes among CD34+ sorted hematopoietic progenitors. **(G)** Heatmap of top 10 differentially expressed genes for each progenitor subset for MDS01-03. **(H)** UMAPs comparing the graph-based clustering output when using the k-nearest neighbors (KNN) algorithm to perform clustering of cells with RNA data alone vs. when using the weighted nearest neighbors (WNN) algorithm that allows for the integration of both RNA and ADT data for clustering cells. The cell-type assignments determined from the KNN-RNA based clustering are projected onto the WNN-RNA+ADT clustering for comparison between the two methods. **(I)** UMAP of CD34+ sorted cells (*n* = 8,879 cells) from samples MDS04-06 with *SF3B1* K700E mutations (*n* = 3), overlaid with cluster cell-type assignments. HSPC, hematopoietic stem progenitor cells; IMP, immature myeloid progenitors; MkP, megakaryocytic progenitors; MEP, megakaryocytic-erythroid progenitors; EP, erythroid progenitors; NP, neutrophil progenitors; E/B/M, eosinophil/basophil/mast progenitor cells; T/B cell progenitors; Mono, monocyte; DC, dendritic cells; Pre-B, precursors B cells; Mono DC, monocyte/dendritic cell progenitors. **(J)** Heatmap of top 10 differentially expressed genes for each progenitor subset for MDS04-06. **(K)** Expression of lineage-specific genes from Velten *et al.*[S1] scored and projected onto the UMAP representation of cells from MDS04-06.

Figure S2

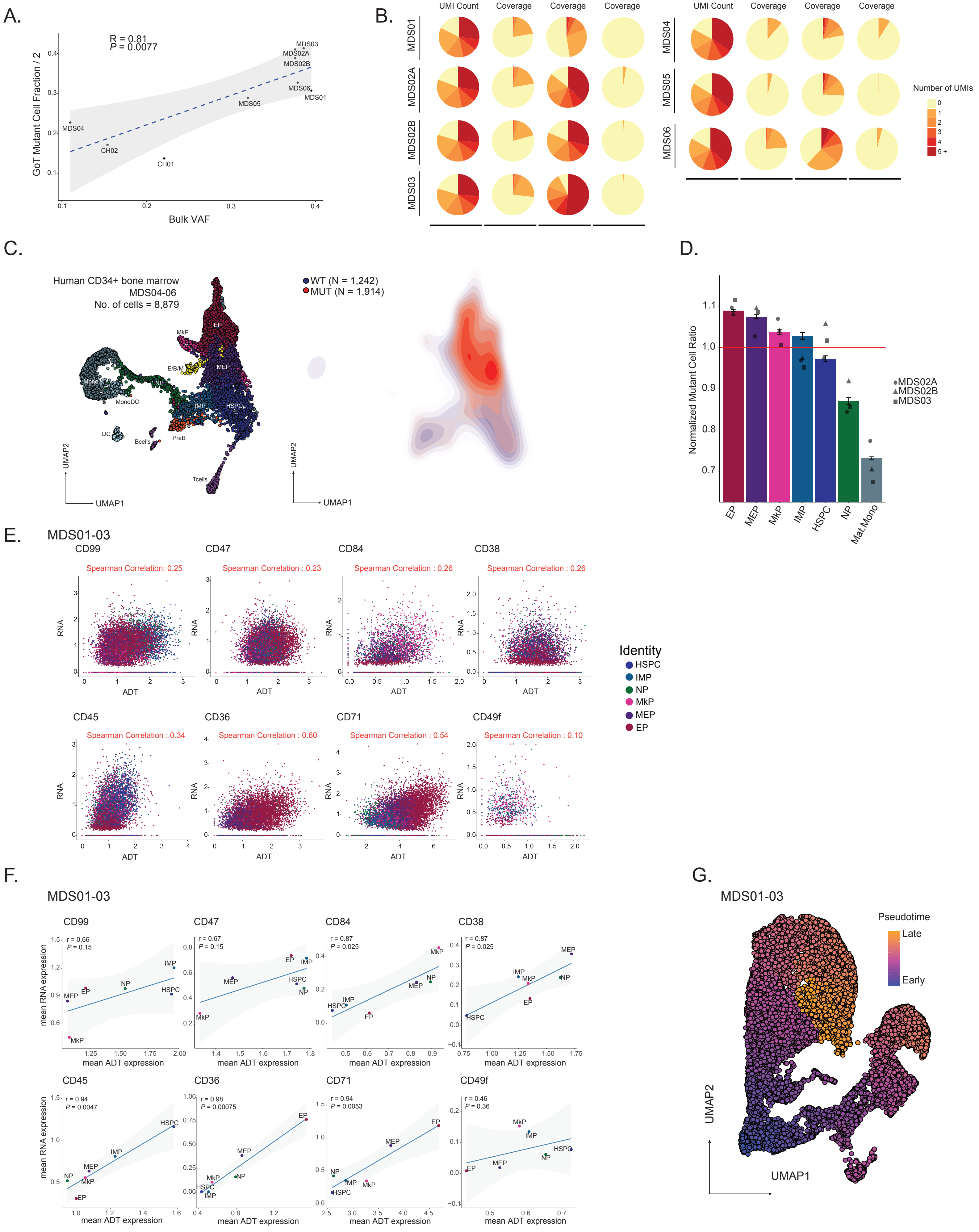
A.
R = 0.81
P = 0.0077
GoT Mutant Cell Fraction / 2
Bulk VAF
MDS03
MDS02A
MDS02B
MDS06
MDS01
MDS05
MDS04
CH02
CH01
B.
10X SF3B1 UMI Count
10X K700E Coverage
GoT K700E Coverage
ONT K700E Coverage
MDS01
MDS02A
MDS02B
MDS03
MDS04
MDS05
MDS06
Number of UMIs
0
1
2
3
4
5 +
C.
Human CD34+ bone marrow
MDS04-06
No. of cells = 8,879
WT (N = 1,242)
MUT (N = 1,914)
EP
MkP
E/B/M
MEP
Mono
NP
MonoDC
IMP
HSPC
DC
Bcells
PreB
Tcells
UMAP2
UMAP1
D.
Normalized Mutant Cell Ratio
MDS02A
MDS02B
MDS03
EP
MEP
MkP
IMP
HSPC
NP
Mat.Mono
E.
MDS01-03
CD99
Spearman Correlation: 0.25
CD47
Spearman Correlation : 0.23
CD84
Spearman Correlation : 0.26
CD38
Spearman Correlation : 0.26
CD45
Spearman Correlation : 0.34
CD36
Spearman Correlation : 0.60
CD71
Spearman Correlation : 0.54
CD49f
Spearman Correlation : 0.10
RNA
ADT
Identity
HSPC
IMP
NP
MkP
MEP
EP
F.
MDS01-03
CD99
r = 0.66
P = 0.15
CD47
r = 0.67
P = 0.15
CD84
r = 0.87
P = 0.025
CD38
r = 0.87
P = 0.025
CD45
r = 0.94
P = 0.0047
CD36
r = 0.98
P = 0.00075
CD71
r = 0.94
P = 0.0053
CD49f
r = 0.46
P = 0.36
mean RNA expression
mean ADT expression
G.
MDS01-03
Pseudotime
Late
Early
UMAP2
UMAP1

Figure S2. MDS samples GoT, CITE-seq, and pseudotime analyses, related to Figure 2.

(A) *SF3B1* K700E (7) and K666N (1) mutant cell fractions determined by GoT in single cells versus *SF3B1* K700E and K666N mutation variant allele frequencies (VAF) determined in bulk sequencing of matched unsorted bone marrow mononuclear cells (MDS) or matched unsorted stem cell product (CH). **(B)** Fraction of cells in MDS samples by number of *SF3B1* UMIs in standard 10x Genomics data without genotyping information, *SF3B1* UMIs with K700E locus coverage in standard 10x data, *SF3B1* UMIs with K700E locus coverage in GoT amplicon library, and *SF3B1* UMIs with K700E locus coverage in ONT library. **(C)** UMAP of CD34+ sorted cells (*n* = 8,879 cells) from samples MDS04-06 with *SF3B1* K700E mutations (*left*) and density plot of *SF3B1*^{mut} vs. *SF3B1*^{wt} cells (*right*). WT, cells with genotype data without *SF3B1* mutation; MUT, cells with genotype data with *SF3B1* mutation. **(D)** Normalized ratio of *SF3B1*^{mut} cells in progenitor subsets with at least 300 genotyped cells. Bars show aggregate analysis of samples MDS01-03 with mean +/- s.e.m. of 100 downsampling iterations to 1 genotyping UMI per cell. Points represent the mean of *n* = 100 downsampling iteration per sample. **(E)** Comparison of the single-cell expression of markers captured in both CITE-seq (x-axes) and RNA-seq (y-axes) libraries. Correlation coefficient r calculated using Spearman's correlation. Cells are colored by each progenitor subset. **(F)** Comparison of the mean expression per progenitor subset of markers captured in both CITE-seq (x-axes) and RNA-seq (y-axes) libraries. Correlation coefficient r calculated using Spearman's Correlation. *P*-values derived from Student's t-distribution. **(G)** UMAP of progenitor cells from MDS01-03 samples overlaid with pseudotemporal ordering.

Figure S3

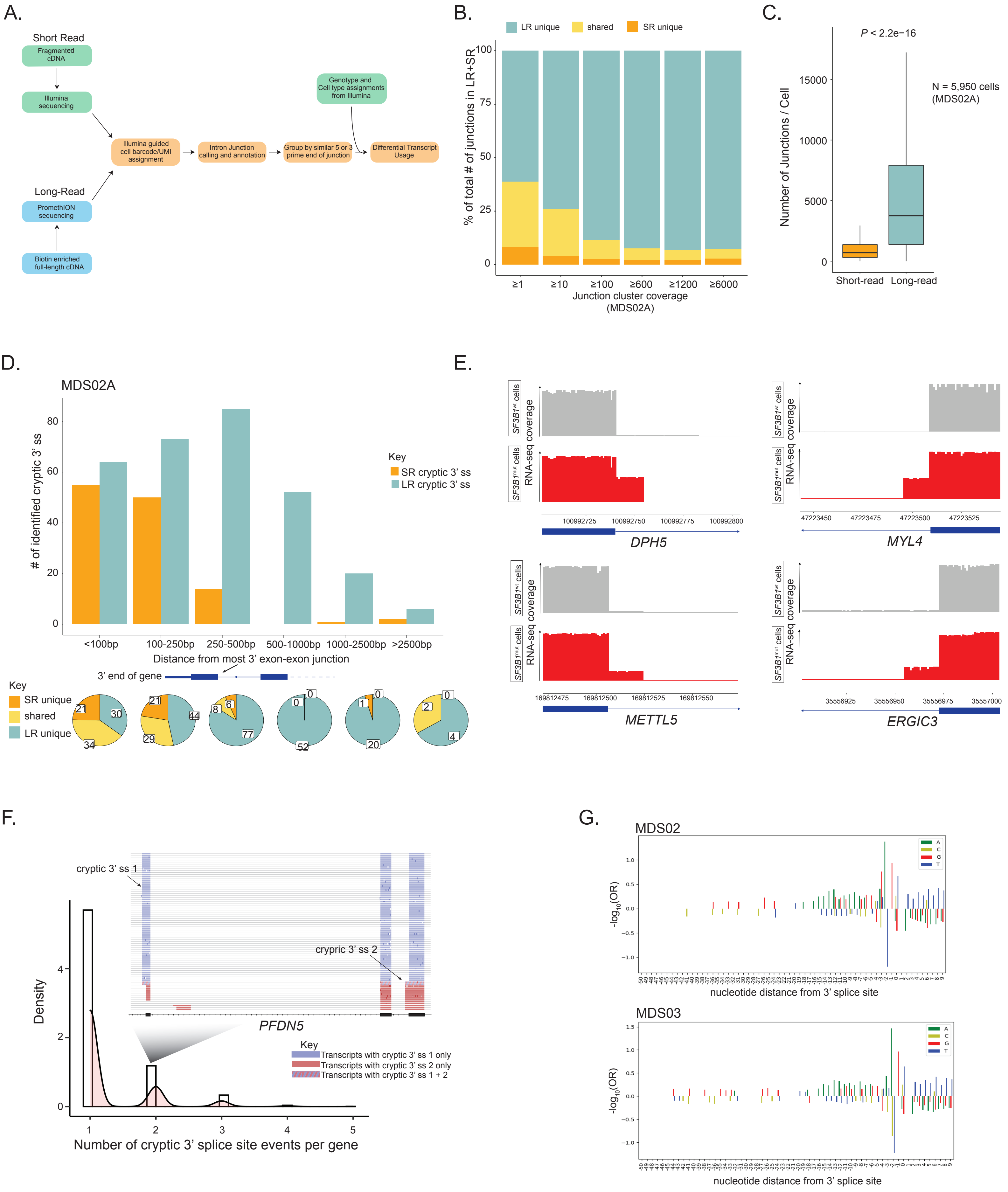
A.
Short Read
Fragmented cDNA
Illumina sequencing
Long-Read
PromethION sequencing
Biotin enriched full-length cDNA
Illumina guided cell barcode/UMI assignment
Intron Junction calling and annotation
Group by similar 5 or 3 prime end of junction
Genotype and Cell type assignments from Illumina
Differential Transcript Usage
B.
LR unique
shared
SR unique
% of total # of junctions in LR+SR
100
75
50
25
0
≥1
≥10
≥100
≥600
≥1200
≥6000
Junction cluster coverage (MDS02A)
C.
P < 2.2e−16
N = 5,950 cells (MDS02A)
Number of Junctions / Cell
15000
10000
5000
0
Short-read
Long-read
D.
MDS02A
# of identified cryptic 3' ss
80
60
40
20
0
Key
SR cryptic 3' ss
LR cryptic 3' ss
<100bp
100-250bp
250-500bp
500-1000bp
1000-2500bp
>2500bp
Distance from most 3' exon-exon junction
3' end of gene
Key
SR unique
shared
LR unique
21
30
34
21
44
29
8
6
77
0
0
52
1
0
20
2
0
4
E.
SF3B1wt cells
SF3B1mut cells
RNA-seq coverage
100992725
100992750
100992775
100992800
DPH5
47223450
47223475
47223500
47223525
MYL4
169812475
169812500
169812525
169812550
METTL5
35556925
35556950
35556975
35557000
ERGIC3
F.
cryptic 3' ss 1
crypric 3' ss 2
PFDN5
Key
Transcripts with cryptic 3' ss 1 only
Transcripts with cryptic 3' ss 2 only
Transcripts with cryptic 3' ss 1 + 2
Density
4
2
0
1
2
3
4
5
Number of cryptic 3' splice site events per gene
G.
MDS02
A
C
G
T
-log10(OR)
1.0
0.5
0.0
−0.5
−1.0
nucleotide distance from 3' splice site
MDS03
1.5
1.0
0.5
0.0
−0.5
−1.0
nucleotide distance from 3' splice site

Figure S3. Long-read splicing and motif analysis, related to Figure 3.

(A) Long-read sequencing processing and splicing analysis pipeline. **(B)** Fraction of cryptic junctions detected in short-reads, long-reads or both when defining distinct cluster coverage thresholds. **(C)** Number of junctions per cell detected in 10x short-read compared to ONT full-length long-read (junction cluster coverage >= 1 read). **(D)** Number of detected cryptic 3’ splice site junctions (in genes with ~1 read/cell in both short and long-read datasets and in clusters with coverage >= 600 reads) along increasing distances to the most 3’ end junction of each transcript (see *inset*) in 10x short versus ONT full-length long-reads. Individual number of events as well as overlaps are itemized in the pie charts (bottom). **(E)** Comparison of the usage of various alternative 3’ splice sites found in our MDS *SF3B1*^{mut} cells vs. a CD34+ sample with no *SF3B1* mutation. **(F)** Bar plot of the number of cryptic 3’ splice sites identified per gene in MDS. *Inset*: Gene example, *PFDN5*, with 2 unique cryptic 3’ splice sites, showing the transcripts that have usage of either site. **(G)** Nucleotide enrichment (measured as log-odds ratio) across the 3’ splice site region comparing cryptic vs. canonical sites in MDS02-03 samples.

Figure S4

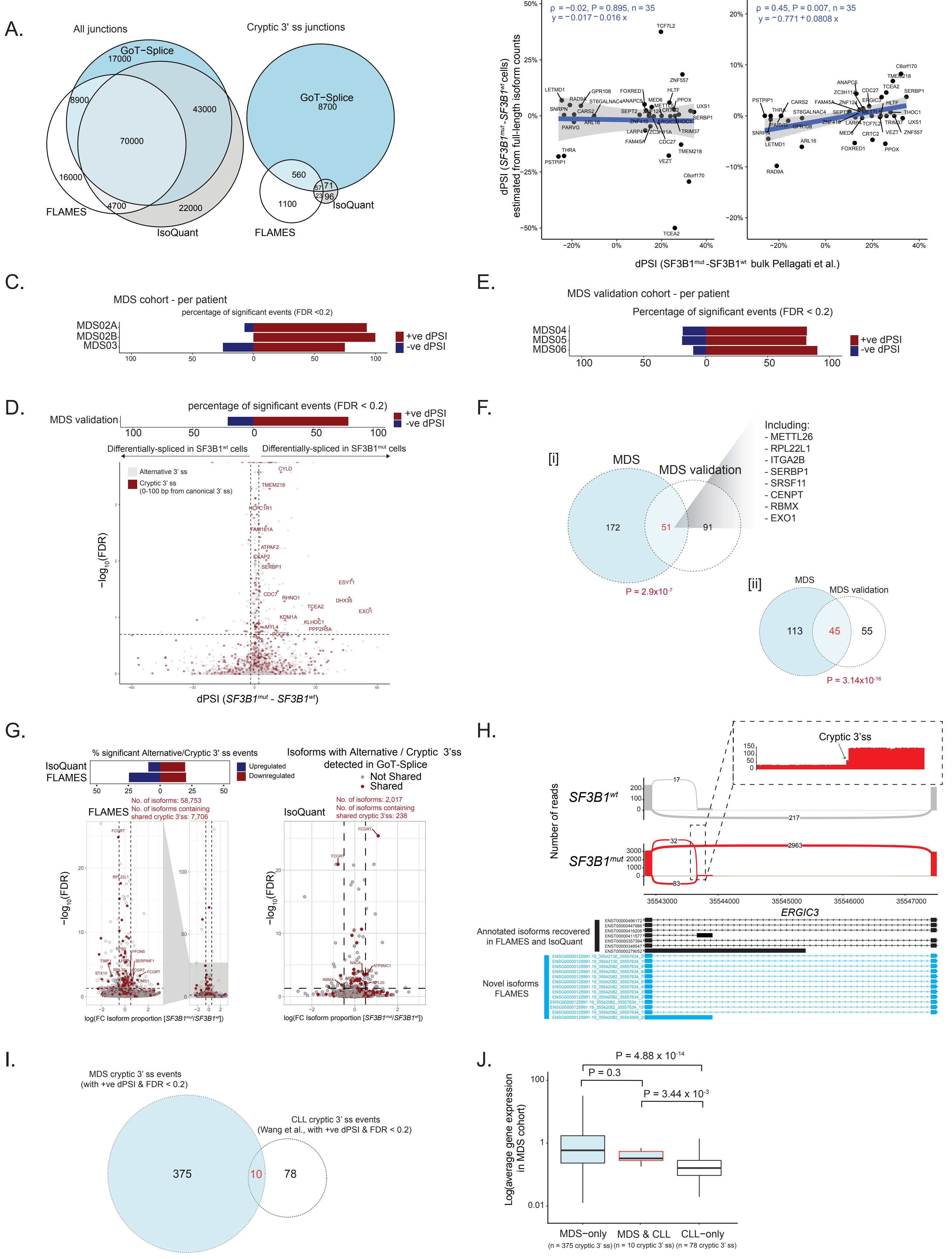
A.
All junctions
GoT−Splice
17000
8900
43000
70000
16000
4700
22000
FLAMES
IsoQuant
Cryptic 3' ss junctions
GoT−Splice
8700
560
57
71
23
96
1100
IsoQuant
FLAMES
B.
IsoQuant
ρ = −0.02, P = 0.895, n = 35
y = −0.017 − 0.016 x
FLAMES
ρ = 0.45, P = 0.007, n = 35
y = −0.771 + 0.0808 x
dPSI (SF3B1mut -SF3B1wt cells) estimated from full-length isoform counts
dPSI (SF3B1mut-SF3B1wt bulk Pellagati et al.)
50%
25%
0%
−25%
−50%
20%
10%
−10%
−20%
−20%
0%
20%
40%
TCF7L2
ZNF557
LETMD1
RAD9A
GPR108
FOXRED1
HLTF
PPOX
UXS1
SERBP1
SNRPN
CARS2
ST6GALNAC4
ANAPC5
MED6
METTL26
SEPT2
ZNF124
CRTC2
ARL16
PARVG
ZNF410
ERGIC3
THOC1
LARP4
ZC3H11A
TRIM37
FAM45A
CDC27
TMEM218
THRA
PSTPIP1
VEZT
C6orf170
TCEA2
C6orf170
TMEM218
TCEA2
SERBP1
ANAPC5
ZC3H11A
CDC27
ERGIC3
HLTF
PSTPIP1
CARS2
FAM45A
ZNF124
THRA
ST6GALNAC4
SEPT2
METTL26
THOC1
SNRPN
PARVG
GPR108
ZNF410
LARP4
TCF7L2
TRIM37
UXS1
LETMD1
ARL16
MED6
CRTC2
VEZT
ZNF557
FOXRED1
PPOX
RAD9A
C.
MDS cohort - per patient
percentage of significant events (FDR <0.2)
MDS02A
MDS02B
MDS03
100
50
0
50
100
+ve dPSI
-ve dPSI
D.
percentage of significant events (FDR < 0.2)
MDS validation
100
50
0
50
100
+ve dPSI
-ve dPSI
Differentially-spliced in SF3B1wt cells
Differentially-spliced in SF3B1mut cells
Alternative 3' ss
Cryptic 3' ss
(0-100 bp from canonical 3' ss)
−log10(FDR)
dPSI (SF3B1mut - SF3B1wt)
CYLD
TMEM216
HCFC1R1
FAM161A
ATPAF2
CKAP2
SERBP1
ESYT1
CDC7
RHNO1
DHX35
TCEA2
EXO1
KDM1A
KLHDC1
MYL4
PPP2R5A
PCGF6
E.
MDS validation cohort - per patient
Percentage of significant events (FDR < 0.2)
MDS04
MDS05
MDS06
100
50
0
50
100
+ve dPSI
-ve dPSI
F.
[i]
MDS
MDS validation
172
51
91
P = 2.9x10-7
Including:
- METTL26
- RPL22L1
- ITGA2B
- SERBP1
- SRSF11
- CENPT
- RBMX
- EXO1
[ii]
MDS
MDS validation
113
45
55
P = 3.14x10-16
G.
% significant Alternative/Cryptic 3' ss events
IsoQuant
FLAMES
50
25
0
25
50
Upregulated
Downregulated
Isoforms with Alternative / Cryptic 3'ss detected in GoT-Splice
Not Shared
Shared
FLAMES
No. of isoforms: 58,753
No. of isoforms containing shared cryptic 3'ss: 7,706
IsoQuant
No. of isoforms: 2,017
No. of isoforms containing shared cryptic 3'ss: 238
−log10(FDR)
log(FC Isoform proportion [SF3B1mut/SF3B1wt])
FCGRT
RPL22L1
PFDN5
SERPINF1
TIMP1
FCGRT
STX10
RPL30
ATP5MC1
NACA
RBMX
H.
Cryptic 3'ss
150
100
50
0
Number of reads
SF3B1wt
200
100
0
17
217
SF3B1mut
3000
2000
1000
0
32
2963
83
35543000
35544000
35545000
35546000
35547000
ERGIC3
Annotated isoforms recovered in FLAMES and IsoQuant
ENST00000496172
ENST00000447986
ENST00000416206
ENST00000411577
ENST00000357394
ENST00000348547
ENST00000279052
Novel isoforms FLAMES
I.
MDS cryptic 3' ss events
(with +ve dPSI & FDR < 0.2)
CLL cryptic 3' ss events
(Wang et al., with +ve dPSI & FDR < 0.2)
375
10
78
J.
P = 4.88 x 10-14
P = 0.3
P = 3.44 x 10-3
Log(average gene expression in MDS cohort)
100
1
0.01
MDS−only
MDS & CLL
CLL−only
(n = 375 cryptic 3' ss)
(n = 10 cryptic 3' ss)
(n = 78 cryptic 3' ss)

Figure S4. Comparison of splicing junctions detected with GoT-Splice vs. other tools, MDS samples, and cancer types, related to Figure 4.

(A) *Left*: Overlapping of all splicing junctions detected in GoT-Splice, FLAMES, and IsoQuant. *Right*: Comparison of splicing junctions supporting cryptic 3' ss between the two full-length isoform detection tools and GoT-Splice. **(B)** Estimated dPSI values from long-read tools FLAMES and IsoQuant compared to reported dPSI values measured in bulk data, like **Figure 2g**. **(C)** Bars showing the percentage of genes differentially spliced in *SF3B1^{mut}* and *SF3B1^{wt}* cells with BH-FDR adjusted *P*-value < 0.2 in each sample of the MDS cohort (MDS02(A/B)-03)**. (D)** Differential splicing analysis between *SF3B1^{mut}* and *SF3B1^{wt}* cells across the aggregate of the MDS validation cohort (MDS04-06). Junctions with an absolute delta percent spliced-in (dPSI) > 2 and BH-FDR adjusted *P*-value < 0.2 were defined as differentially spliced. Bars (top) showing the percentage of genes differentially spliced in *SF3B1^{mut}* and *SF3B1^{wt}* cells of MDS validation cohort. **(E)** Bars showing the percentage of genes differentially spliced in *SF3B1^{mut}* and *SF3B1^{wt}* cells with BH-FDR adjusted *P*-value < 0.2 in each sample of the MDS validation cohort (MDS04-06)**. (F)** Venn Diagram for the overlap of differentially spliced genes used more highly in *SF3B1^{mut}* cells (*P*-values < 0.05, dPSI >0) from the bulk comparison of *SF3B1^{mut}* vs. *SF3B1^{wt}* cells in the MDS and MDS validation cohorts [i]. Increasing the read coverage threshold for the differentially spliced genes showed a more significant overlap between cohorts [ii]. *P*-values for the overlap from Fisher's exact test. **(G)** Differential isoform proportion in *SF3B1*$^{\text{mut}}$ vs *SF3B1*$^{\text{wt}}$ cells. Isoforms containing alternative/cryptic 3' ss overlapping with those detected in GoT-Splice are highlighted in red. *Top*: Quantification of significant (FDR adjusted *P*-value < 0.05) full-length isoforms detected containing alternative/cryptic 3'ss, split by direction of the logFC in proportion between MUT and WT cells. **(H)** Example of cryptic 3' ss in the gene *ERGIC3* missed by IsoQuant and FLAMES but detected and quantified by GoT-Splice. **(I)** Venn Diagram showing the overlap of genes with cryptic 3' ss splicing events (dPSI > 0 and FDR adjusted *P*-value < 0.2) found in the MDS discovery cohort vs. those found in the bulk RNA sequencing of a set of published *SF3B1^{mut}* CLL samples. **(J)** Box plot of the average expression levels found across all cells in the MDS discovery cohort for the genes cryptically mis-spliced in the MDS dataset only, in both the MDS and CLL datasets, and the CLL dataset only. *P*-values were computed using Wilcox rank sum test.

Figure S5

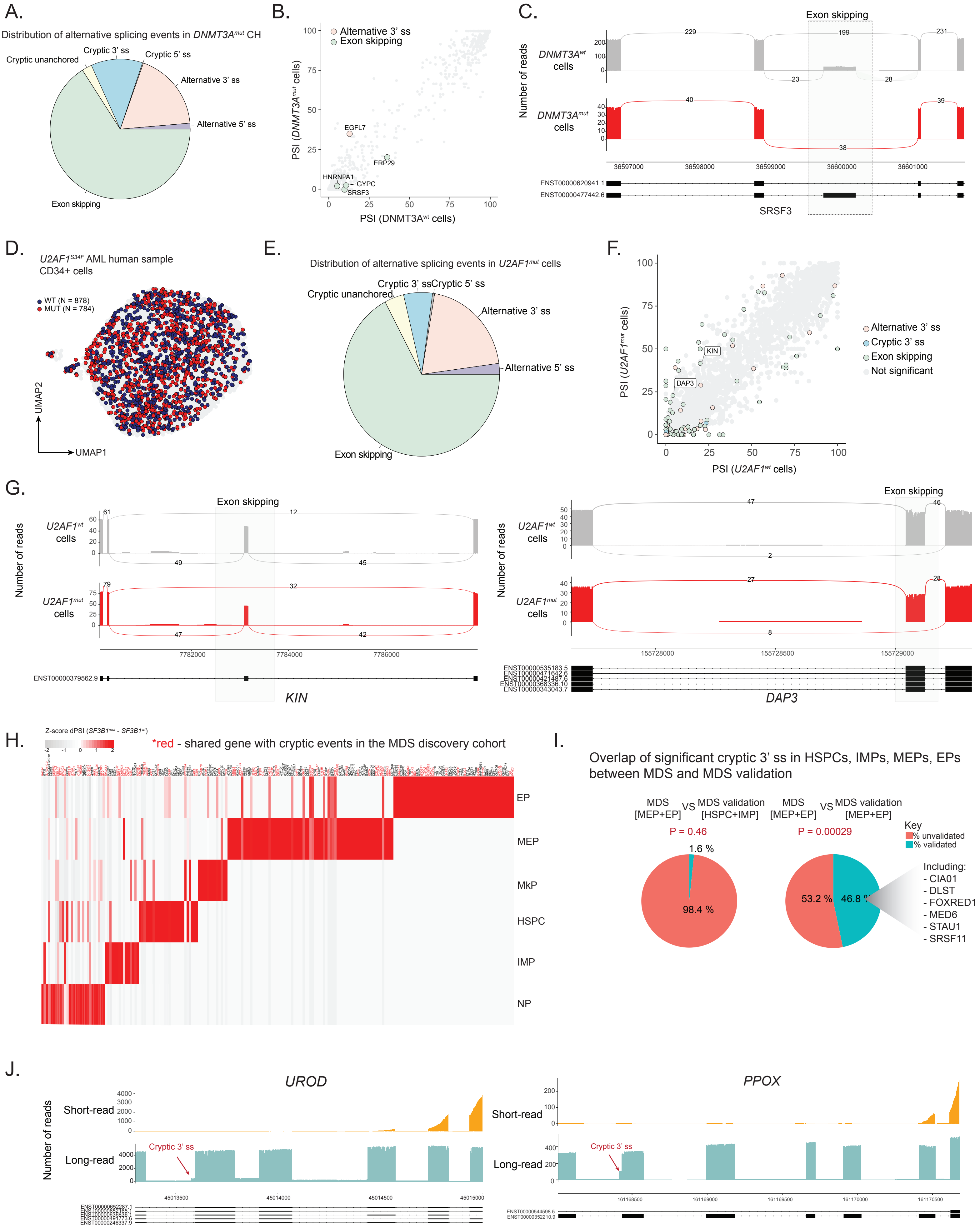
A.
Distribution of alternative splicing events in DNMT3A^mut CH
Cryptic unanchored
Cryptic 3' ss
Cryptic 5' ss
Alternative 3' ss
Alternative 5' ss
Exon skipping
B.
Alternative 3' ss
Exon skipping
EGFL7
ERP29
HNRNPA1
GYPC
SRSF3
PSI (DNMT3A^mut cells)
PSI (DNMT3A^wt cells)
100
75
50
25
0
C.
Exon skipping
Number of reads
DNMT3A^wt cells
DNMT3A^mut cells
229
199
231
23
28
40
39
38
36597000
36598000
36599000
36600000
36601000
ENST00000620941.1
ENST00000477442.6
SRSF3
D.
U2AF1^S34F AML human sample
CD34+ cells
WT (N = 878)
MUT (N = 784)
UMAP1
UMAP2
E.
Distribution of alternative splicing events in U2AF1^mut cells
Cryptic unanchored
Cryptic 3' ss
Cryptic 5' ss
Alternative 3' ss
Alternative 5' ss
Exon skipping
F.
Alternative 3' ss
Cryptic 3' ss
Exon skipping
Not significant
KIN
DAP3
PSI (U2AF1^mut cells)
PSI (U2AF1^wt cells)
G.
Exon skipping
Number of reads
U2AF1^wt cells
U2AF1^mut cells
61
12
49
45
79
32
47
42
7782000
7784000
7786000
ENST00000379562.9
KIN
47
46
2
27
28
8
155728000
155728500
155729000
ENST00000535183.5
ENST00000471642.6
ENST00000421487.6
ENST00000368336.10
ENST00000343043.7
DAP3
H.
Z-score dPSI (SF3B1^mut - SF3B1^wt)
-2 -1 0 1 2
*red - shared gene with cryptic events in the MDS discovery cohort
EP
MEP
MkP
HSPC
IMP
NP
I.
Overlap of significant cryptic 3' ss in HSPCs, IMPs, MEPs, EPs between MDS and MDS validation
MDS [MEP+EP] VS MDS validation [HSPC+IMP]
MDS [MEP+EP] VS MDS validation [MEP+EP]
P = 0.46
P = 0.00029
1.6 %
98.4 %
53.2 %
46.8 %
Key
% unvalidated
% validated
Including:
- CIA01
- DLST
- FOXRED1
- MED6
- STAU1
- SRSF11
J.
UROD
PPOX
Number of reads
Short-read
Long-read
Cryptic 3' ss
45013500
45014000
45014500
45015000
ENST00000652287.1
ENST00000246337.9
161168500
161169000
161169500
161170000
161170500
ENST00000544598.5
ENST00000352210.9

Figure S5. Application of GoT-Splice to other mutations and identification of cell-type specific splicing changes, related to Figure 5.

(A) Pie chart summarizing the distribution of different alternative splicing events detected after junction annotation in the *DNMT3A^{mut}* CH sample. **(B)** Comparison of percent spliced-in (dPSI) values of alternative splicing events identified in *DNMT3A^{mut}* vs. *DNMT3A^{wt}* CH cells. Significant alternative splicing events are highlighted. **(C)** Sashimi plot of *SRSF3* exon skipping event showing the expected increase in the PSI value in *DNMT3A^{wt}* cells. (**D)** UMAP of progenitor cells from *U2AF1^{S34F}* AML sample overlaid with genotyping data. WT, cells with genotype data without *U2AF1* mutation; MUT, cells with genotype data with *U2AF1* mutation. **(E)** Pie chart summarizing the distribution of different alternative splicing events detected after junction annotation in the *U2AF1^{mut}* AML sample. **(F)** Comparison of percent spliced-in (dPSI) values of alternative splicing events identified in *U2AF1^{mut}* vs. *U2AF1^{wt}* cells. Significant alternative splicing events are highlighted. **(G)** Sashimi plot of *KIN* and *DAP3* exon skipping events showing the expected increase in the exon skipping in *U2AF1^{mut}* cells. **(H)** Heatmap of dPSI values between *SF3B1^{mut}* and *SF3B1^{wt}* cells for cryptic 3' splicing events identified in the main progenitor subsets across MDS validation samples. Rows (z-score normalized) correspond to cryptic 3' junctions found to be differentially spliced in at least one cell-type, with *P*-value <= 0.05 and dPSI >= 2. Columns correspond to cell-type. Genes with an *SF3B1^{mut}* associated cryptic 3' splice site found in the MDS cohort are highlighted in red. **(I)** Pie chart showing the percent overlap of cryptically 3' spliced genes unique to MEPs and EPs in the primary MDS cohort that are also cryptically 3' spliced and unique to earlier progenitor cells (HSPCs and IMPs) in the MDS validation cohort (left) as well as the percent overlap with genes cryptically 3' spliced and unique to the MEPs and EPs in the MDS validation cohort (right). *P*-value for the overlap from Fisher's exact test. **(J)** Plots highlighting the differences in sequencing coverage along the transcript, between short- and long-read sequencing, for *PPOX* and *UROD*, demonstrating the ability for long-read sequencing to uncover splicing events than remain undetected in the 3' biased short-read sequencing.

Figure S6

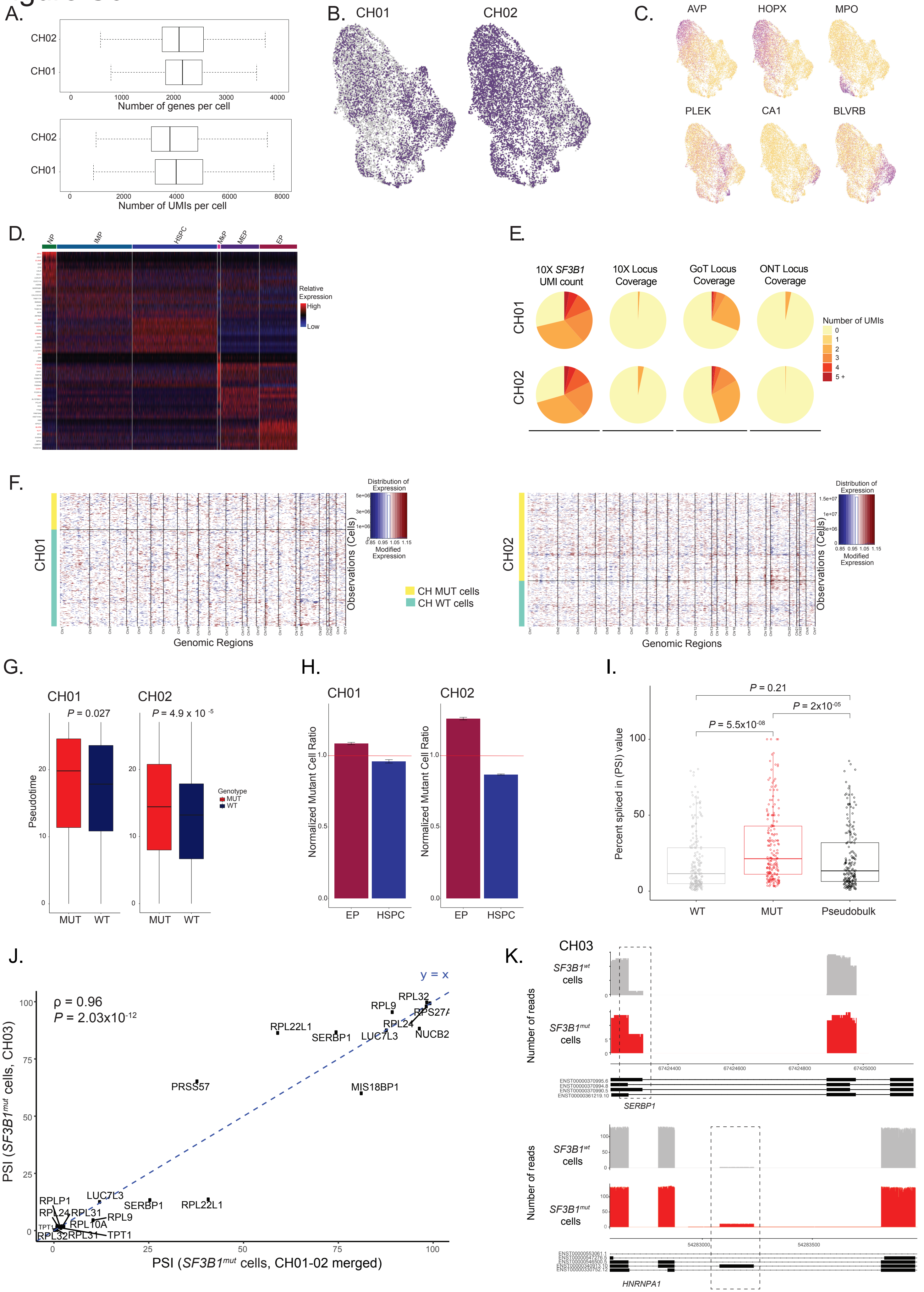
A.
CH02
CH01
0
1000
2000
3000
4000
Number of genes per cell
CH02
CH01
0
2000
4000
6000
8000
Number of UMIs per cell
B.
CH01
CH02
C.
AVP
HOPX
MPO
PLEK
CA1
BLVRB
D.
NP
IMP
HSPC
MkP
MEP
EP
Relative Expression
High
Low
E.
10X SF3B1 UMI count
10X Locus Coverage
GoT Locus Coverage
ONT Locus Coverage
CH01
CH02
Number of UMIs
0
1
2
3
4
5 +
F.
CH01
Distribution of Expression
5e+06
3e+06
1e+06
0.85 0.95 1.05 1.15
Modified Expression
Observations (Cells)
Genomic Regions
CH MUT cells
CH WT cells
CH02
Distribution of Expression
1.5e+07
1e+07
5e+06
0
0.85 0.95 1.05 1.15
Modified Expression
Observations (Cells)
Genomic Regions
G.
CH01
CH02
P = 0.027
P = 4.9 x 10 -5
Pseudotime
0
10
20
Genotype
MUT
WT
MUT
WT
MUT
WT
H.
CH01
CH02
Normalized Mutant Cell Ratio
0.0
0.5
1.0
EP
HSPC
EP
HSPC
I.
P = 0.21
P = 2x10-05
P = 5.5x10-08
Percent spliced in (PSI) value
0
50
100
WT
MUT
Pseudobulk
J.
y = x
ρ = 0.96
P = 2.03x10-12
PSI (SF3B1mut cells, CH03)
PSI (SF3B1mut cells, CH01-02 merged)
0
25
50
75
100
RPL32
RPL9
RPS27A
RPL24
NUCB2
RPL22L1
SERBP1
LUC7L3
PRSS57
MIS18BP1
LUC7L3
RPLP1
SERBP1
RPL22L1
RPL24
RPL31
RPL9
TPT1
RPL10A
RPL32
RPL31
TPT1
K.
CH03
SF3B1wt cells
SF3B1mut cells
Number of reads
67424400
67424600
67424800
67425000
ENST00000370995.6
ENST00000370994.8
ENST00000370990.5
ENST00000361219.10
SERBP1
SF3B1wt cells
SF3B1mut cells
Number of reads
54283000
54283500
ENST00000553061.1
ENST00000547276.5
ENST00000546500.5
ENST00000340913.10
ENST00000330752.12
HNRNPA1

Figure S6. CH cohort QC, integration, and GoT-Splice analyses, related to Figure 6.

(A) Number of genes per cell (top) and number of UMIs per cell (bottom) in CD34+ sorted hematopoietic progenitors from samples CH01-02 after QC filters, shown by each patient sample. **(B)** UMAP of CD34+ sorted progenitor cells for each individual sample of CH01-02 after integration using the Seurat package. **(C)** Expression of lineage-specific genes from Velten *et al.*[S1] scored and projected onto the UMAP representation of cells from CH01-02. **(D)** Heatmap of top 10 differentially expressed genes for each progenitor subset for CH01-02. **(E)** Fraction of cells in CH01-02 by number of *SF3B1* UMIs in standard 10x Genomics data without genotyping information, *SF3B1* UMIs with K666N (CH01) or K700E (CH02) locus coverage in standard 10x data, *SF3B1* UMIs with K666N (CH01) or K700E (CH02) locus coverage in GoT amplicon library, and *SF3B1* UMIs with K666N (CH01) or K700E (CH02) locus coverage in ONT library**.** **(F)** Per sample heatmap of relative expression of genes ordered by chromosome/chromosomal position following copy number variation analysis using the InferCNV package (see STAR Methods). Cells (y-axis) are stratified by *SF3B1* genotype status. **(G)** Pseudotime in $SF3B1^{mut}$ vs. $SF3B1^{wt}$ cells per CH sample. *P*-value for comparison of means from Wilcoxon rank sum test. **(H)** Normalized ratio of $SF3B1^{mut}$ cells in HSPC and EP cells for CH01 and CH02. Bars show the mean of *n* = 100 downsampling iterations to 1 genotyping UMI per cell. **(I)** Comparison of the PSI values of identified cryptic junctions in WT cells only (gray) vs. MUT cells only (red) vs. all cells in pseudobulk in CH01-02. *P*-values for comparison of means from Wilcoxon rank sum test. **(J)** Comparison of the $SF3B1^{mut}$ cells PSI values for alternative splicing events that were found to be significantly differentially used (*P*-value < 0.05) between $SF3B1^{mut}$ and $SF3B1^{wt}$ in the primary CH cohort and that were also present in the CH validation sample. Spearman's rank correlation coefficient and *P*-value derived from Student's t-distribution are shown. **(K)** Sashimi plot examples of alternative splicing events in *SERBP1* (top, cryptic 3' ss event) and *HNRNPA1* (bottom, exon skipping/inclusion event) found to be significantly differentially spliced between $SF3B1^{mut}$ and $SF3B1^{wt}$ cells in the CH validation sample.

REFERENCES

[S1] 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. 10.1038/ncb3493.