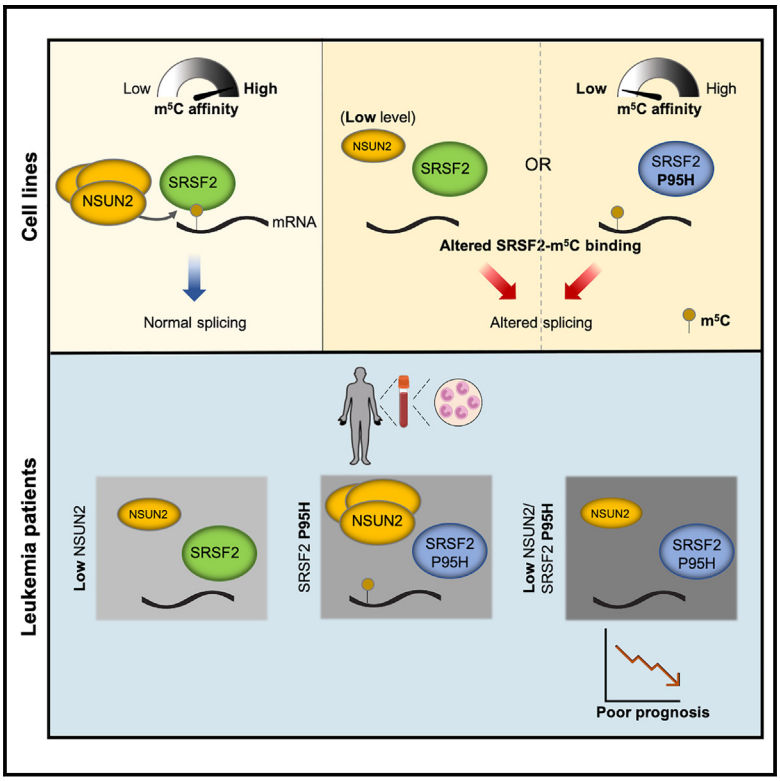
Molecular Cell

Article

SRSF2 plays an unexpected role as reader of m^{5}C on mRNA, linking epitranscriptomics to cancer

Graphical abstract



Highlights

- SRSF2 preferentially binds m^{5}C-marked mRNA, whereas SRSF2^{P95H} mutant impairs binding
- NSUN2 depletion reduces mRNA m^{5}C levels and alters SRSF2 RNA binding and splicing
- NSUN2 loss and SRSF2^{P95H} alter SRSF2 binding to key leukemia-related transcripts
- In leukemia patients, low *NSUN2* levels and SRSF2^{P95H} mutation predict poor outcomes

Authors

Hai-Li Ma, Martin Bizet, Christelle Soares Da Costa, ..., Omar Abdel-Wahab, Yun-Gui Yang, François Fuks

Correspondence

ygyang@big.ac.cn (Y.-G.Y.), francois.fuks@ulb.be (F.F.)

In brief

Ma et al. report that the RNA-splicing factor SRSF2 is an mRNA m^{5}C reader, that a frequent leukemia-associated mutation impairs SRSF2-m^{5}C binding, and that this is associated with leukemogenesis. This work uncovers a mechanistic link between epitranscriptomics and a key driver of oncogenesis.

Ma et al., 2023, Molecular Cell *83*, 4239–4254
December 7, 2023
https://doi.org/10.1016/j.molcel.2023.11.003

CellPress

Molecular Cell

CellPress

Article

SRSF2 plays an unexpected role as reader of m^{5}C on mRNA, linking epitranscriptomics to cancer

Hai-Li Ma,[1,14] Martin Bizet,[1,14] Christelle Soares Da Costa,[1,14] Frédéric Murisier,[1] Eric James de Bony,[1,8] Meng-Ke Wang,[2] Akihide Yoshimi,[3,9] Kuan-Ting Lin,[7] Kristin M. Riching,[4] Xing Wang,[2] John I. Beckman,[5] Shailee Arya,[5] Nathalie Droin,[6] Emilie Calonne,[1] Bouchra Hassabi,[1] Qing-Yang Zhang,[2,10] Ang Li,[2,11] Pascale Putmans,[1] Lionel Malbec,[1] Céline Hubert,[1] Jie Lan,[1,12] Frédérique Mies,[1] Ying Yang,[2] Eric Solary,[6] Danette L. Daniels,[4,13] Yogesh K. Gupta,[5] Rachel Deplus,[1] Omar Abdel-Wahab,[3] Yun-Gui Yang,[2,*] and François Fuks[1,15,*]

[1]Laboratory of Cancer Epigenetics, Faculty of Medicine, ULB-Cancer Research Center (U-CRC), Université libre de Bruxelles (ULB), Institut Jules Bordet, Brussels 1070, Belgium

[2]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics and Chinese Academy of Sciences, China National Center for Bioinformation, Beijing 100101, China

[3]Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[4]Promega Corporation, Madison, WI 53711, USA

[5]Greehey Children's Cancer Research Institute, Department of Biochemistry and Structural Biology, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

[6]Université Paris-Saclay, INSERM U1287, and Department of Hematology, Gustave Roussy Cancer Center, Villejuif 94800, France

[7]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

[8]Present address: OncoRNALab, Cancer Research Institute Ghent (CRIG), Department of Biomolecular Medicine, Ghent University, Ghent 9000, Belgium

[9]Present address: Division of Cancer RNA Research, National Cancer Center Research Institute, Tokyo 1040045, Japan

[10]Present address: State Key Laboratory of Liver Research, Department of Pathology, The University of Hong Kong, Hong Kong 999077, China

[11]Present address: Shandong Provincial Key Laboratory of Animal Cell and Developmental Biology, School of Life Sciences, Shandong University, Qingdao 266237, China

[12]Present address: Institute for Genetics, Justus Liebig University, Giessen 35392, Germany

[13]Present address: Foghorn Therapeutics, 500 Technology Square, Cambridge, MA 02139, USA

[14]These authors contributed equally

[15]Lead contact

*Correspondence: ygyang@big.ac.cn (Y.-G.Y.), francois.fuks@ulb.be (F.F.)

https://doi.org/10.1016/j.molcel.2023.11.003

SUMMARY

A common mRNA modification is 5-methylcytosine (m^{5}C), whose role in gene-transcript processing and cancer remains unclear. Here, we identify serine/arginine-rich splicing factor 2 (SRSF2) as a reader of m^{5}C and impaired SRSF2 m^{5}C binding as a potential contributor to leukemogenesis. Structurally, we identify residues involved in m^{5}C recognition and the impact of the prevalent leukemia-associated mutation SRSF2^{P95H}. We show that SRSF2 binding and m^{5}C colocalize within transcripts. Furthermore, knocking down the m^{5}C writer NSUN2 decreases mRNA m^{5}C, reduces SRSF2 binding, and alters RNA splicing. We also show that the SRSF2^{P95H} mutation impairs the ability of the protein to read m^{5}C-marked mRNA, notably reducing its binding to key leukemia-related transcripts in leukemic cells. In leukemia patients, low *NSUN2* expression leads to mRNA m^{5}C hypomethylation and, combined with SRSF2^{P95H}, predicts poor outcomes. Altogether, we highlight an unrecognized mechanistic link between epitranscriptomics and a key oncogenesis driver.

INTRODUCTION

RNA modifications are important in the regulation of eukaryotic cells.[1] Of the 170 different RNA modifications known to date, approximately 80% are methylations. N6-methyladenosine (m^{6}A) is the most abundant modification on higher-eukaryote mRNAs, with substantial links to human pathologies.[2,3] Another modification, 5-methylcytosine (m^{5}C), has also been found on a wide range of RNAs, such as tRNA, rRNA, non-coding RNA (ncRNA), and mRNA.[4] The presence of the m^{5}C modification on mRNA has attracted increasing attention, and several m^{5}C regulators have been identified.[5] The m^{5}C methyltransferases (writers), NOP2/Sun RNA methyltransferase (NSUN)2 and NSUN6, and the demethylases (erasers), ten–eleven

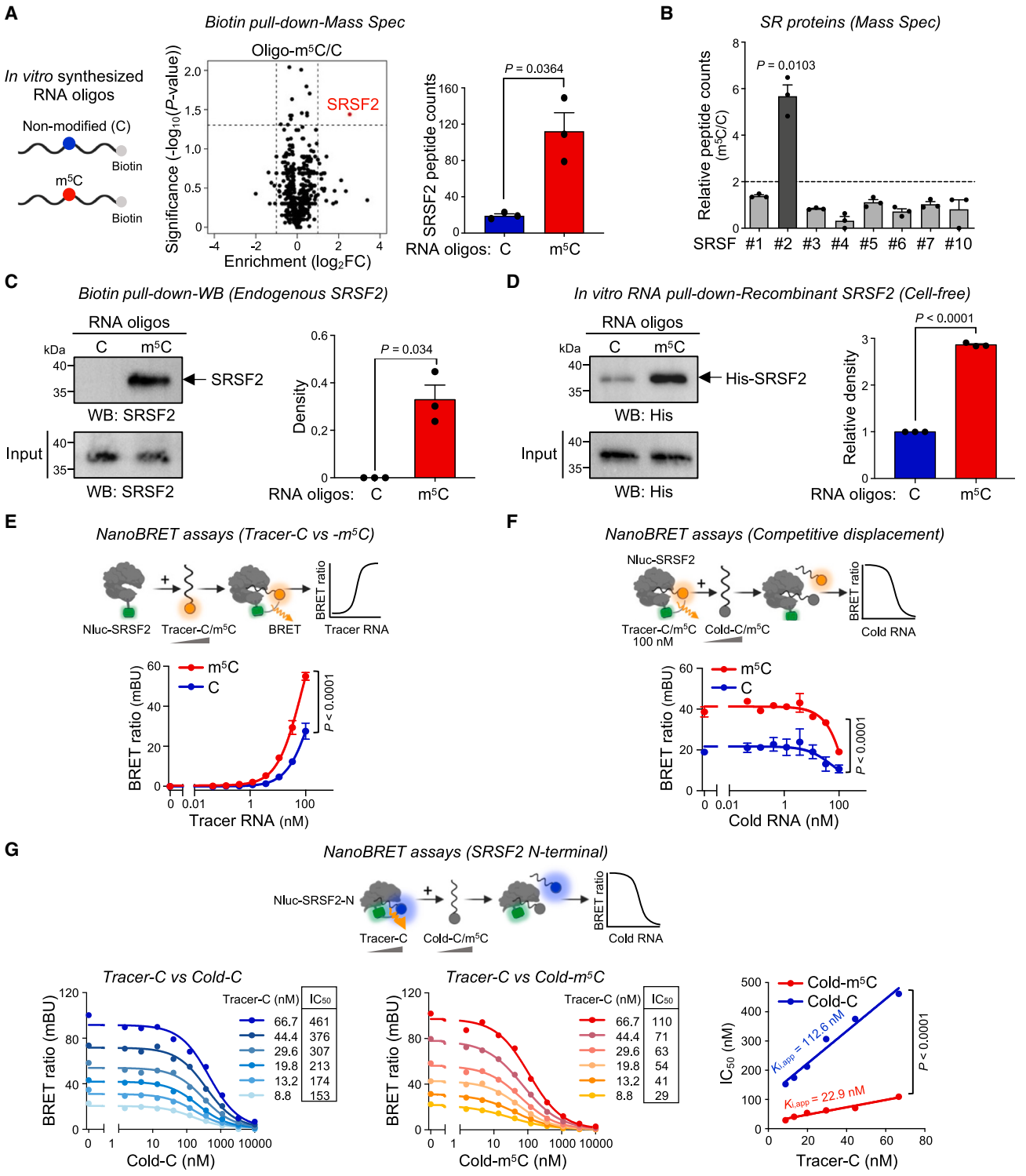


Figure 1. SRSF2 binds preferentially to m^{5}C-modified RNAs

(A) SRSF2 binds to m^{5}C-RNA with higher affinity than to the unmodified control (n = 3).

(B) Among the SR-family proteins, only SRSF2 preferentially binds m^{5}C-modified RNA (n = 3).

(C) Biotin pull-down followed by western blotting shows that endogenous SRSF2 binds to oligo-m^{5}C with higher affinity than to oligo-C (n = 3).

(D) *In vitro* RNA pull-down with recombinant His-tagged SRSF2 demonstrates the direct binding of SRSF2 to m^{5}C (n = 3).

(E) NanoBRET assays in cells transiently transfected with Nluc-SRSF2 protein and treated with varying concentrations of RNA tracer-m^{5}C or tracer-C (n = 3).

(legend continued on next page)

translocation family member 2 (TET2) and alkylated DNA repair protein alkB homolog 1 (ALKBH1), are regulators of m^5C levels.[6–10] To date, the proteins known to bind m^5C-marked RNA transcripts (readers) are Aly/REF export factor (ALYREF), Y-box binding protein (YBX)1, YBX2, YTH domain-containing family protein 2 (YTHDF2), radiation sensitive 52 (RAD52), and fragile X mental retardation protein (FMRP).[9,11–16] mRNA m^5C modifications have been implicated in various biological processes and multiple diseases through reader proteins. For example, YBX1 recognizes and maintains the stability of its target m^5C-marked mRNAs, thereby mediating oncogene activation in the pathogenesis of human bladder urothelial carcinoma.[15] Whether other cell proteins likewise recognize and bind m^5C is unknown. The discovery of m^5C reader proteins will help elucidate the mechanisms affecting the fate and functions of m^5C-modified RNAs.

Serine/arginine-rich (SR) proteins are RNA-binding proteins acting as core regulators of RNA splicing. The family comprises 12 unique members, SR splicing factor (SRSF)1–12,[17] including SRSF2. As a splicing factor, SRSF2 binds exonic splicing enhancer (ESE) motifs and facilitates both constitutive and alternative splicing.[18–20] SRSF2 is essential to the functional integrity of the hematopoietic system, and its mutations can alter the RNA-splicing profiles of a wide panel of genes involved in carcinogenesis.[21] SRSF2 mutations occur in ~15% of the patients with acute myeloid leukemia (AML), 20%–30% of the patients with myelodysplastic syndrome (MDS), and 47% of those with chronic myelomonocytic leukemia (CMML).[22–24] SRSF2 heterozygous mutations occur frequently at position 95, with the most common mutation being proline-to-histidine (P95H).[24] Although the motif for SRSF2 is SSNG (S = C/G, N = A/C/G/U), $SRSF2^{P95H}$ shows a higher binding affinity for CCNG than the GGNG motif, which alters the RNA-binding activity to specific ESE motifs.[25,26] Despite these important findings, the mechanisms underlying the altered binding preference and aberrant splicing conferred by the P95H mutation in leukemia remain elusive.

Here, we unexpectedly find that SRSF2 exhibits preferential direct binding to m^5C-modified RNAs. By mapping the transcriptome-wide SRSF2 RNA-binding profile and m^5C methylome in HeLa cells, we reveal changes in m^5C levels, RNA binding, and splicing upon NSUN2 depletion. Strikingly, the prevalent leukemia-associated $SRSF2^{P95H}$ mutation decreases the affinity of SRSF2 binding to mRNA m^5C. In leukemia cells, this mutation results in reduced binding to many leukemia-related transcripts and leads to alterations in global RNA-splicing patterns, similar to those seen with NSUN2 loss. Moreover, by means of RNA m^5C modification landscape analysis in CMML patients, we find overall decreased m^5C levels in patients with low *NSUN2* levels. We have evidenced an association between low *NSUN2* expression combined with $SRSF2^{P95H}$ and poor prognosis in AML patients. By linking epitranscriptomics to a frequent leukemia-associated mutation, our findings open potential therapeutic avenues for hematologic malignancies.

RESULTS

SRSF2 binds preferentially to m^5C-modified RNAs

Identifying m^5C-binding proteins is an important step toward better understanding the biological consequences of m^5C on RNA. To find m^5C-binding proteins, we performed RNA pull-down assays followed by mass spectrometry. We found that only one protein showed a significant binding preference for biotinylated m^5C RNA oligos: SRSF2 (Figure 1A; Table S1). For other SR proteins detected by mass spectrometry, we found each SR protein to have a binding motif showing at least 50% identity to the RNA probe (Figure S1A). However, although SRSF2 showed a significant increase in binding to m^5C RNA oligos, the other SR proteins showed no significant changes (Figures 1B and S1B). This confirms the reliability of the experiment and the specificity of SRSF2 for m^5C in the C(m^5C)GG context. Furthermore, enrichment of the pull-down mixture in endogenous and overexpressed SRSF2 appeared greater with the m^5C bait than with the control (Figures 1C and S1C). Consistently, recombinant SRSF2 exhibited a strong preference for m^5C probes in a cell-free environment, suggesting direct binding (Figure 1D). Additionally, we evaluated the integrity of the RNA probes before and after pull-down and found no difference in probe stability (Figure S1D). As further controls, complexes pulled down by probes containing m^6A or 5-hydroxymethylcytosine (hm^5C, an oxidation product of m^5C) modifications did not appear enriched in SRSF2, indicating that SRSF2-m^5C interaction is specific (Figure S1E). To examine which domain(s) of SRSF2 mediate its preferential binding to m^5C-decorated RNA, we performed pull-down with SRSF2 fragments. On western blots (Figure S1F), the N-terminal fragment (SRSF2-N) containing the RNA recognition motif (RRM) and linker region exhibited a binding profile similar to that of the full-length protein, suggesting that the N terminus of SRSF2 is essential to m^5C recognition and binding.

To confirm on live cells the above-described *in tubo* SRSF2-m^5C interaction and monitor this interaction quantitatively, we conducted nanoluciferase-based bioluminescence resonance energy transfer (NanoBRET) assays. First, we tested the suitability of m^5C-marked and unmarked RNA tracer probes (called tracer-m^5C and tracer-C) for NanoBRET. At all concentrations, tracer-m^5C gave rise to a stronger BRET signal than tracer-C (Figure 1E). We then used cold (unlabeled) RNA for competitive binding and found that cold RNA attenuated the BRET signal in a concentration-dependent manner, which indicates that the BRET signal was generated by a specific, reversible interaction of the tracer with the Nanoluciferase (Nluc)-fused SRSF2 (Figure 1F).

(F) Concentration-dependent attenuation of BRET from Nluc-SRSF2 upon titration with cold-C or cold-m^5C in the presence of a fixed concentration of the corresponding tracer (n = 2).
(G) The SRSF2 N terminus binds to m^5C with higher affinity than to C. IC_{50}, half-maximal inhibitory concentration. Pooled data in (A)–(F) are represented as mean ± SEM. p values in (A)–(C) and in (D) were calculated using paired or unpaired two-tailed Student's t test, respectively. p values in (E)–(F) and in (G) were determined using extra sum-of-squares F test and two-tailed F test, respectively.
See also Figure S1 and Table S1.

Displacing the tracer RNA with cold RNA made it possible to determine the relative affinities of SRSF2 binding to different RNA sequences. We performed NanoBRET assays using different concentrations of cold-C or cold-m^{5}C RNA for competitive displacement in the presence of different concentrations of tracer-C. Higher affinity binding (apparent dissociation constant, $K_{i,app}$ = 84.39 nM) was observed with cold-m^{5}C than with cold-C ($K_{i,app}$ = 322.1 nM) (Figure S1G). This strengthens our finding that SRSF2 preferentially binds m^{5}C. We also estimated this preferential binding on other different RNA probes and found that SRSF2 had a higher affinity to all Cm^{5}CNG-containing probes but very weak binding to the A(C/m^{5}C)AA-containing probes (Figures S1H–S1K). These results suggest a role for m^{5}C in increasing the binding of SRSF2 to its target RNAs, at least in all the sequence contexts tested. Using Nluc-fused SRSF2 N- and C-terminal fragments to test the binding to our RNA probes, only the former gave rise to a significant BRET signal (Figure S1L). Competition experiments using SRSF2-N showed that cold-m^{5}C ($K_{i,app}$ = 22.9 nM) displayed a significantly higher ability than cold-C ($K_{i,app}$ = 112.6 nM) to compete with tracer-C (Figure 1G). Taken together, these results support the view that SRSF2, *in vitro* and in live cells, preferentially binds m^{5}C-bearing RNAs.

Transcriptome-wide SRSF2-binding profile

To study SRSF2-RNA-binding sites comprehensively at the transcriptome-wide level, we performed photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation sequencing (PAR-CLIP-seq) in HeLa cells (Figures S2A and S2B; STAR Methods). A total of 10,928 SRSF2-binding sites within 6,844 transcripts were identified (Figure 2A; Table S2). SRSF2 was found mainly enriched in exonic regions (Figure 2B), consistent with the protein's known preferential binding to ESE.[25,26] The majority of SRSF2-binding transcripts were found to be protein-coding (86.25%), particularly enriched in coding sequence (CDS) region (73%) (Figures S2C and S2D). Subsequent motif analyses revealed at peak centers the presence of CAG(C/G)CUGR motif (Figure 2C) and of other SSNG-containing motifs such as (G/C)AG(G/A)AG and U(C/G)C(U/A)G (Figure S2E). Exemplary SRSF2-binding sites containing SSNG sequences are displayed in Figure 2D and validated by RNA immunoprecipitation-qPCR (RIP-qPCR) (Figures 2E, and S2F–S2H). Functional annotation analysis showed that SRSF2-binding targets are enriched in "RNA splicing" and "chromatin remodeling" categories (Figure S2I). Finally, we found that most of the SRSF2-binding targets are unique to this protein, with very little overlap with SRSF1- or SRSF3-binding sites (Figure S2J).

Overall, we find that SRSF2 binds mainly to the CDS regions of exons, preferentially at cytosine-guanine (CG)-rich SSNG motifs.

In the transcriptome, SRSF2 binds m^{5}C-bearing mRNAs

We then assessed the transcriptome-wide m^{5}C landscape. First, we evaluated and confirmed the binding and specificity of the m^{5}C antibody by performing m^{5}C-methylated RNA immunoprecipitation (m^{5}C MeRIP) followed by RT-qPCR (Figure S2K). We then conducted m^{5}C MeRIP followed by next-generation sequencing (Figure S2L; STAR Methods). A total of 6,913 m^{5}C peaks within 4,684 transcripts were identified (Figure 2F; Table S2), and among these m^{5}C peaks, the majority were located in protein-coding transcripts (92.72%, Figure S2M). In mRNA, the most abundant m^{5}C peaks were found in CDSs, accumulating in regions immediately downstream of translation initiation sites (Figure 2G). Interestingly, by integrating in-house SRSF2 PAR-CLIP-seq and RNA m^{5}C MeRIP-seqdata, we found that SRSF2-binding sites appeared very frequently at m^{5}C peak centers (Figure 2H). We also observed, by MeRIP-seq or published RNA bisulfite sequencing (RNA-BisSeq) data,[9] that among the m^{5}C-containing transcripts, around 40% were SRSF2 targets (Figure S2N). Furthermore, the percentage of m^{5}C sites associated with SRSF2-binding transcripts was the highest for the high-stoichiometry group (Figure S2O). The SRSF2-associated m^{5}C-methylated transcripts mainly involved biological processes such as "chromatin organization," "mRNA processing," and "RNA splicing" (Figure S2P). The top biological processes overrepresented in this analysis were likewise overrepresented among SRSF2-binding transcripts (Figure S2I). Together, these results provide evidence that SRSF2 binds directly to a subset of m^{5}C-marked sequences within the transcriptome.

NSUN2 depletion reduces m^{5}C levels and alters the RNA-binding affinity of SRSF2

Since SRSF2 binds preferentially to m^{5}C-modified RNAs, we next wondered how reduced m^{5}C marking might affect transcriptome-wide SRSF2 binding. To investigate this, we first verified that mRNA m^{5}C levels were significantly reduced in NSUN2 knockdown (KD) HeLa cells by m^{5}C mass spectrometry, dot blot, and m^{5}C MeRIP-seq (Figures 3A, 3B, and S3A). We then performed, on control and NSUN2 KD cells, SRSF2 PAR-CLIP followed either by RNA biotin-labeling assay or high-throughput sequencing. The RNA biotin-labeling assay revealed significantly reduced SRSF2 RNA binding upon NSUN2 KD (Figures S3B and S3C). As with HeLa control cells (Figure S2B), PAR-CLIP-seq on NSUN2 KD cells identified highly reproducible SRSF2-binding sites (Figure S3D). Differential binding analysis between NSUN2 KD and control revealed a total of 3,426 SRSF2 differential binding sites, of which approximately 65% showed loss of binding (called "siNSUN2-loss sites" in what follows) and 35% displayed gain of binding (referred to as "siNSUN2-gain sites") after NSUN2 KD (Figure 3C; Table S2). To better understand the gain in SRSF2 binding, we first evaluated the expression of the genes encoding another mRNA m^{5}C writer, NSUN6, and the m^{5}C erasers TET2 and ALKBH1. None of these genes showed differential expression after NSUN2 KD (Figure S3E). Hence, this does not support the hypothesis that the gain in SRSF2 binding is due to compensatory alteration of the expression of other m^{5}C regulators when NSUN2 is low. We next wondered how SRSF2 binding to SSNG motifs might change when m^{5}C levels are low. Although enrichment in the same motifs was observed, we found a C-containing motif (GCAG) to rank lower in gain sites than in loss sites, whereas a non-C-containing motif (GGGG) ranked higher (Figures 3D and S3F). This finding suggests redirection of SRSF2 toward non-C-containing binding sites when NSUN2 is reduced.

The majority of siNSUN2-loss sites were mapped to protein-coding transcripts, where they were mainly present in the CDS region (Figures S3G and S3H). siNSUN2-gain sites appeared

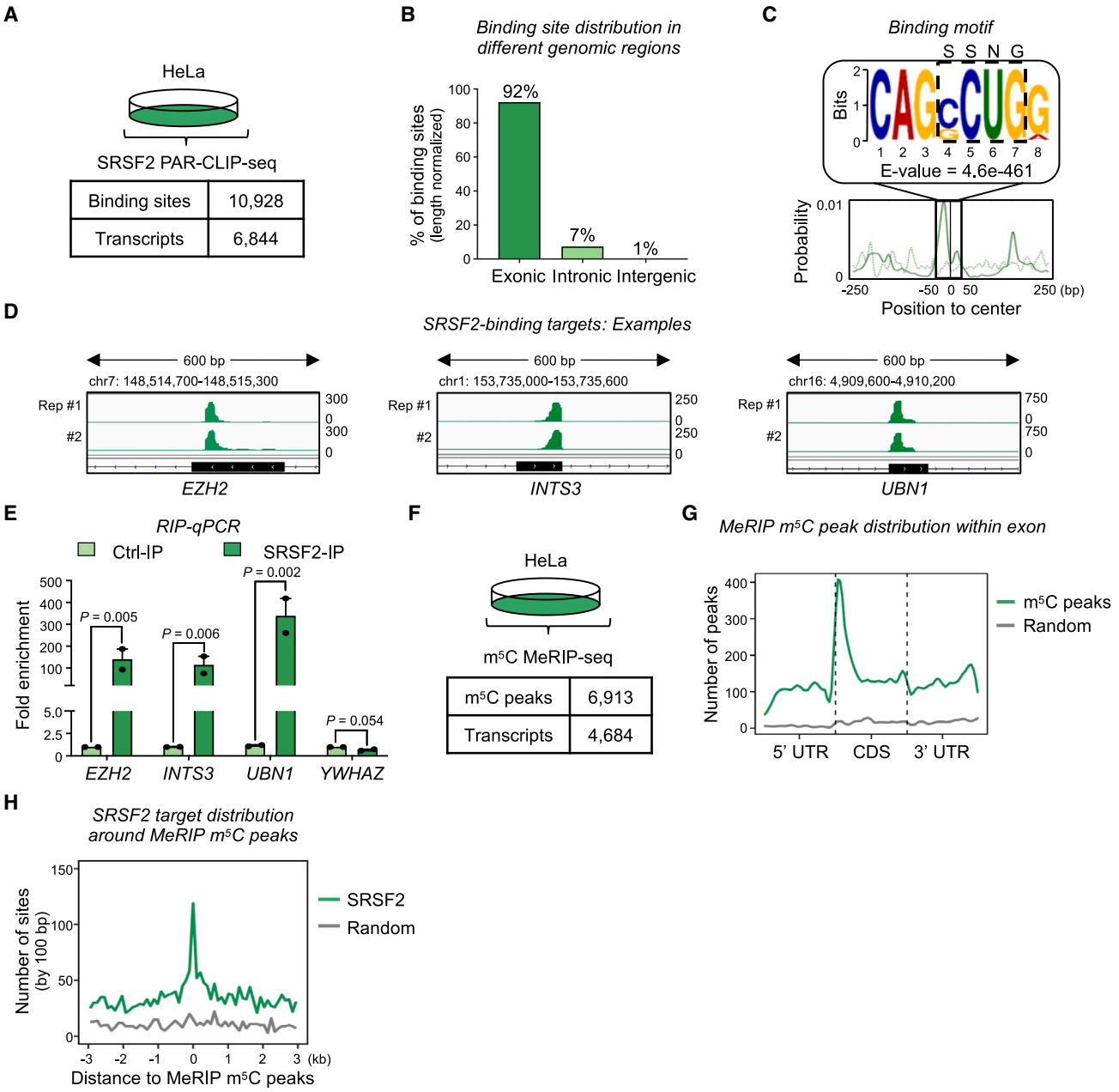


Figure 2. Transcriptome-wide SRSF2-binding profile, mRNA m^{5}C landscape, and co-occurrence of SRSF2 binding and m^{5}C

(A) RNA-binding sites and transcripts of SRSF2 identified by PAR-CLIP-seq in HeLa cells (n = 2).

(B) SRSF2 preferentially binds exons. The percentages in the bar chart were scaled using the total region length of each genomic region as the normalization factor.

(C) Canonical SSNG motif enriched at the centers of SRSF2-binding sites. Top: enriched motif, the E value is the enrichment p value (Fisher's exact test) times the number of candidate motifs tested.

(D) Integrative Genomics Viewer (IGV) tracks displaying exemplary SRSF2-binding sites.

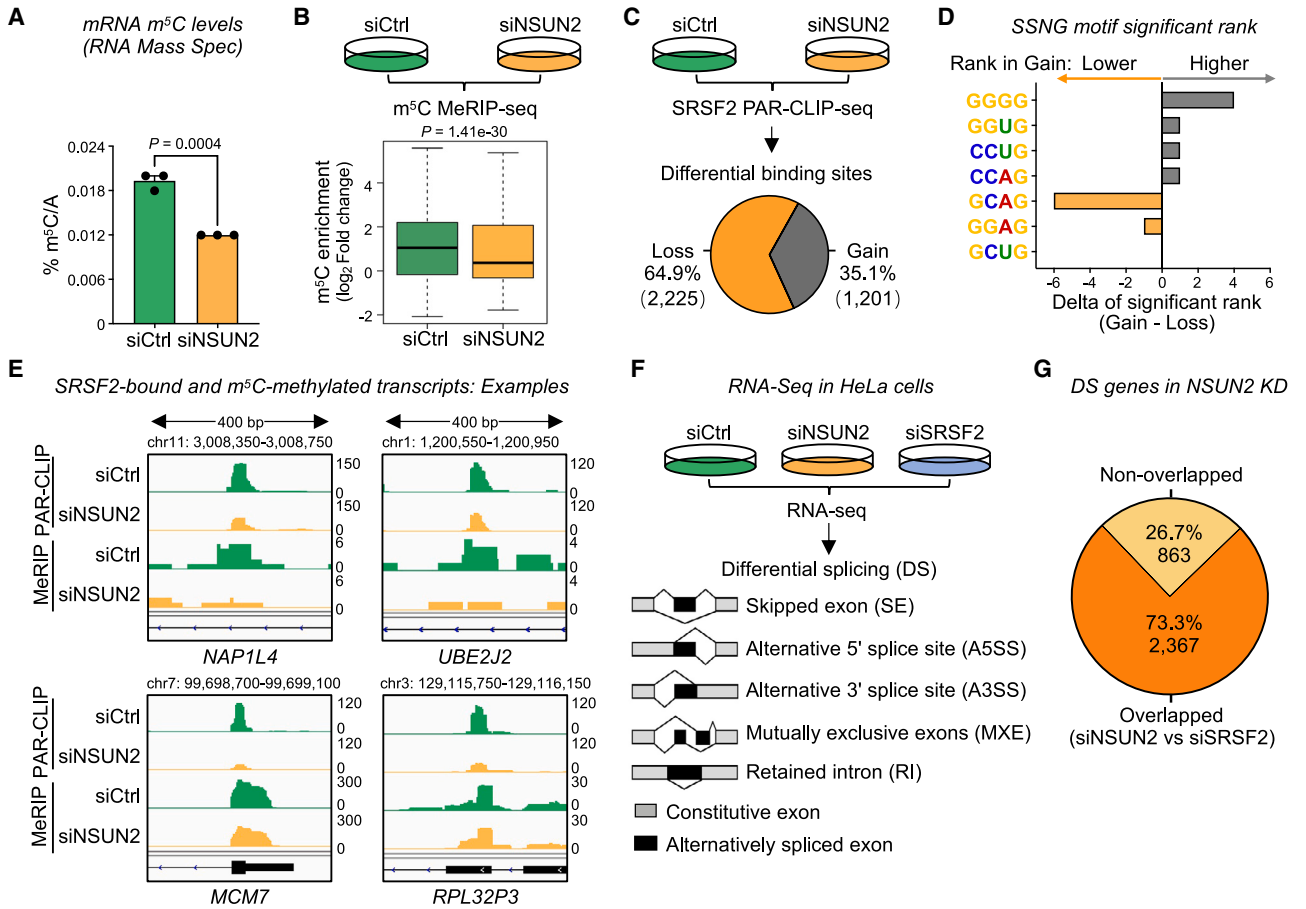
(E) RIP-qPCR validation of SRSF2 binding (n = 2, mean ± SEM, unpaired two-tailed Student's t test).

(F) RNA m^{5}C MeRIP-seq revealed the presence of m^{5}Cs within many transcripts (n = 2).

(G) mRNA m^{5}C peaks were found mainly in CDS regions, particularly those immediately downstream of translation start sites.

(H) Frequent proximity of SRSF2-binding sites and m^{5}C peak centers.

See also Figure S2 and Table S2.



(legend on next page)

comparably distributed between the CDS and 3′ UTR regions (Figure S3H). Representative coverage tracks for sites having lost SRSF2 binding and showing lower m^5C levels upon NSUN2 KD are displayed in Figure 3E. Transcripts containing siNSUN2-loss sites were enriched in categories for cell biology such as RNA splicing and oncogenesis-associated categories like "AML." In contrast, transcripts containing siNSUN2-gain sites showed an over-representation of the "ribosome," "rRNA processing," and "regulation of mRNA stability" categories (Figure S3I). We found that the differentially bound transcripts showed no significant difference in translation efficiency (Figure S3J), suggesting that altered SRSF2-binding profiles observed in NSUN2-depleted cells do not affect translation.

NSUN2 and SRSF2 depletion similarly alters RNA splicing, with enrichment of m^5C sites and SRSF2-binding sites near altered splicing events

Concerning the well-known function of SRSF2 in RNA splicing[18–20] and the fact that the RNA-splicing category is over-represented among SRSF2-binding targets showing siNSUN2-related loss of binding (Figure S3I), we hypothesized that NSUN2, by adding the m^5C mark to RNAs, affects SRSF2-driven alternative splicing. If so, NSUN2 depletion should result in alternative splicing pattern alterations similar to those caused by SRSF2 depletion. To test this hypothesis, we conducted RNA sequencing (RNA-seq) and analyzed RNA splicing (Figures 3F and S3K–S3M; Table S3). Notably, we observed a strong positive correlation of the differential splicing (DS) events between NSUN2 KD and SRSF2 KD (Figures S3N and S3O). Between siNSUN2 and siSRSF3 or siSRSF10, used as a negative control, the correlation was very weak (Figures S3N and S3O). Consistently, 73.3% of the DS genes identified in NSUN2 KD were also identified in SRSF2 KD cells (Figure 3G), and exemplary splicing events are represented in Figure 3H. Collectively, these data suggest that NSUN2 depletion has effects on alternative splicing similar to those of SRSF2 depletion.

Given this observation, we next investigated whether the m^5C modification and SRSF2 binding might occur at NSUN2- and SRSF2-associated splicing events. Our analysis using in-house m^5C MeRIP-seq data and publicly available RNA-BisSeq data[9] consistently showed a close proximity of SRSF2-binding sites and m^5C sites to the splicing events (Figures 3I and S3P). This strongly supports our finding that SRSF2 acts as an m^5C-binding protein and suggests an association between m^5C, SRSF2, and RNA splicing.

We further overlapped the co-occurring differentially spliced genes (2,367 genes) with SRSF2-binding targets. A significant subset of 1,058 SRSF2-binding targets were also differentially spliced (Figure 3J). These differentially spliced SRSF2-binding targets showed, notably, enrichment in "cell cycle," "gene expression," and "DNA repair" pathways (Figure S3Q). These observations, along with our findings that SRSF2 binds the m^5C mark, suggest that SRSF2 contributes to the alternative splicing effects of NSUN2-mediated m^5C through its reader function.

The prevalent disease-associated P95H mutation reduces the binding affinity of SRSF2 for RNA m^5C

Various somatic SRSF2 mutations are frequently reported in leukemia, and these alterations are crucial to pathogenesis.[23,27] The discovery that SRSF2 binds m^5C-containing RNA drove us to investigate whether these disease-associated mutations alter the preferential binding of SRSF2 to m^5C. To answer this question, we tested several mutations in the N-terminal region of SRSF2: T51A, K52A, P95H, H99A, and P107H.[28] Intriguingly, we found the other SRSF2 mutant forms assessed to maintain a preference for m^5C, in contrast to the P95H variant (Figure 4A).

Using NanoBRET, we found that compared with $SRSF2^{WT}$ ($K_{i,app}$ = 22.9 nM; Figure 1G), $SRSF2^{P95H}$ ($K_{i,app}$ = 43.4 nM; Figure 4B) showed a higher $K_{i,app}$ value, i.e., a lower affinity for the methylated RNA. These results concur to indicate that the P95H mutation reduces the affinity of SRSF2 binding to RNA m^5C.

Structural modeling of the interaction between m^5C and either WT or mutant SRSF2 and validation by equilibrium-binding affinity measurements

A previous NMR structure uncovered the mode of SRSF2 N-terminal domain and RNA interaction (PDB: 2LEB).[29] A single-stranded hexanucleotide RNA (5′-$U_1C_2C_3A_4G_5U_6$-3′) fits into a groove formed by positively charged and aromatic amino acids emanating from the central β sheet and hinge region (Lys91-His99) of SRSF2 (Figure 4C, left). Two direct hydrogen bonds between the Watson-crick edge of the C3 base and the side chain of Arg61 confer the base specificity for the second cytosine (C3).[29] Interestingly, the opposite face of the C3 base is stabilized by the van der Waals (vdW) contacts with Pro95. We modeled a m^5C at this position (C3) (Figure 4C, middle).

Figure 3. Depletion of NSUN2 reduces m^5C levels, alters the mRNA-binding affinity of SRSF2, and results in RNA-splicing changes similar to SRSF2 depletion

(A) Overall decrease in mRNA m^5C levels upon NSUN2 knockdown detected by quantitative liquid chromatography-mass spectrometry (LC-MS) analysis (n = 3, mean ± SEM).

(B) m^5C MeRIP-seq from control and NSUN2 KD HeLa cells (n = 2).

(C) Pie chart depicting the percentage and number of SRSF2-binding sites lost or gained in NSUN2 KD cells (n = 2).

(D) Preferential SRSF2 binding to SSNG-containing sequences was altered after NSUN2 knockdown.

(E) IGV tracks showing a decrease in SRSF2-RNA binding and m^5C levels in NSUN2 KD versus control cells.

(F) RNA-seq experimental design using siCtrl, siNSUN2, and siSRSF2 cells (n = 2).

(G) Majority of NSUN2 KD-mediated DS genes are associated with SRSF2.

(H) Exemplary sashimi plots showing concerted alternative splicing changes that occurred in cells depleted of SRSF2 or NSUN2.

(I) SRSF2-binding sites and m^5C sites occur frequently around NSUN2- and SRSF2-associated splicing events.

(J) Significant overlap between SRSF2-binding targets and overlapped DS genes identified in both siNSUN2 and siSRSF2 cells (genes from dark orange region in G). p values in (A), (B), and (J) were calculated using unpaired two-tailed Student's t test and hypergeometric test, respectively.

See also Figure S3 and Tables S2 and S3.

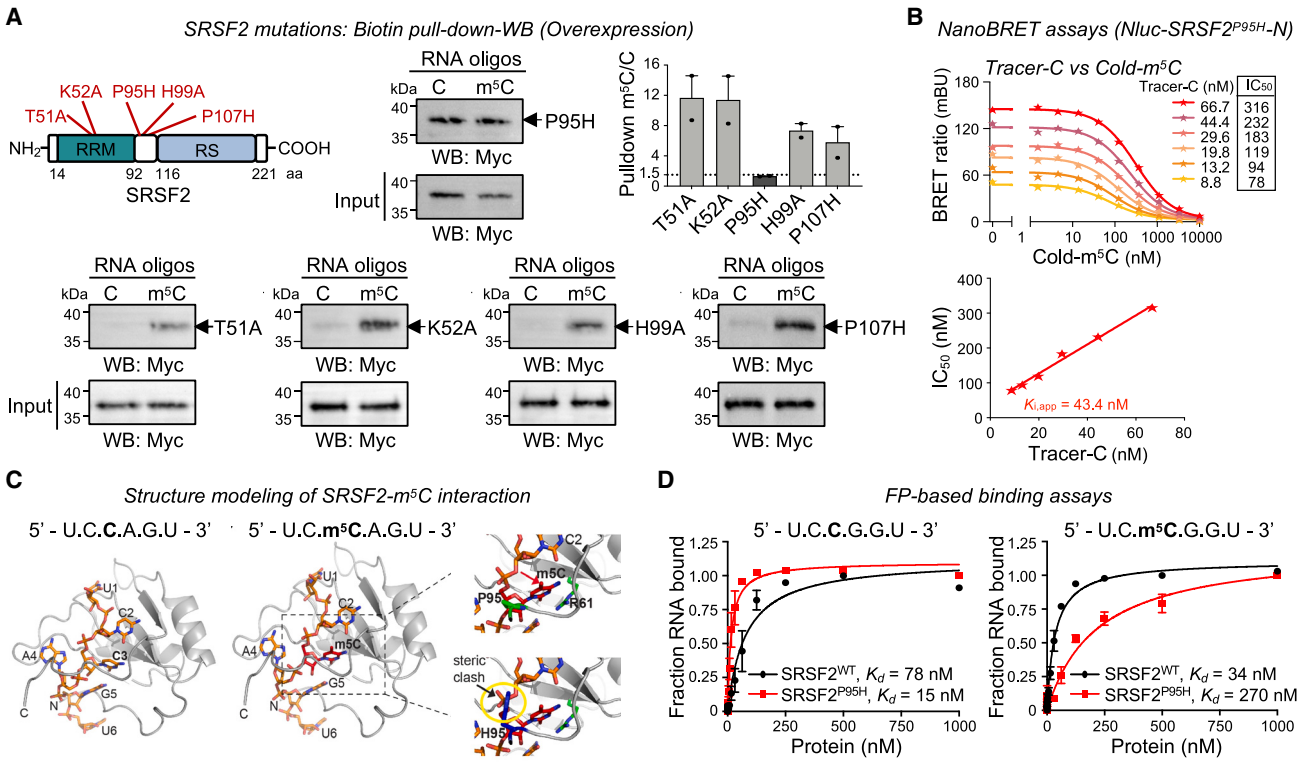


Figure 4. The SRSF2^{P95H} mutation reduces the m^{5}C-binding affinity of SRSF2

(A) Only the SRSF2^{P95H} mutant protein shows a decreased binding preference for m^{5}C-RNA (n = 2, mean ± SEM).

(B) NanoBRET target engagement assays using N-terminal SRSF2^{P95H} and titration with cold-m^{5}C in the presence of serial dilutions of tracer-C.

(C) Left: NMR structure of SRSF2/RNA complex, protein, gray cartoon; RNA, orange sticks. Middle: an m^{5}C base (red stick) is modeled at the position of C3 base. Right: close-up view of the m^{5}C-binding pocket of wild-type SRSF2 (upper) and P95H mutant (modeled histidine, blue).

(D) Binding isotherms from FP assays show preferential binding of the N-terminal domain of SRSF2 WT and P95H mutant to methylated and unmethylated RNA hexanucleotide, respectively (n = 3, mean ± SEM).

See also Figure S4 and Table S1.

Interestingly, the methyl moiety of m^{5}C appears to be stabilized by additional vdW contacts with protein backbone atoms of Arg94 and Pro95 from one side and ribose moiety of the first cytosine (C2) of RNA from the other (Figure 4C, right upper). Importantly, modeling with the other three variants of the UCm^{5}CNGU sequence also revealed stabilization of m^{5}C binding to SRSF2 via additional vdW contacts (Figure S4A). These modeling results are consistent with the NanoBRET data (Figures S1H–S1J). Furthermore, it is conceivable that a bulkier histidine residue at position 95 would disrupt these contacts, resulting in weaker binding of the P95H mutant protein to an m^{5}C-containing RNA. Consistently, the side chain of a modeled histidine sterically clashes with the phosphate backbone of RNA (Figure 4C, right lower). In addition, we modeled the interaction of other SRSF2 mutants (Figure S4B). Arg95 (R95), a less frequent mutation in AML/CMML patients than H95,[24,30] may also sterically clash with m^{5}C. Ala95 (A95), a rare mutation in AML/CMML patients,[24,30] might be less detrimental and would appear not to clash with the m^{5}C base or the RNA backbone. Together, these data highlight the crucial role of Pro95 in the SRSF2-m^{5}C interaction. Finally, fluorescence polarization (FP)-based assays experimentally confirm observations that the wild-type (WT) SRSF2 RRM binds more tightly to an m^{5}C-containing RNA, whereas the P95H mutant prefers the unmethylated RNA sequence (Figure 4D).

Thus, our structural studies together with FP assays suggest a molecular mechanism of specific recognition of m^{5}C-modified RNA by SRSF2 and thereby might explain how WT SRSF2 binds more tightly to an m^{5}C-containing RNA, whereas leukemia-associated Pro95 mutants, such as the P95H mutant, prefer the unmethylated RNA sequence.

RNA-binding profile of SRSF2 in NSUN2 KD and P95H-mutant leukemic cells

We then explored in a leukemic cell model how the P95H mutation and RNA hypomethylation affect SRSF2 binding to mRNA. To characterize the intracellular effects of low m^{5}C levels, we generated a stable NSUN2 KD (shNSUN2) chronic myeloid leukemia cell line (K562) and verified the overall low mRNA m^{5}C abundance (Figures S5A and S5B). We then performed PAR-CLIP-seq on shNSUN2 and SRSF2^{P95H} K562 cells to identify SRSF2-binding targets on mRNA (Figures S5C and S5D). Focusing on the sites showing differential SRSF2 binding, we found a total of 1,933 SRSF2-binding sites, identified in control

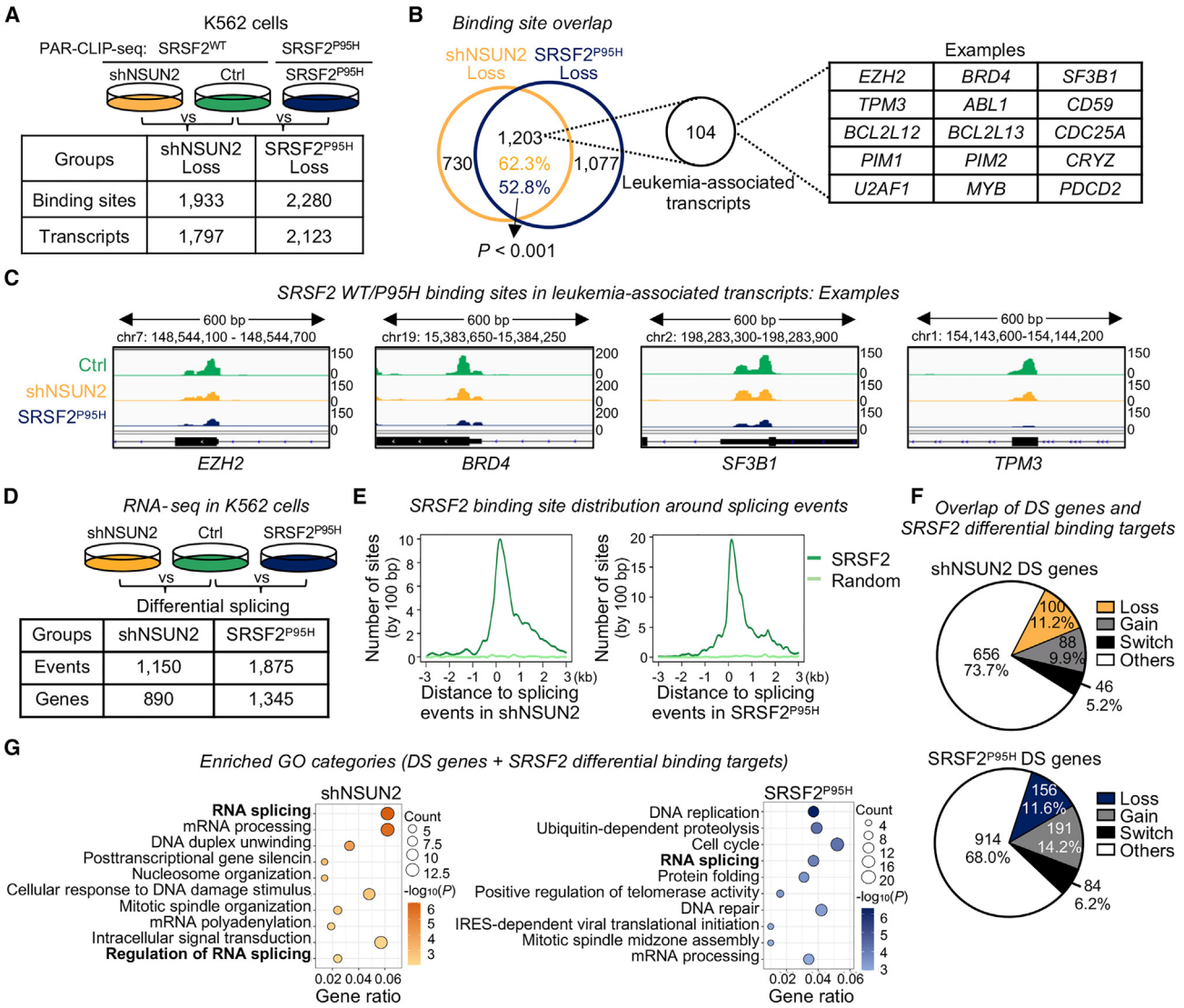


Figure 5. Involvement of mRNA m^{5}C regulatory transcripts in leukemia

(A) SRSF2 RNA-binding sites and transcripts identified by PAR-CLIP-seq in K562 cells (n = 2).

(B) Many SRSF2^{WT} preferential binding sites are NSUN2-dependent binding and 104 of the corresponding transcripts are leukemia-associated.

(C) IGV profiles show reduced binding of SRSF2 in NSUN2 KD or SRSF2^{P95H} mutant K562 cells.

(D) Schematic of RNA-seq experimental design using K562 cells (n = 2).

(E) SRSF2-binding sites occur preferentially around NSUN2- and SRSF2^{P95H}-associated splicing events.

(F) Pie chart displaying the percentage of DS genes that are differentially bound by SRSF2 in NSUN2-depleted or SRSF2 mutant cells.

(G) Differentially spliced SRSF2-binding targets in NSUN2-depleted or SRSF2 mutant cells are significantly enriched in the RNA-splicing category.

See also Figure S5 and Tables S4 and S5.

cells, to be lost in shNSUN2 cells and 2,280 binding sites to be lost in SRSF2^{P95H}-mutant cells (Figure 5A; Table S4). We next compared the distributions of the following subsets of sites: shNSUN2-loss or -gain sites (loss or gain upon NSUN2 depletion) and P95H-loss or -gain sites (loss or gain in SRSF2^{P95H} cells). The majority of differential binding sites aligned to exonic regions in protein-coding transcripts (Figures S5E and S5F), in keeping with the results obtained for HeLa cells.

Previous experiments have shown the relative binding affinity of SRSF2^{P95H} for the different SSNG variants is CCNG > GCNG > CGNG > GGNG.[26] Consistent with these *in vitro* findings, P95H-gain sites were more enriched in CCNG and GCNG motifs, but not in CGNG and GGNG motifs (Figure S5G). Therefore, our intracellular binding motif analyses provide evidence that the SRSF2^{P95H} mutation causes alteration rather than loss of the protein's normal SSNG motif-binding activity.

When we compared the sites showing a loss of binding under these two conditions, we observed an overlap of 1,203 binding sites, corresponding to 62.3% of the shNSUN2-loss sites (Figure 5B). This result suggests that NSUN2 depletion

and the SRSF2^{P95H} mutation might similarly affect SRSF2 binding to some targets. Strikingly, 104 binding sites in the overlap zone are known to encode leukemia-related genes, e.g., enhancer of zeste homolog 2 (*EZH2*), bromodomain protein 4 (*BRD4*), splicing factor 3B subunit 1 (*SF3B1*), and tropomyosin 3 (*TPM3*) (Figures 5B and 5C; Table S4). The fact that both NSUN2 KD and the SRSF2^{P95H} mutation alter SRSF2 binding to mRNA, particularly to leukemia-associated targets, highlights a potential involvement of m^{5}C recognition in leukemogenesis.

NSUN2 depletion leads to global RNA-splicing alterations comparable to that of SRSF2 mutations

We first examined whether altered SRSF2-binding profiles observed in NSUN2-depleted or SRSF2^{P95H} mutant cells are associated with translation. We found that the translation was not affected (Figure S5H). It has been shown that SRSF2^{P95H} mutant switches the RNA-splicing profile on a large panel of genes involved in cancer development.[26,31,32] Therefore, we performed RNA-seq in NSUN2 KD and SRSF2^{P95H} K562 cells to analyze the RNA-splicing patterns (Figure 5D; Table S5). We observed a strong positive correlation of splicing events (Figures S5I and S5J), suggesting that NSUN2 depletion leads to a global RNA-splicing alteration comparable with the SRSF2 mutation. 7 of the top 12 enriched pathways for those DS genes were overrepresented in both contexts (Figure S5K). This suggests that NSUN2 depletion- and SRSF2 mutation-mediated RNA-splicing alterations co-impact many downstream biological functions.

We next investigated the distance of SRSF2-binding sites from alternative splicing event locations. In agreement with findings in HeLa cells (Figure 3I), we found SRSF2-binding sites identified in control cells, but not randomly selected sites, to be located preferentially around splicing events identified in NSUN2-depleted or SRSF2^{P95H} mutant cells (Figure 5E). Furthermore, we found that approximately 26%–32% of differentially spliced genes were SRSF2-binding targets that were altered upon NSUN2 depletion or SRSF2 mutation (Figure 5F). Intriguingly, these differentially spliced SRSF2-binding targets were significantly enriched in the RNA-splicing category (Figure 5G). These results suggest that NSUN2 depletion and SRSF2 mutation led to alternative splicing of the direct SRSF2-binding targets and the indirect targets by affecting the binding and splicing of other RNA-splicing factors.

Distribution of RNA m^{5}C in monocytes of CMML patients with high or low *NSUN2* levels

To profile transcriptome-wide m^{5}C methylation in leukemia patients at single-base resolution, we isolated peripheral blood monocytes from eight CMML patients and performed RNA-BisSeq on ribo-depleted RNAs (Figures 6A and S6A; Table S6; STAR Methods). We found that NSUN2-low patients had a significantly lower number of m^{5}C sites than NSUN2-high patients (Figure 6B). The majority of m^{5}C sites were mapped to protein-coding transcripts (Figure S6B). The median methylation level of all identified mRNA m^{5}C sites was 16.7%, with more than 30% of m^{5}C sites showing methylation level over 20% (Figure S6C), in agreement with previous observations on human bladder urothelial carcinoma tissues.[15] A sequence frequency logo showed the m^{5}C sites to be embedded in environments with high CG content (Figure S6D). The distribution profiles of m^{5}C sites in mRNA were then examined, and the most highly m^{5}C-associated region was found to be the CDS, particularly the region immediately downstream of the translation initiation site (Figures 6C and S6E). These patterns are consistent with our m^{5}C MeRIP-seq data for HeLa cells and with previous reports on mouse tissues and both normal and tumor-derived human tissues.[9,15,33] Remarkably, NSUN2-low patients showed a less frequent occurrence of m^{5}C sites in mRNA exonic regions (especially CDSs) than NSUN2-high patients (Figure 6C).

Given the above observation that m^{5}C site counts were lower in NSUN2-low patients, we further compared methylation levels in m^{5}C-marked mRNA transcripts. We observed a significant reduction of m^{5}C levels in NSUN2-low patients (Figure 6D). Consistently, the heatmap showed that most m^{5}C-modified transcripts were hypomethylated in NSUN2-low patients (Figure 6E). These results indicate that a low NSUN2 level leads to low m^{5}C levels in CMML patient monocytes. Gene set enrichment analysis (GSEA) showed that the inflammatory response pathway was significantly overrepresented, and showed a strong negative correlation with m^{5}C differences in NSUN2-low patients compared to NSUN2-high patients (Figure 6F). Of note, the transcriptional signature of CMML monocytes has been reported to be highly inflammatory, contributes to malignant expansion, and reflects leukemia-specific and age-related alterations.[34]

Low expression of *NSUN2*, but not *NSUN6*, is significantly associated with poor prognosis in AML patients with the SRSF2^{P95H} mutation

We next explored the expression levels of *NSUN2* in a larger number of leukemia patients and found that *NSUN2* expression was significantly downregulated in CMML and AML patients (Figure 7A). Expression of *NSUN6* showed no significant differences (Figure S7A). The overall low expression of *NSUN2* in patients prompted us to investigate the clinical role of *NSUN2*.

To explore the clinical relevance of m^{5}C-related genes in leukemia, we first performed survival analysis on a public dataset consisting of 246 AML patients (tagged "Bamopoulos et al.").[35] SRSF2^{P95H} patients had shorter overall survival (OS) than non-P95H mutant (referred to as "WT") patients (Figure S7B), as previously reported.[35] We then investigated the relationship between the abundance of m^{5}C writer NSUN2 and patient prognosis. The *NSUN2*-high and -low WT patients were found not to differ significantly in OS. Strikingly, however, the *NSUN2*-low SRSF2^{P95H} group showed a significantly worse prognosis, with a 1-year survival rate of only 20% (Figure 7B). Consistently, single Cox proportional hazards regression analysis showed that for *NSUN2*-low SRSF2^{P95H} mutant patients, the average risk of death exceeded that of patients with *NSUN2*-high WT by approximately 251% (Figure 7C). These findings were validated by the analysis of another cohort (the Beat AML cohort,[36] containing 451 samples) (Figures 7D, 7E, and S7C). We next evaluated the expression of key leukemia-associated genes in the four groups of patients in both AML cohorts. In *NSUN2*-low patients with the SRSF2^{P95H} mutation, importantly, orosomucoid 1 (ORM1) and lipocalin-1 (LCN2),[37,38] oncogenes known to be associated with leukemia development and progression showed significantly higher

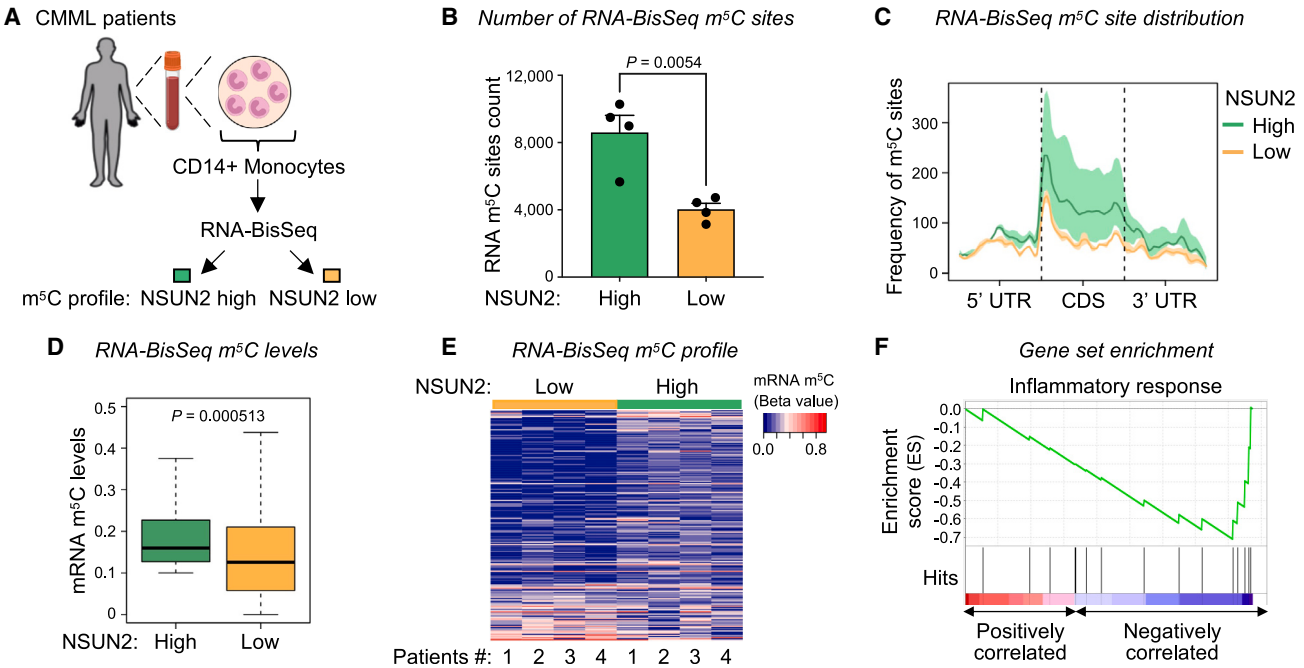


Figure 6. Transcriptome-wide distribution of RNA m^{5}C in monocytes of CMML patients with high or low *NSUN2* levels

(A) RNA-BisSeq experimental design using ribo-depleted RNAs from peripheral blood CD14$^+$ monocytes of eight CMML patients.

(B) NSUN2-low patients have a significantly lower number of m^{5}C sites than NSUN2-high patients (mean ± SEM).

(C) mRNA m^{5}C sites occur more frequently in CDS regions than in UTR regions.

(D) Boxplot showing the median m^{5}C levels of methylated protein-coding transcripts in NSUN2-high patients are significantly higher than that of the same transcripts in NSUN2-low patients.

(E) Heatmap showing correlation of mRNA m^{5}C levels in NSUN2-high and -low patients.

(F) Genes with differential m^{5}C levels are associated with inflammatory response pathways. The p values in (B) and (D) were calculated with the unpaired two-tailed Student's t test.

See also Figure S6 and Table S6.

expression (Figures 7F and 7G). We also investigated the relationship between *NSUN6* expression and patient prognosis. However, the prognosis of *NSUN6*-low SRSF2^{P95H} patients was not consistent between the two cohorts, and the oncogenes *ORM1* and *LCN2* were not overexpressed (Figures S7D–S7I). This could be due to the fact that NSUN2 and NSUN6 have different sets of RNA substrates, since two different types of m^{5}C sites are reported to exist in mRNAs, targeted by NSUN2 or NSUN6, respectively.[10]

Altogether, these results show that low expression of *NSUN2*, but not *NSUN6*, is reproducibly associated with poor prognosis and low expression of some oncogenes in patients with SRSF2^{P95H} mutation. This suggests a potential role for NSUN2 as a prognostic marker in SRSF2^{P95H} mutant AML patients and highlights an unrecognized link between NSUN2, SRSF2^{P95H}, and oncogenesis.

DISCUSSION

Modifications of mRNA control the fate of the modified mRNAs, mainly by recruiting binding proteins. Only a few mRNA m^{5}C-binding proteins have been identified so far, and we are only beginning to understand the m^{5}C machinery and its biological functions. Our findings add a player, SRSF2, to the list of m^{5}C readers. Our results suggest that the role of NSUN2-dependent m^{5}C mRNA, mediated in part through SRSF2 binding, is an important, previously underestimated, feature in the context of leukemia.

Using structural modeling, we found the cytosine bearing the methyl group to be stabilized by two hydrogen bonds and specifically recognized by Arg61 of SRSF2. Proline 95 further stabilizes this methyl group of m^{5}C from the other side, but in SRSF2^{P95H}, the side chain of His95 moves the phosphate of RNA away from the methyl group resulting in the loss of a critical stabilizing contact. These results might explain how WT SRSF2 binds more tightly to an m^{5}C-modified RNA and why proline 95 is critical in stabilizing the interaction. The preferential binding of SRSF2 to m^{5}C is similar to that of other readers, such as another RNA m^{5}C reader YBX1 and DNA 5mC readers methyl-cytosine binding domain protein 4 (MBD4) and kinesin superfamily protein member 4 (Klf4),[15,16,39,40] which also show binding to both unmodified and modified targets but prefer the latter. One should note that SRSF2 does not always show a preference for m^{5}C-marked sites. Sajini et al. report that SRSF2 is repelled by m^{5}C on a vault RNA.[41] This suggests that the role of SRSF2 as a reader of m^{5}C is part of a more complex picture.

SRSF2 is a multifunctional protein involved in regulating RNA splicing, transcriptional elongation, and RNA stability.[42–44] The m^{5}C mark, on the other hand, has been shown to

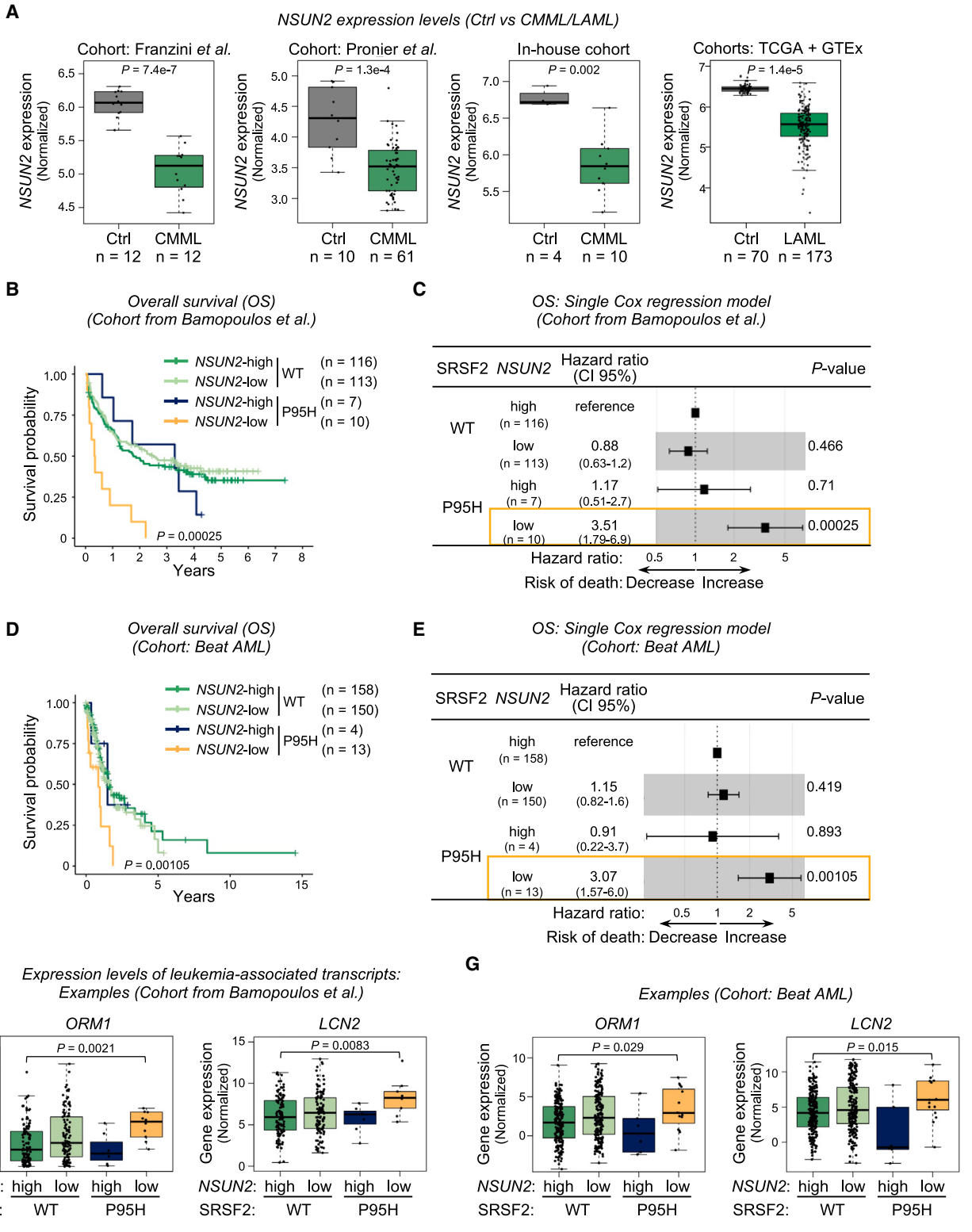


Figure 7. Low *NSUN2* expression is associated with poor prognosis in AML patients with the SRSF2^{P95H} mutation

(A) *NSUN2* expression is lower in CMML and AML patients than in healthy controls. The p value comparing the data from the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) database was computed by the web server Gene Expression Profiling Interactive Analysis 2 (GEPIA2). All other p values were calculated with the Wilcoxon test.

(legend continued on next page)

promote mRNA export and enhance RNA stability.[9,15,16] Our findings suggest a possible role for m^5C in regulating alternative splicing through the recruitment of SRSF2. It has indeed been shown that an NSUN2 deficiency and concomitant loss of m^5C residues can dysregulate HIV-1 mRNA splicing.[45] Of note, there is a similar finding in m^6A that depletion of writer methyltransferase-like protein 3 (METTL3) and reader heterogeneous nuclear ribonucleoprotein A2/B1 (HNRNPA2B1) causes similar changes to alternative splicing.[46] The association of RNA modification with SR family proteins in modulating RNA splicing has been reported, for example, m^6A modification appears to affect the RNA-binding ability of SRSF2 and thus influence the splicing outcome of genes regulated by SRSF2.[47] Evidence from previous and current studies highlights the importance of RNA modifications as an additional layer of RNA-splicing regulation on top of *cis*-regulatory sequences and *trans*-acting factors. A thorough mechanistic understanding of the interplay between m^5C, m^6A, and SR proteins will be a challenge for future studies.

NSUN2 has been shown to be highly expressed in multiple tumor types, such as hepatocellular carcinoma, gastric cancer, and prostate cancer.[48] Here, we find that *NSUN2* is lowly expressed and that low *NSUN2* levels correlate positively with RNA m^5C hypomethylation in CMML patients. TET2 is an m^5C eraser that is frequently found to be mutated in patients with myeloid malignancies, and notably in approximately 50% of CMML cases, 30% of MDS cases, and 10% of AML cases.[49] As *TET2* is a tumor-suppressor gene, *TET2* mutations are associated with myeloid expansion and tumor progression.[50] However, the correlation of *TET2* mutations with RNA methylation levels in leukemia needs to be investigated. Along with the frequently found SRSF2 mutation, m^5C writer, eraser, and reader dysregulation have all been linked to leukemia. Further studies are needed to gain insights into the mechanisms through which these regulators coordinate to contribute to the role of m^5C in leukemia.

When associated with MDS, SRSF2 mutations portend a poor outcome.[51] Here, in AML patients, we demonstrate an association between poor prognosis and a combination of *NSUN2* downregulation and the presence of $SRSF2^{P95H}$ mutation. To explain this finding, it is worth mentioning that the P95H mutation in patients was heterozygous, i.e., a WT copy of SRSF2 was retained in the genome. We speculate that in SRSF2-mutant patients with high *NSUN2* levels, these retained WT SRSF2 proteins are sufficient to bind to some m^5C-associated transcripts and thus partially maintain some essential biological functions. However, when *NSUN2* levels are low, only a small fraction of transcripts is m^5C-methylated, and this results in reduced WT SRSF2 binding. On the basis of the survival results obtained for two independent cohorts, it appears that the combination of these two factors (loss of m^5C affinity for the P95H mutant and reduced m^5C levels due to low *NSUN2* levels) is required to produce a significantly poor prognosis. As we have further demonstrated that this combination favors increased expression of leukemia-related oncogenes, our data strongly suggest a link between aberrant NSUN2-associated m^5C marking and hematologic malignancies. This warrants an in-depth investigation of the underlying mechanisms, with a view to developing new therapies.

In conclusion, we have discovered a previously unrecognized reader of m^5C on mRNA: the protein SRSF2, well known for its involvement in splicing and whose mutation at residue 95 (P95H) is strongly associated with hematologic malignancies. Furthermore, we have uncovered a previously unknown association between NSUN2/m^5C and SRSF2-mediated RNA splicing. Strikingly, in leukemia patients, *NSUN2* is lowly expressed, and this correlates with low m^5C methylation levels. The co-occurrence of low NSUN2 with SRSF2 mutation predicts poor prognosis. Although the path from mutation to disease remains to be fully elucidated, our work suggests that impairment of the SRSF2 m^5C reader function can contribute to leukemia progression. Overall, our data identify unrecognized mechanistic crosstalk between RNA modifications and an important mutation-dependent factor.

Limitations of the study

First, it could be that immortalized cell lines, used here to identify the SRSF2 RNA-binding profile, m^5C landscape, and RNA splicing, do not fully recapitulate what happens *in vivo*. In CMML patients, we have identified m^5C methylation profiles and observed low m^5C levels due to low NSUN2 levels. The *in vivo* consequences of RNA splicing and whether SRSF2 binds directly to the identified m^5C-modified targets remain to be investigated. Second, although we found a strong positive correlation between low NSUN2 expression, poor prognosis, and overexpression of several oncogenes in $SRSF2^{P95H}$-mutated AML patients, the underlying mechanisms are unclear because of potential confounding effects from multiple pathways. Therefore, characterization of relevant pathways and factors is crucial to fully understanding such mechanisms and to undertaking therapeutic targeting efforts. In conclusion, our work on AML and CMML patients provides a framework that can be broadened in the future to include other types of leukemia.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

(B and D) AML patients with $SRSF2^{P95H}$ and low *NSUN2* expression have worse overall survival in the Bamopoulos et al. (B) and Beat AML (D) cohorts. p values were determined with the log-rank test.
(C and E) $SRSF2^{P95H}$ with low *NSUN2* expression is associated with higher risk of death (log-rank test).
(F and G) High leukemia-associated oncogene expression in AML patients with $SRSF2^{P95H}$ and low *NSUN2* expression (Wilcoxon test).
See also Figure S7.

- Cell lines
- Human specimens

- METHOD DETAILS
  - RNA interference and transfection
  - Expression plasmids and site-directed mutagenesis
  - Histidine-tagged protein purification
  - FP-based binding assay
  - Biotinylated RNA pull-down assay
  - *In vitro* RNA pull-down assay and bioanalyzer analysis
  - NanoBRET assay
  - m^5C dot blotting
  - Mass spectrometry analysis of m^5C
  - Reverse-transcriptase quantitative PCR
  - RNA immunoprecipitation-qPCR
  - RNA-seq
  - PAR-CLIP
  - m^5C MeRIP-RT-qPCR
  - m^5C MeRIP-seq
  - RNA-BisSeq of CMML patients
  - RNA-seq and m^5C MeRIP-seq preprocessing
  - PAR-CLIP, m^5C MeRIP-seq, and RNA-BisSeq annotation
  - RNA-seq data analysis
  - m^5C MeRIP-seq analysis
  - PAR-CLIP-seq data analysis
  - Functional enrichment analysis
  - Overlaps between SRSF1-3 binding sites
  - Similarities between SR protein motifs and RNA probe
  - RNA-BisSeq analysis of the CMML cohort
  - Integrated analysis of SRSF2 targets and m^5C sites
  - Distribution of SRSF2 targets, m^5C, and DS events
  - Translation efficiency analysis
  - *NSUN2*/*NSUN6* expression level analysis
  - Overall survival analysis
  - Leukemia-related gene expression analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2023.11.003.

ACKNOWLEDGMENTS

H.-L.M., C.S.D.C., and L.M. were supported by "Télévie." J.L. was supported by BELSPO and Télévie. E.J.d.B. was supported by the Belgian F.R.I.A. and the Jean and Rose Hoguet Foundation. M.B., E.C., C.H., and B.H. were supported by the FNRS. F. Murisier was supported by the Walloon region (Win2Wal). F.F. is a ULB professor. R.D. is a ULB lecturer. F.F.'s lab was funded by grants from the FNRS and Télévie, the "Action de Recherche Concertée" (ARC) (AUWB-2018-2023 ULB-No 7), Wallon Region grants U-CAN-REST and INTREPID (1710179-WALLINOV, INTREPID RW 7787), an FNRS Welbio grant (FNRS-WELBIO-CR-2017A-04 and FNRS-WELBIO-CR-2019A-04R), the ULB Foundation, H2020-MSCA-ITN ROPES, and the Belgian Foundation against Cancer (FCC 2016-086, FAF-F/2016/872). Y.K.G.'s lab was supported by grants from NIH NIAID (1R01AI161363) and the Welch Foundation (AQ-2101). Y.-G.Y.'s lab was funded by the National Natural Science Foundation of China (NSFC) Science Fund for Creative Research Groups (32121001), the National Natural Science Foundation of China (92053115), and CAS Key Research Projects of the Frontier Science (QYZDY-SSW-SMC027).

AUTHOR CONTRIBUTIONS

Conceptualization, H.-L.M., M.B., C.S.D.C., E.J.d.B., and F.F.; methodology, H.-L.M., C.S.D.C., and M.B.; investigation, H.-L.M., E.J.d.B., E.C., K.M.R., J.I.B., S.A., B.H., C.S.D.C., A.L., C.H., F. Mies, P.P., J.L., and X.W.; formal analysis, M.B., H.-L.M., F. Murisier, M.-K.W., Q.-Y.Z., Y.K.G., and K.-T.L.; visualization, H.-L.M., M.B., and F. Murisier; writing – original draft, H.-L.M., M.B., and F.F.; writing – review & editing, H.-L.M., M.B., F. Murisier, L.M., E.S., Y.-G.Y., and F.F.; project administration, R.D., Y.Y., and D.L.D.; funding acquisition, F.F. and Y.-G.Y.; resources, A.Y., O.A.-W., N.D., and E.S.; supervision, Y.-G.Y. and F.F.

DECLARATION OF INTERESTS

F.F. is a co-founder of Epics Therapeutics (Gosselies, Belgium).

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: July 29, 2022
Revised: September 6, 2023
Accepted: November 3, 2023
Published: December 7, 2023

REFERENCES

1. Shi, H., Chai, P., Jia, R., and Fan, X. (2020). Novel insight into the regulatory roles of diverse RNA modifications: re-defining the bridge between transcription and translation. Mol. Cancer *19*, 78.
2. Murakami, S., and Jaffrey, S.R. (2022). Hidden codes in mRNA: control of gene expression by m6A. Mol. Cell *82*, 2236–2251.
3. Huang, H., Weng, H., and Chen, J. (2020). m6A modification in coding and non-coding RNAs: roles and therapeutic implications in cancer. Cancer Cell *37*, 270–288.
4. Xue, C., Chu, Q., Zheng, Q., Jiang, S., Bao, Z., Su, Y., Lu, J., and Li, L. (2022). Role of main RNA modifications in cancer: N6-methyladenosine, 5-methylcytosine, and pseudouridine. Signal Transduct. Target. Ther. *7*, 142.
5. Guo, G., Pan, K., Fang, S., Ye, L., Tong, X., Wang, Z., Xue, X., and Zhang, H. (2021). Advances in mRNA 5-methylcytosine modifications: detection, effectors, biological functions, and clinical relevance. Mol. Ther. Nucleic Acids *26*, 575–593.
6. Selmi, T., Hussain, S., Dietmann, S., Heiß, M., Borland, K., Flad, S., Carter, J.M., Dennison, R., Huang, Y.L., Kellner, S., et al. (2021). Sequence- and structure-specific cytosine-5 mRNA methylation by NSUN6. Nucleic Acids Res. *49*, 1006–1022.
7. Shen, Q., Zhang, Q., Shi, Y., Shi, Q., Jiang, Y., Gu, Y., Li, Z., Li, X., Zhao, K., Wang, C., et al. (2018). Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation. Nature *554*, 123–127.
8. Kawarada, L., Suzuki, T., Ohira, T., Hirata, S., Miyauchi, K., and Suzuki, T. (2017). ALKBH1 is an RNA dioxygenase responsible for cytoplasmic and mitochondrial tRNA modifications. Nucleic Acids Res. *45*, 7401–7415.
9. Yang, X., Yang, Y., Sun, B.F., Chen, Y.S., Xu, J.W., Lai, W.Y., Li, A., Wang, X., Bhattarai, D.P., Xiao, W., et al. (2017). 5-methylcytosine promotes mRNA export-NSUN2 as the methyltransferase and ALYREF as an m 5 C reader. Cell Res. *27*, 606–625.
10. Liu, J., Huang, T., Zhang, Y., Zhao, T., Zhao, X., Chen, W., and Zhang, R. (2021). Sequence- and structure-selective mRNA m5C methylation by NSUN6 in animals. Natl. Sci. Rev. *8*, nwaa273.
11. Wang, X., Wang, M., Dai, X., Han, X., Zhou, Y., Lai, W., Zhang, L., Yang, Y., Chen, Y., Wang, H., et al. (2022). RNA 5-methylcytosine regulates YBX2-dependent liquid-liquid phase separation. Fundamental Research *2*, 48–55.
12. Dai, X., Gonzalez, G., Li, L., Li, J., You, C., Miao, W., Hu, J., Fu, L., Zhao, Y., Li, R., et al. (2020). YTHDF2 binds to 5-methylcytosine in RNA and modulates the maturation of ribosomal RNA. Anal. Chem. *92*, 1346–1354.

13. Chen, H., Yang, H., Zhu, X., Yadav, T., Ouyang, J., Truesdell, S.S., Tan, J., Wang, Y., Duan, M., Wei, L., et al. (2020). m5C modification of mRNA serves a DNA damage code to promote homologous recombination. Nat. Commun. *11*, 2834.

14. Yang, H., Wang, Y., Xiang, Y., Yadav, T., Ouyang, J., Phoon, L., Zhu, X., Shi, Y., Zou, L., and Lan, L. (2022). FMRP promotes transcription-coupled homologous recombination via facilitating TET1-mediated m5C RNA modification demethylation. Proc. Natl. Acad. Sci. USA *119*, e2116251119.

15. Chen, X., Li, A., Sun, B.F., Yang, Y., Han, Y.N., Yuan, X., Chen, R.X., Wei, W.S., Liu, Y., Gao, C.C., et al. (2019). 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. Nat. Cell Biol. *21*, 978–990.

16. Yang, Y., Wang, L., Han, X., Yang, W.L., Zhang, M., Ma, H.L., Sun, B.F., Li, A., Xia, J., Chen, J., et al. (2019). RNA 5-methylcytosine facilitates the maternal-to-zygotic transition by preventing maternal mRNA decay. Mol. Cell *75*, 1188–1202.e11.

17. Jeong, S. (2017). SR proteins: binders, regulators, and connectors of RNA. Mol. Cells *40*, 1–9.

18. Moon, H., Cho, S., Loh, T.J., Jang, H.N., Liu, Y., Choi, N., Oh, J., Ha, J., Zhou, J., Cho, S., et al. (2017). SRSF2 directly inhibits intron splicing to suppresses cassette exon inclusion. BMB Rep. *50*, 423–428.

19. Moon, H., Jang, H.N., Liu, Y., Choi, N., Oh, J., Ha, J., Zheng, X., and Shen, H. (2019). Activation of cryptic 3′ splice-sites by SRSF2 contributes to cassette exon skipping. Cells *8*, 696.

20. Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M., and Fu, X.D. (2013). Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. Mol. Cell *50*, 223–235.

21. Komeno, Y., Huang, Y.-J., Qiu, J., Lin, L., Xu, Y., Zhou, Y., Chen, L., Monterroza, D.D., Li, H., DeKelver, R.C., et al. (2015). SRSF2 is essential for hematopoiesis, and its myelodysplastic syndrome-related mutations dysregulate alternative Pre-mRNA splicing. Mol. Cell. Biol. *35*, 3071–3082.

22. Bonner, E.A., and Lee, S.C. (2023). Therapeutic targeting of RNA splicing in cancer. Genes (Basel) *14*, 1378.

23. 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.

24. Meggendorfer, M., Roller, A., Haferlach, T., Eder, C., Dicker, F., Grossmann, V., Kohlmann, A., Alpermann, T., Yoshida, K., Ogawa, S., et al. (2012). SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). Blood *120*, 3080–3088.

25. Zhang, J., Lieu, Y.K., Ali, A.M., Penson, A., Reggio, K.S., Rabadan, R., Raza, A., Mukherjee, S., and Manley, J.L. (2015). Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. Proc. Natl. Acad. Sci. USA *112*, E4726–E4734.

26. Kim, E., Ilagan, J.O., Liang, Y., Daubner, G.M., Lee, S.C.W., Ramakrishnan, A., Li, Y., Chung, Y.R., Micol, J.B., Murphy, M.E., et al. (2015). SRSF2 mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. Cancer Cell *27*, 617–630.

27. Yoshimi, A., Lin, K.T., Wiseman, D.H., Rahman, M.A., Pastore, A., Wang, B., Lee, S.C.W., Micol, J.B., Zhang, X.J., de Botton, S., et al. (2019). Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. Nature *574*, 273–277.

28. Pangallo, J., Kiladjian, J.J., Cassinat, B., Renneville, A., Taylor, J., Polaski, J.T., North, K., Abdel-Wahab, O., and Bradley, R.K. (2020). Rare and private spliceosomal gene mutations drive partial, complete, and dual phenocopies of hotspot alterations. Blood *135*, 1032–1043.

29. Daubner, G.M., Cléry, A., Jayne, S., Stevenin, J., and Allain, F.H.T. (2012). A syn-anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. EMBO J. *31*, 162–174.

30. Grimm, J., Jentzsch, M., Bill, M., Backhaus, D., Brauer, D., Küpper, J., Schulz, J., Franke, G.N., Vucinic, V., Niederwieser, D., et al. (2021). Clinical implications of SRSF2 mutations in AML patients undergoing allogeneic stem cell transplantation. Am. J. Hematol. *96*, 1287–1294.

31. Liang, Y., Tebaldi, T., Rejeski, K., Joshi, P., Stefani, G., Taylor, A., Song, Y., Vasic, R., Maziarz, J., Balasubramanian, K., et al. (2018). SRSF2 mutations drive oncogenesis by activating a global program of aberrant alternative splicing in hematopoietic cells. Leukemia *32*, 2659–2671.

32. Smeets, M.F., Tan, S.Y., Xu, J.J., Anande, G., Unnikrishnan, A., Chalk, A.M., Taylor, S.R., Pimanda, J.E., Wall, M., Purton, L.E., et al. (2018). Srsf2P95H initiates myeloid bias and myelodysplastic/ myeloproliferative syndrome from hemopoietic stem cells. Blood *132*, 608–621.

33. Amort, T., Rieder, D., Wille, A., Khokhlova-Cubberley, D., Riml, C., Trixl, L., Jia, X.Y., Micura, R., and Lusser, A. (2017). Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain. Genome Biol. *18*, 1.

34. Franzini, A., Pomicter, A.D., Yan, D., Khorashad, J.S., Tantravahi, S.K., Than, H., Ahmann, J.M., O'Hare, T., and Deininger, M.W. (2019). The transcriptome of CMML monocytes is highly inflammatory and reflects leukemia-specific and age-related alterations. Blood Adv. *3*, 2949–2961.

35. Bamopoulos, S.A., Batcha, A.M.N., Jurinovic, V., Rothenberg-Thurley, M., Janke, H., Ksienzyk, B., Philippou-Massier, J., Graf, A., Krebs, S., Blum, H., et al. (2020). Clinical presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with acute myeloid leukemia. Leukemia *34*, 2621–2634.

36. Tyner, J.W., Tognon, C.E., Bottomly, D., Wilmot, B., Kurtz, S.E., Savage, S.L., Long, N., Schultz, A.R., Traer, E., Abel, M., et al. (2018). Functional genomic landscape of acute myeloid leukaemia. Nature *562*, 526–531.

37. Klairmont, M.M., Carroll, W.L., Aifantis, I., and Park, C.Y. (2022). High ORM1 expression marks a subset of genetically adverse-risk B-ALL characterized by MDSC enrichment, T-cell dysfunction, and inferior overall survival. Blood *140*, 6360.

38. Tillmann, S., Olschok, K., Schröder, S.K., Bütow, M., Baumeister, J., Kalmer, M., Preußger, V., Weinbergerova, B., Kricheldorf, K., Mayer, J., et al. (2021). The unfolded protein response is a major driver of lcn2 expression in bcr–abl-and jak2v617f-positive mpn. Cancers (Basel) *13*, 4210.

39. Walavalkar, N.M., Cramer, J.M., Buchwald, W.A., Scarsdale, J.N., and Williams, D.C. (2014). Solution structure and intramolecular exchange of methyl-cytosine binding domain protein 4 (MBD4) on DNA suggests a mechanism to scan for mCpG/TpG mismatches. Nucleic Acids Res. *42*, 11218–11232.

40. Liu, Y., Olanrewaju, Y.O., Zheng, Y., Hashimoto, H., Blumenthal, R.M., Zhang, X., and Cheng, X. (2014). Structural basis for Klf4 recognition of methylated DNA. Nucleic Acids Res. *42*, 4859–4867.

41. Sajini, A.A., Choudhury, N.R., Wagner, R.E., Bornelöv, S., Selmi, T., Spanos, C., Dietmann, S., Rappsilber, J., Michlewski, G., and Frye, M. (2019). Loss of 5-methylcytosine alters the biogenesis of vault-derived small RNAs to coordinate epidermal differentiation. Nat. Commun. *10*, 2550.

42. Qian, W., Iqbal, K., Grundke-Iqbal, I., Gong, C.X., and Liu, F. (2011). Splicing factor SC35 promotes tau expression through stabilization of its mRNA. FEBS Lett. *585*, 875–880.

43. Li, K., and Wang, Z. (2021). Splicing factor srsf2-centric gene regulation. Int. J. Biol. Sci. *17*, 1708–1715.

44. Zhong, X.Y., Wang, P., Han, J., Rosenfeld, M.G., and Fu, X.D. (2009). SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. Mol. Cell *35*, 1–10.

45. Courtney, D.G., Tsai, K., Bogerd, H.P., Kennedy, E.M., Law, B.A., Emery, A., Swanstrom, R., Holley, C.L., and Cullen, B.R. (2019). Epitranscriptomic addition of m5C to HIV-1 transcripts regulates viral gene expression. Cell Host Microbe *26*, 217–227.e6.

46. Alarcón, C.R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S.S.F., and Tavazoie, S.F. (2015). HNRNPA2B1 is a mediator of m6A-dependent nuclear RNA processing events. Cell *162*, 1299–1308.

47. Zhao, X., Yang, Y.G.Y., Sun, B.F., Shi, Y., Yang, X., Xiao, W., Hao, Y.J., Ping, X.L., Chen, Y.S., Wang, W.J., et al. (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell Res. *24*, 1403–1419.

48. Gu, X., Ma, X., Chen, C., Guan, J., Wang, J., Wu, S., and Zhu, H. (2023). Vital roles of m5C RNA modification in cancer and immune cell biology. Front. Immunol. *14*, 1207371.

49. Delhommeau, F., Dupont, S., Valle, V. Della, James, C., Trannoy, S., Massé, A., Kosmider, O., Le Couedic, J.-P., Robert, F., Alberdi, A., et al. (2009). Mutation in TET2 in myeloid cancers. N. Engl. J. Med. *360*, 2289–2301.

50. Ferrone, C.K., Blydt-Hansen, M., and Rauh, M.J. (2020). Age-associated TET2 mutations: common drivers of myeloid dysfunction, cancer and cardiovascular disease. Int. J. Mol. Sci. *21*, 626.

51. Thol, F., Kade, S., Schlarmann, C., Löffeld, P., Morgan, M., Krauter, J., Wlodarski, M.W., Kölking, B., Wichmann, M., Görlich, K., et al. (2012). Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes. Blood *119*, 3578–3584.

52. Xiao, W., Adhikari, S., Dahal, U., Chen, Y.S., Hao, Y.J., Sun, B.F., Sun, H.Y., Li, A., Ping, X.L., Lai, W.Y., et al. (2016). Nuclear m6A reader YTHDC1 regulates mRNA splicing. Mol. Cell *61*, 507–519.

53. Choe, J., Lin, S., Zhang, W., Liu, Q., Wang, L., Ramirez-Moya, J., Du, P., Kim, W., Tang, S., Sliz, P., et al. (2018). mRNA circularization by METTL3–eIF3h enhances translation and promotes oncogenesis. Nature *561*, 556–560.

54. Karmakar, S., Ramirez, O., Paul, K.V., Gupta, A.K., Kumari, V., Botti, V., de los Mozos, I.R., Neuenkirchen, N., Ross, R.J., Karanicolas, J., et al. (2022). Integrative genome-wide analysis reveals EIF3A as a key downstream regulator of translational repressor protein Musashi 2 (MSI2). NAR Cancer *4*, zcac015.

55. Pronier, E., Imanci, A., Selimoglu-Buet, D., Badaoui, B., Itzykson, R., Roger, T., Jego, C., Naimo, A., Francillette, M., Breckler, M., et al. (2022). Macrophage migration inhibitory factor is overproduced through EGR1 in TET2low resting monocytes. Commun. Biol. *5*, 110.

56. Liu, Y., Luo, M., Jin, Z., Zhao, M., and Qu, H. (2018). DbLGL: an online leukemia gene and literature database for the retrospective comparison of adult and childhood leukemia genetics with literature evidence. Database (Oxford) *2018*, bay062.

57. Anders, S. (2010). Babraham bioinformatics – FastQC a quality control tool for high throughput sequence data. Soil 5. http://www.bioinformatics.babraham.ac.uk/projects/.

58. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10.

59. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

60. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

61. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

62. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

63. Corcoran, D.L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R.L., Keene, J.D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol. *12*, R79.

64. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

65. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

66. Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., and Chang, W. (2022). David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. *50*, W216–W221.

67. Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using David bioinformatics resources. Nat. Protoc. *4*, 44–57.

68. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. USA *111*, E5593–E5601.

69. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., and Gu, J. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics *18*, 80.

70. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

71. Antanaviciute, A., Baquero-Perez, B., Watson, C.M., Harrison, S.M., Lascelles, C., Crinnion, L., Markham, A.F., Bonthron, D.T., Whitehouse, A., and Carr, I.M. (2017). M6aViewer: software for the detection, analysis, and visualization of N6-methyladenosine peaks from m6A-seq/ME-RIP sequencing data. Rna *23*, 1493–1501.

72. Rieder, D., Amort, T., Kugler, E., Lusser, A., and Trajanoski, Z. (2016). MeRanTK: methylated RNA analysis ToolKit. Bioinformatics *32*, 782–785.

73. Moon, Y., Kim, M.H., Kim, H.R., Ahn, J.Y., Huh, J., Huh, J.Y., Han, J.H., Park, J.S., and Cho, S.R. (2018). The 2016 WHO versus 2008 WHO criteria for the diagnosis of chronic myelomonocytic leukemia. Ann. Lab. Med. *38*, 481–483.

74. Cheng, Y.-C., and Prusoff, W.H. (1973). Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem. Pharmacol. *22*, 3099–3108.

75. Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N 6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

76. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44*, D710–D716.

77. Volders, P.J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). Lncipedia 5: Towards a reference set of human long non-coding RNAS. Nucleic Acids Res. *47*, D135–D139.

78. Zhu, S., Xiang, J.F., Chen, T., Chen, L.L., and Yang, L. (2013). Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences. BMC Genomics *14*, 206.

79. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

80. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523.

81. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

82. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

83. Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res. *47*, W556–W560.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal anti-m^{5}C (for RNA dot bot)	Abcam	Cat# ab214727; RRID: AB_2802117
Mouse monoclonal anti-m^{5}C (for RNA m^{5}C MeRIP)	Diagenode	Cat# C15200003
Mouse monoclonal anti-Flag	Sigma-Aldrich	Cat# F3165; RRID: AB_259529
Mouse monoclonal anti-His	Abcam	Cat# ab18184; RRID: AB_444306
Mouse monoclonal anti-Myc	Cell Signaling	Cat# 2276; RRID: AB_2148465
Rabbit polyclonal anti-NSUN2	Proteintech	Cat# 20854-1-AP; RRID: AB_10693629
Mouse monoclonal anti-ACTIN	Sigma-Aldrich	Cat# A5316; RRID: AB_476743
anti-Mouse IgG, HRP-linked secondary antibody	GE Healthcare	Cat# NXA931V; RRID: AB_2721110
anti-Rabbit IgG, HRP-linked secondary antibody	GE Healthcare	Cat# NA934V; RRID: AB_772191
Bacterial and virus strains		
BL21(DE3) Competent E. coli	NEB	Cat# C2530H
Chemicals, peptides, and recombinant proteins		
Acid-Phenol:Chloroform	Thermo Fisher	Cat# AM9722
TURBO™ Dnase	Thermo Fisher	Cat# AM2239
RNasin	Promega	Cat# N251B
4-thiouridine	Sigma-Aldrich	Cat# T4509
Protease inhibitor cocktail	Sigma-Aldrich	Cat# P8340
Protease K	Sigma-Aldrich	Cat# P2308
RNase T1	Fermentas	Cat# EN0542
T4 Polynucleotide Kinase (T4 PNK)	NEB	Cat# M0201L
Adenosine 5′-Triphosphate (ATP)	NEB	Cat# P0756S
Alkaline Phosphatase, Calf Intestinal (CIP)	NEB	Cat# M0290L
SuperScript™ II Reverse Transcriptase	Invitrogen	Cat# 18064014
LightCycler® 480 SYBR Green	Roche	Cat# 4887352001
Chemiluminescent Nucleic Acid Detection Module	Thermo Fisher	Cat# 89880
Dynabeads Protein A beads	Invitrogen	Cat# 10001D
Streptavidin Magnetic Beads	NEB	Cat# S1420S
Anti-FLAG® M2 Magnetic Beads	Millipore	Cat# M8823
Critical commercial assays		
QuikChange™ site-directed mutagenesis kit	Stratagene	Cat# 200518
NEBNext® Multiplex Small RNA Library Prep Set for Illumina	NEB	Cat# E7300S
SMARTer smRNA-seq Kit for Illumina	Takara	Cat# 635030
RNA 3′ end biotinylation kit	Thermo Fisher	Cat# 20160
Deposited data		
Raw and processed high-throughput sequencing data	This paper	GEO: GSE207643
The original imaging data and source dataset deposited in Mendeley Data	This paper	Mendeley Data: https://doi.org/10.17632/zv3fyzh4tr.1
m^{5}C RNA-BisSeq data in HeLa cells	Yang et al.[9]	GEO: GSE93749
SRSF1 and SRSF3 PAR-CLIP-seq data	Xiao et al.[52]	GEO: GSE71096
SRSF3 and SRSF10 RNA-seq data	Xiao et al.[52]	GEO: GSE71095
Polysome profiling sequencing data in HeLa	Choe et al.[53]	GEO: GSE117299
Polysome profiling sequencing data in K562	Karmakar et al.[54]	https://academic.oup.com/narcancer/article/4/2/zcac015/6576546#supplementary-data

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
AML cohort: Bamopoulos et al.	Bamopoulos et al.[35]	GEO: GSE146173
AML cohort: Beat AML	Tyner et al.[36]	http://www.vizome.org/
CMML cohort: Franzini et al.	Franzini et al.[34]	GEO: GSE135902
CMML cohort: Pronier et al.	Pronier et al.[55]	GEO: GSE165305, GSE188624
Leukemia gene and literature (LGL) database	Liu et al.[56]	http://soft.bioinfo-minzhao.org/lgl/
Experimental models: Cell lines		
Human: K562 cells	This paper	N/A
Human: HeLa cells	ATCC	RRID: CVCL_0030
Human: HEK293GP cells	ATCC	RRID: CVCL_E072
Oligonucleotides		
RNA sequences used for biotinylated pull-down assays and NanoBRET assays, see Table S1	This paper	N/A
Primers for RT-qPCR, MeRIP-RT-qPCR, RIP-qPCR, see Table S7	This paper	N/A
siRNA/shRNA sequence, see Table S7	This paper	N/A
Recombinant DNA		
Plasmid: pcDNA3.1-Myc-His-SRSF2	Addgene	Cat# 44721
Plasmid: pcDNA3.1-Myc-His-SRSF2^{P95H}	This paper	N/A
Plasmid: pcDNA3.1-Myc-His-SRSF2^{T51A}	This paper	N/A
Plasmid: pcDNA3.1-Myc-His-SRSF2^{K52A}	This paper	N/A
Plasmid: pcDNA3.1-Myc-His-SRSF2^{H99A}	This paper	N/A
Plasmid: pcDNA3.1-Myc-His-SRSF2^{P107H}	This paper	N/A
Plasmid: pET30a(+)-His-SRSF2	This paper	N/A
Plasmid: pET30a(+)-His-SRSF2-N (1-115)	This paper	N/A
Plasmid: pET30a(+)-His-SRSF2-C (115-221)	This paper	N/A
Plasmid: pCMV-Flag-SRSF2	This paper	N/A
Plasmid: pCMV-Flag-SRSF2^{P95H}	This paper	N/A
Plasmid: pCMV-Myc-SRSF2	This paper	N/A
Plasmid: pCMV-Myc-SRSF2^{P95H}	This paper	N/A
Software and algorithms		
FastQC v0.11.5	Andrews[57]	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Cutadapt v1.9.1	Martin[58]	https://cutadapt.readthedocs.io/en/stable/
Trimmomatic v0.33	Bolger et al.[59]	http://www.usadellab.org/cms/index.php?page=trimmomatic
Bowtie v2.3.4,1	Langmead and Salzberg[60]	http://bowtie-bio.sf.net.
STAR v2.6.1d	Dobin et al.[61]	https://github.com/alexdobin/STAR
Bedtools v2.25.0	Quinlan and Hall[62]	https://bedtools.readthedocs.io/en/latest/
PARalyzer v1.5	Corcoran et al.[63]	https://ohlerlab.mdc-berlin.de/software/PARalyzer_85/
IGV v2.9.4	Thorvaldsdóttir et al.[64]	https://software.broadinstitute.org/software/igv/
MEME (Web-based)	Bailey et al.[65]	http://meme-suite.org/tools/meme
DAVID v2021q4 (Web-based)	Sherman et al.[66] and Huang et al.[67]	https://david.ncifcrf.gov
rMARTs v4.1.2	Shen et al.[68]	https://github.com/Xinglab/rmats-turbo
rmats2sashimiplot v2.0.4	Xing Lab	https://github.com/Xinglab/rmats2sashimiplot
Python v2.7	Python Software Foundation	https://www.python.org
R v4.0.4	The R Foundation	https://www.r-project.org
GraphPad Prism 9	GraphPad Software, Inc.	https://www.graphpad.com/scientific-software/prism/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
AfterQC v0.9.6	Chen et al.[69]	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1469-3
HTSeq count v0.9.1	Anders et al.[70]	https://academic.oup.com/bioinformatics/article/31/2/166/2366196
m6aViewer v1.6.1	Antanaviciute et al.[71]	https://pubmed.ncbi.nlm.nih.gov/28724534/
meRanTK v1.2.1b	Rieder et al.[72]	https://icbi.i-med.ac.at/software/meRanTK/
Biorender	Biorender	https://biorender.com
Other		
NanoBRET assay	Promega	https://www.promega.com
Mass spectrometry	Promega	https://www.promega.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, François Fuks (francois.fuks@ulb.be).

Materials availability

All unique reagents including plasmids generated in this study are available from the lead contact without any restrictions for academic research purposes.

Data and code availability

- PAR-CLIP-seq, RNA-seq, m^5C MeRIP-seq data in cell lines and RNA-BisSeq data in CMML patients supporting the findings of this study have been deposited at Gene Expression Omnibus (GEO) database under accession number GEO: GSE207643 and are publicly available as of the date of publication. The unprocessed western blot images and source dataset have been deposited in Mendeley Data (https://doi.org/10.17632/zv3fyzh4tr.1). This paper also analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- This paper does not report original code. A detailed description of the use of publicly available programs is mentioned in the methods, and also listed in 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 STUDY PARTICIPANT DETAILS

Cell lines

HeLa, K562, and HEK293GP cell lines were originally purchased from ATCC. The K562 $SRSF2^{P95H/WT}$ knockin cell line (engineered to express $SRSF2^{P95H}$ from an endogenous locus) were from Horizon Discovery Inc. All cells were authenticated by short tandem repeat (STR) analysis and routinely checked for mycoplasma contamination. HeLa and HEK293GP cells were maintained in DMEM (Gibco) supplemented with 10% fetal bovine serum (FBS, Gibco) and 1% penicillin and streptomycin (Pen Strep, Gibco). K562 cells were cultured in IMDM (Gibco) supplemented with 10% FBS and 1% Pen Strep. All cells were cultured at 37 °C in a humidified atmosphere containing 5% CO_2.

Human specimens

Peripheral blood samples were collected from 8 CMML patients with informed consent in compliance with guidelines of the ethics committee Ile-de-France (MYELOMONO cohort, DC-2014-2091). Patients with CMML were diagnosed according to the 2016 WHO criteria[73] and their clinical-biological characteristics are summarized in Table S6. Peripheral blood mononuclear cells (PBMC) were sorted out by density centrifugation Pancoll (Pan Biotech) and CD14+ monocytes were isolated by negative selection with magnetic beads and the AutoMacs system (Miltenyi Biotech).

METHOD DETAILS

RNA interference and transfection

For transfection with small interfering RNA (siRNA), HeLa cells were cultured to 50%–60% confluency. The cells were then transfected by electroporation with control siRNA (universal negative control) or siRNA for NSUN2 or SRSF2 (See Table S7 for siRNA

sequences), using the LONZA Kit (VCA-1001, Lonza, Germany) according to the manufacturer's instructions. After 48 h, the transfected cells were washed in PBS and RNA or protein was isolated.

Stable NSUN2 knockdown K562 cell line was generated by inserting the target sequence for NSUN2 or the scramble control into the pSUPER.retro.puro (pRS) vector (OligoEngine, VEC-PRT-0002) to form short hairpin RNAs for RNA interference. To produce the lentivirus, HEK293GP cells were grown to 40%-50% confluency and transfected with 5 μg pRS plasmid and 1 μg plasmid encoding the glycoprotein of vesicular stomatitis virus (VSV-G, BD Biosciences Clontech) using polyethylenimine (PEI). The transfection mixture was replaced with fresh growth medium after 5 h. 48 h post-transfection, viral supernatants were harvested, sterile filtered, mixed with 8 μg/ml polybrene and incubated with target K562 cells. After 48 h, infected cells were selected with 2 μg/ml puromycin.

For plasmid transient transfection in HeLa cells, cells were grown to 80% confluency and then transfected with plasmids and lipofectamine 2000 (Thermo Fisher) at a ratio of 1:3 (m/v) according to the manufacturer's protocol. Cells were collected 48 h after transfection.

For plasmid transient transfection in K562 cells, cells were suspended in IMDM medium without FBS or antibiotics at a concentration of 10^7 cells/ml. A volume of 0.3 ml was transferred to a sterile electroporation cuvette (Bio-Rad Gene Pulser cuvette, 0.4 cm) and kept at room temperature for 15 min in the presence of 50 μg plasmid. Electroporation was performed using the Gene Pulser Xcell System (Bio-Rad) with 875 V/cm, 500 μF capacitance, and infinite resistance. After receiving the electric pulse, cells were transferred to culture flasks and incubated with complete IMDM medium for 48 h before harvesting.

Expression plasmids and site-directed mutagenesis

We obtained Myc-tagged SRSF2 full-length pcDNA3.1 plasmid from the Addgene plasmid repository (cat #44721). The mutant plasmids were generated by introducing point mutations (either P95H, T51A, K52A, H99A or P107H) into wild-type SRSF2 plasmids using the QuickChange Site-directed Mutagenesis kit (Stratagene) according to the manufacturer's instructions. SRSF2 full length and fragments (amino acids 1-115 and 115-221) were amplified by PCR and subcloned into the pET30a vector (Addgene). All plasmids were verified by Sanger sequencing and prepared with the Qiagen Plasmid Plus Midi Kit. All the primers used for plasmid cloning are listed in Table S7.

Histidine-tagged protein purification

BL21 competent E. coli were transformed with His-tagged SRSF2 plasmids and grown overnight at 37 °C in 50 ml of LB culture medium containing kanamycin. One hour before induction with isopropyl β-D-1-thiogalactopyranoside (IPTG), the cell suspension was diluted to 400 ml. The production step was carried out at 16 °C for 20 h. Cells were then pelleted and resuspended in lysis buffer (TBS-Triton supplemented with 10 mM imidazole (Sigma-Aldrich) and antiprotease cocktail (Promega)). After sonication, the supernatant was clarified by centrifugation (5000 rpm, 10 min at 4 °C) and incubated with 400 μl of nickel-nitrilotriacetic acid (Ni-NTA) agarose beads (Qiagen) on a rotating wheel for 2 h at 4 °C. Beads were then spun down, washed with lysis buffer, and eluted with 1-3 ml of TBS 400 mM Imidazole. The eluted protein was concentrated using Amicon centrifugal filters (EMD Millipore, Billerica, MA) with a molecular weight cut-off of 3 kDa. Protein purity was confirmed by Coomassie staining and western blotting with anti-His antibody (Abcam #18814).

FP-based binding assay

For FP-based binding assays, we expressed the RRM domain of SRSF2 from a plasmid pET-26b(+) capable of encoding histidine tagged SRSF2 RRM domain (amino acids 1-101). This plasmid was a kind gift from James Manley (Columbia University). The P95H mutation was introduced by site-directed mutagenesis. Proteins were expressed in E. coli and purified by successive passage of filtered lysates form Ni-NTA affinity and size-exclusion chromatography columns. The final proteins in buffer containing 20 mM HEPES pH 7.5, 0.2 M NaCl were used for subsequent binding experiments. FP-based binding assays were carried out in a buffer containing 0.01 M HEPES pH 7.5 and 0.05 M KCl. A constant 5 nM concentration of the fluorescein-labeled oligo was used with increasing concentrations of SRSF2 RRM (WT or P95H) proteins in a 384-well plate. Significant changes observed in FP upon increasing protein concentrations were indicative of direct binding. The FP (emission wavelength = 530 nm, excitation wavelength = 485 nm) value for each dilution was measured using PHERAstar FS (BMG Labtech). The buffer corrected values were used to calculate the equilibrium dissociation constant (K_d) using a simple 1:1 specific binding model. Data were fitted in GraphPad Prism (GraphPad Software, San Diego, CA).

Biotinylated RNA pull-down assay

Biotin-labeled RNA oligos were obtained from Integrated DNA Technologies (IDT) (oligo sequences were listed in Table S1). For detection of endogenous SRSF2, 1×10^7 cells were used per condition (no probe, A, m^6A, C, m^5C or hm^5C), and for detection of overexpressed Myc-tagged protein, 5×10^6 cells were used per condition. Cells were lysed by rotating at 4 °C for 30 min in 500 μl lysis buffer (10 mM NaCl, 2 mM EDTA, 0.5% Triton X-100, 0.5 mM DTT, 10 mM Tris-HCl pH 7.5, 1 × protease inhibitor cocktail, 40 U/ml RNase inhibitor) and centrifuged at 15,000 g for 15 min. Total cell extracts were then supplemented with 500 μl of binding buffer (150 mM KCl, 1.5 mM $MgCl_2$, 0.05% NP-40, 0.5 mM DTT, 10 mM Tris-HCl pH 7.5) and pre-cleared with 20 μl of streptavidin-conjugated magnetic beads (NEB) for 1 h at 4 °C. The beads were removed, and the supernatant was collected. 5% of the pre-cleared cell lysate was saved as input and the rest was incubated with 2 μg of RNA probes for 30 min at room temperature and

then for 1.5 h at 4 °C on a wheel. Meanwhile, 50 μl of beads were blocked in binding buffer containing 5 μg/ml yeast tRNA and 1% BSA for 1.5 h at 4 °C. The pull-down mixture was then incubated with pre-blocked beads for 1 h at 4 °C with rotation. After washing three times with ice-cold binding buffer, the RNA-protein-bead mixture was heated in 1 × NuPAGE LDS sample buffer (Invitrogen) at 95 °C for 5 min. For western blot analysis, the eluted RNA-protein complexes were separated on 10% polyacrylamide gels and immunoblotted with antibodies.

For mass spectrometry, the beads were dried after washing steps and shipped on dry ice to Promega (Madison, Wisconsin, United States) for further processing. Briefly, captured proteins were separated on SDS-PAGE gel and stained with Coomassie brilliant blue. The protein-containing gel slices were digested with trypsin on an automated ProGest Protein Digestion Station (Digilab, Marlborough, MA). Gel digests were analyzed directly by nano LC-MS/MS with a NanoAcquity HPLC (Waters) interfaced with an Orbitrap Velos Pro (Thermo Fisher) tandem mass spectrometer. The data were searched against the Mascot database (Matrix Science) and filtered by Scaffold software (Proteome Software). To avoid false positive, a protein was considered identified only if at least two unique peptides from this protein were identified. The volcano plot was based on average counts of peptides detected by mass spectrometry at least twice in three independent experiments. Statistical significance ($-\log_{10}$(p-value); y-axis) was plotted against fold change ($\log_2$(oligo-m^{5}C/C); x-axis). Only p values < 0.05 and $|\text{fold change}| \geq 2$ were considered significant changes in binding.

In vitro RNA pull-down assay and bioanalyzer analysis

1 mg of recombinant protein and 2 μg of RNA probes with or without m^{5}C (5′-UUU CAG CUC (C/m^{5}C)GG UCA CGC UC-biotin-3′) were incubated with 15 μl streptavidin-conjugated magnetic beads (NEB) in 1 ml binding buffer (50 mM Tris-HCl pH 7.5, 250 mM NaCl, 0.4 mM EDTA, 0.1% NP-40, 1 mM DTT, 40 U/ml RNase inhibitor) for 1 h at 4 °C with rotation. After washing three times with ice-cold binding buffer, the protein-RNA-bead mixture was subjected to western-blot or bioanalyzer analysis. For western-blot analysis, the mixture was heated in 1 × NuPAGE LDS sample buffer (Invitrogen) at 95 °C for 5 min and the eluted RNA-protein complexes were separated on 10% polyacrylamide gels and immunoblotted with anti-His antibody. For bioanalyzer analysis, the mixture was incubated with 400 μl Proteinase K solution (4 mg/ml) for 1 h at 55°C with rotation at 1000 rpm/min on a Thermoblock. The supernatant was then collected and subjected to RNA extraction with phenol:chloroform:iso-amyl alcohol (125:24:1, pH 4.5, Invitrogen). Finally, the purified pull-down RNA probes and 1 μg of each input probe were analyzed by bioanalyzer, using the Agilent small RNA Kit to check the stability of these RNA probes.

NanoBRET assay

FuGENE HD (Promega) was used according to the manufacturer's protocol to transfect HEK293 cells with plasmid DNA containing Nluc-SRSF2 fusion constructs. Briefly, Nluc-target fusion constructs were diluted in Transfection Carrier DNA (Promega) at a mass ratio of 1:10, after which FuGENE HD was added at a ratio of 1:3 (μg DNA: μl FuGENE HD). One vol transfection mixture was combined with 20 vol HEK293 cell suspension (density: 2 × 10^5 cells/ml) and then incubated for 20 h. Following transfection, the cells were trypsinized and resuspended in Opti-MEM containing a 1:1000 dilution of RNasin (Promega). This mixture was then dispensed at 28 μl/well into white 384-well plates (Corning) (cell density: 5.6 × 10^3 cells/well). Serial dilutions of unlabeled oligo-C (IDT) or oligo-m^{5}C (called "cold" RNA, see sequences in Table S1) were prepared at 20 × working concentration in Opti-MEM. So were the fluorescently labeled "tracer" RNAs, identical in sequence to the C and m^{5}C oligomers but additionally labeled in 5′ with Alexa594 dye. Cold RNAs and tracer RNAs contained the same sequence as the probes used in the biotin pull-down experiments unless specified. To permeabilize the cells, 4 μl of 20 × digitonin was added to the plate (final concentration: 50 μg/ml). Four microliters each of prepared serial dilutions of cold RNA and tracer RNA were then added to the plate. Background control wells received no tracer RNA. Forty microliters of 2 x NanoBRET™ Nano-Glo® Substrate was then added to each well and the plate was briefly mixed using vibrational mixing. NanoBRET measurements were immediately collected on a GloMax Discover luminometer equipped with a 450-nm BP filter (donor) and a 600-nm LP filter (acceptor) using a 0.3-s integration time. Background-subtracted BRET ratios were calculated by first dividing the acceptor signal by the donor signal and then subtracting the BRET ratio of background control wells lacking tracer RNA. BRET ratios were then expressed in milli-BRET units (mBU) by multiplying the background-corrected ratios by 1000. The IC_{50} values were determined using a four-parameter dose-response curve fit in Prism 9 (GraphPad Software, Inc., La Jolla, CA). Linearized Cheng–Prusoff analysis[74] yielded a linear plot with a y-intercept equal to the apparent dissociation constant ($K_{i,app}$).

m^{5}C dot blotting

Total RNA was extracted from control and siNSUN2 HeLa cells with the RNeasy kit (Qiagen) and treated with the RNase-Free DNase Set (Qiagen) to remove the residual DNA. Enrichment of mRNA from total RNA was performed using GenElute™ mRNA Miniprep Kit (Sigma-Aldrich). The mRNAs were heat-denatured for 2 min at 70 °C, cooled on ice for 2 min and then spotted on a nylon membrane (GE Healthcare Hybond-N+) in an assembled Bio-Dot apparatus (Bio-Rad) according to the manufacturer's instructions. The membrane was dried and subsequently cross-linked twice with 200,000 μJ/cm^2 UV. It was then blocked in 5% bovine serum albumin (BSA) in PBST (PBS + 0.1% Tween-20) for 1 h at room temperature and incubated with an anti-m^{5}C monoclonal antibody (diluted 1:500, Abcam #ab214727) overnight at 4 °C. Thereafter, the membrane was washed three times with PBST for a total of 30 min and incubated with an HRP-linked anti-rabbit IgG secondary antibody (diluted 1:5000, GE Healthcare #NA934V) for 1 h at room

temperature, washed three times with PBST, and developed with the Western Lightning Plus-ECL (Perkin-Elmer) or SuperSignal West Femto Chemiluminescent Substrate (Thermo Fisher) according to the manufacturer's protocols. To ensure equal loading of RNA on the membrane, the same membrane was rinsed with PBST for 10 min and stained with methylene blue staining buffer (0.02% methylene blue in 0.4 M sodium acetate and 0.4 M acetic acid).

Mass spectrometry analysis of m^5C

Total RNA was extracted from HeLa or K562 cells with the RNeasy kit (Qiagen) and treated with the RNase-Free DNase Set (Qiagen). Two rounds of mRNA enrichment were performed with the RNeasy Pure mRNA Bead Kit (Qiagen) to ensure no contamination from other RNA species. For detection of m^5C, 500 ng of mRNA per sample was sent to Tamaserv (Germany) for liquid chromatography coupled to mass spectrometry (LC-MS) analysis of methylated nucleotides.

Reverse-transcriptase quantitative PCR

Total RNA was purified with the RNeasy kit (Qiagen) and treated with the RNase-Free DNase Set (Qiagen) to remove the residual DNA. One μg of DNase-free RNA was reverse transcribed using the SuperScript II Reverse Transcriptase and oligo (dT) primers (Invitrogen). qPCR was performed for each cDNA (25 ng) sample in triplicate using the LightCycler 480 Probes Master Kit (Roche). The housekeeping genes GAPDH and β-ACTIN were used as the internal reference genes. The fold change in expression of the target gene relative to the reference genes was assessed. The RT-qPCR data were presented as the fold-change in gene expression normalized to the reference genes and relative to the control. The sequences of all primers used in this study are listed in Table S7.

RNA immunoprecipitation-qPCR

The RIP experimental procedure was adapted from the previously reported method.[75] Briefly, Flag-SRSF2 overexpressing HeLa cells were lysed by rotating at 4 °C for 30 min in 2 vol lysis buffer (150 mM KCl, 10 mM HEPES pH 7.6, 2 mM EDTA, 0.5% NP-40, 0.5 mM DTT, 1 × protease inhibitor cocktail, 40 U/ml RNase inhibitor) and centrifuged at 15,000 g for 15 min. The supernatant was collected and divided into 2 aliquots, of which 1/10 was used as input and 9/10 for immunoprecipitation. The cell lysate was incubated with anti-Flag M2 magnetic beads (Sigma-Aldrich, 10 μl per mg lysate) at 4 °C for 4 h in 2 vol NT2 buffer (200 mM NaCl, 50 mM HEPES pH 7.6, 2 mM EDTA, 0.05% NP-40, 0.5 mM DTT, 40 U/ml RNase inhibitor) with rotation. After washing eight times with 1 ml ice-cold NT2 buffer, the protein-RNA-bead mixture was incubated with 400 μl Proteinase K solution (4 mg/ml) for 1 h at 55 °C with rotation at 1000 rpm/min on a thermoblock. The supernatant was then collected and subjected to RNA extraction with phenol:chloroform:iso-amyl alcohol (125:24:1, pH 4.5, Invitrogen). The input RNA was extracted from the input cell lysate in the same way as IPed RNA using Phenol/Chloroform-based method. Equal amounts of input and IPed RNAs were subjected to reverse transcription and downstream qPCR analysis (primers listed in Table S7). The relative binding enrichment of bound RNAs in IP was normalized to input. The p values were determined using an unpaired two-tailed Student's t test.

RNA-seq

Total RNA was purified with the RNeasy kit (Qiagen) and treated with the RNase-Free DNase Set (Qiagen) to remove the residual DNA. Total RNA samples from HeLa cells was then subjected to rRNA depletion using Ribominus Human/Mouse Transcriptome Isolation Kit (Invitrogen). Enrichment of mRNA from total RNA sample in K562 cells was performed using GenElute™ mRNA Miniprep Kit (Sigma-Aldrich). The RNA-seq library preparation was performed using the KAPA Stranded mRNA-seq kit according to the manufacturer's instructions. High-throughput sequencing was performed on Illumina HiSeq2500 sequencing system (RNA-seq in HeLa cells) or Illumina NextSeq500 system (RNA-seq in K562 cells).

PAR-CLIP

We followed previously reported procedures.[9] Briefly, HeLa cells were co-transfected by electroporation with siRNA and the pCMV-Flag-SRSF2 plasmid. Control and NSUN2-knockdown K562 cells were co-transfected with the pCMV-Flag-SRSF2 and pCMV-Myc-SRSF2 plasmids. $SRSF2^{P95H}$ mutant K562 cells were co-transfected with the pCMV-Flag-$SRSF2^{P95H}$ and pCMV-Myc-$SRSF2^{P95H}$ plasmids. Transfected cells were cultured in a medium supplemented with 200 μM 4-thiouridine (4-SU) (Sigma-Aldrich) for 14 h and then irradiated once with 400 mJ/cm^2 at 365 nm. The cells were then lysed, digested with 1 U/μl RNase T1 at 22 °C for 8 min, and immunoprecipitated with anti-Flag M2 magnetic beads (Sigma-Aldrich). The protein-RNA-bead complex was digested with 10 U/μl RNase T1 again at 22 °C for 8 min and incubated with 0.5 U/μl Alkaline Phosphatase Calf Intestinal (CIP, NEB) for 10 min at 37 °C. The beads were washed and then incubated with 0.5 U/μl T4 Polynucleotide Kinase (NEB) for 15 min at 37 °C. After washing, one-sixth of the beads were resuspended in 1 × NuPAGE LDS sample buffer (Invitrogen), boiled at 95 °C for 10 min, and the mixture was resolved by SDS-PAGE to detect the immunoprecipitation efficiency. One-sixth of the beads were labeled with biotin using the RNA 3′ End Biotinylation kit (Thermo Fisher) and visualized with the Chemiluminescent Nucleic Acid Detection Module kit (Thermo Fisher) following the manufacturer's instructions. The rest of the beads were also boiled and the mixture electrophoresed through a NuPAGE Bis-Tris protein gel. Parts containing target protein-RNA complexes were cut from the gel according to the protein-RNA-biotin signal. The protein-bound RNA in the gel pieces was recovered by D-Tube™ Dialyzer Midi (Merck-Millipore), digested with proteinase K (Roche), and extracted with phenol-chloroform. The purified RNA was used for library construction with

the Takara SMARTer smRNA-seq kit (PAR-CLIP in HeLa cells) or the NEBNext® Multiplex Small RNA Library Prep kit (PAR-CLIP in K562 cells) for Illumina, following the manufacturer's instructions.

m^{5}C MeRIP-RT-qPCR

To evaluate the specificity and efficiency of the m^{5}C antibody (Diagenode #C15200003), we performed MeRIP-RT-qPCR. Briefly, we first obtained the unmethylated, m^{5}C-, or hm^{5}C-methylated Renilla luciferase RNA transcripts by *in vitro* transcription. The MEGAscript T7 transcription kit (Invitrogen) was used for *in vitro* transcription according to the manufacturer's instructions. For transcripts containing m^{5}C or hm^{5}C, CTP nucleotides were replaced with m^{5}CTP or hm^{5}CTP (TriLink Biotechnologies) during the reaction. Next, 2.38 μg of each RNA transcript was immunoprecipitated and purified as described in the "m^{5}C MeRIP-seq" section. Reverse transcription of immunoprecipitated and input RNAs was carried out using SuperScript II Reverse Transcriptase (Invitrogen) with primers specific for Renilla luciferase (see Table S7). LightCycler 480 Probes Master Kit (Roche) was employed for qPCR. IP-versus-input enrichment in transcripts was determined by the percentage of input method.

m^{5}C MeRIP-seq

Total RNA was first extracted from siCtrl and siNSUN2 HeLa cells with the RNeasy kit (Qiagen) and then 1000 μg DNA-free total RNA was subjected to mRNA enrichment through oligo-dT selection with the GenElute™ mRNA Miniprep Kit (Sigma). The obtained mRNA was fragmented into 200-300 nucleotide-long fragments in RNA fragmentation buffer at 94°C for 18s. Fragmented RNA was then precipitated with ethanol and resuspended in RNase-free water. The amount and size of the fragmented RNA were tested with an Invitrogen Qubit Fluorometer and an Agilent 2100 Bioanalyzer and, respectively, with the Qubit RNA Assay kit (Thermo Fisher) and the RNA 6000 Nano kit (Agilent). Then 10-50 ng fragmented RNA was stored at -80°C to serve as input. The remaining RNA was first denatured at 70°C for 5 min and then incubated overnight at 4°C with 0.5 mg/ml anti-m^{5}C monoclonal antibody (Diagenode #C15200003) on a rotating wheel in IP buffer (50 mM Tris-HCl, 750 mM NaCl, 0.5% Igepal CA-630, RNasin 400 U/ml and ribonucleoside vanadyl complex 2 mM) supplemented with protease inhibitors (cOmplete, Mini, EDTA-free, Roche). The next day, 50 μl Dynabeads™ Protein G (Invitrogen) were washed three times with IP buffer and blocked by incubating with 0.5 mg/ml BSA in IP buffer for 1 h on a rotating wheel. The blocked beads were washed twice, then added to the IP mix and incubated for 2 h at 4°C with gentle rotation. After extensive washing with IP buffer, bound RNA was purified with TriPure Isolation Reagent (Roche) and resuspended in RNase-free water. cDNA libraries were constructed with the SMARTer® Stranded Total RNA-seq kit v2 - Pico Input Mammalian (Takara) for the input and IP samples. Sequencing was performed on the Illumina NextSeq500 platform.

RNA-BisSeq of CMML patients

Total RNA from CMML monocytes was extracted with the RNA Purification Plus Kit (Norgen Biotek). All total RNA samples had RIN (RNA integrity number) values > 7. The Ribo-off rRNA Depletion Kit (Vazyme) was used to remove ribosomal RNA. About 500 ng ribo-depleted RNA was mixed with 1.5 ng *in vitro* transcribed firefly luciferase spike-in RNA and cut into fragments approximately 150 nucleotides in length with the RNA fragmentation reagent (Ambion). Bisulfite treatment was performed with the EZ RNA methylation Kit (Zymo Research), with some modifications. Briefly, fragmented RNA was converted by means of two cycles of 5 min at 70°C followed by 45 min at 64°C. RNA desulfonation and purification were also performed with this kit. RNA quantity was determined with Qubit. cDNA libraries were constructed with the KAPA Stranded mRNA-seq Kit (KK8421) according to the manufacturer's instructions. High-throughput sequencing was performed on the Illumina NovaSeq 6000 platform.

RNA-seq and m^{5}C MeRIP-seq preprocessing

Sequencing data from RNA-seq and m^{5}C MeRIP-seq were pre-processed as follows. First, the raw sequencing data were analysed with FastQC.[57] Low-complexity reads were removed with the AfterQC tool using default parameters.[69] To exclude reads originating from rRNA or tRNA, the reads were mapped to human tRNA and rRNA sequences with Bowtie2.[60] The rRNA and tRNA sequences were downloaded from https://www.ncbi.nlm.nih.gov/nuccore using "Homo sapiens [Organism] AND (biomol_rrna [PROP] OR biomol_trna [PROP])" as search parameters. Reads that did not map to tRNA or rRNA sequences were further processed with Trimmomatic using default parameters to remove adapter sequences.[59] The resulting fastq data were again analysed with FastQC to ensure that no further processing was needed. The clean reads were aligned with the hg19 genome, with the STAR algorithm[61] using the reference transcriptome based on Ensembl v85[76] and LNCipedia v5.2[77] (hereafter referred to as Ensembl + LNCipedia).

PAR-CLIP, m^{5}C MeRIP-seq, and RNA-BisSeq annotation

The sites identified by PAR-CLIP and m^{5}C MeRIP-seq were annotated with the Ensembl + LNCipedia reference transcriptome. The percentage of binding sites mapping to mRNA, lncRNA, sncRNA, pseudogenes, and others were plotted with GraphPad Prism 9. Sites were assigned to one or several transcripts and to annotated structural elements: to an exon when the peak summit was inside an annotated exon, to an intron when the peak summit was outside the exon but inside the transcript, and counted as intergenic when the peak could not be associated with a coding gene. The same rules were used to categorize peaks according to their association with coding sequences (CDS) or flanking regions (5′ UTR and 3′ UTR).

RNA-seq data analysis

RNA sequencing data for SRSF3 and SRSF10 knockdown and corresponding controls were downloaded from the GEO database under accession number GEO: GSE71095.[52] Published data and in-house paired-end RNA-seq data were processed in the same way, described in the "RNA-seq and m^5C MeRIP-seq preprocessing" section. Then read count was computed with the HTSeq tool[70] and converted to Transcripts Per Million (TPM). Heatmap was plotted using the R package pheatmap.

rMATS (replicate Multivariate Analysis of Transcript Splicing)[68] was applied to analyze 5 different types of alternative splicing events, namely skipped exon (SE), alternative 5′ splice site (A5SS), alternative 3′ splice site (A3SS), mutually exclusive exons (MXE) and retained intron (RI). The differences in the exon inclusion level (delta "percent spliced in"; ΔPSI) between knockdown and control samples were used as a measure of modulations in alternative splicing events upon depletion of each of these genes. Differential splicing events with FDR < 0.1 and ΔPSI >10% were considered significant. The correlations of differential splicing events were calculated using Spearman's correlation analysis. The Spearman's correlation coefficients (r_s) and p values were calculated using the "cor.test" function in the statistical language R. Representative splicing events were represented with rmats2sashimiplot (https://github.com/Xinglab/rmats2sashimiplot). A "bedtools merged"-based in-house script was used to identify overlaps of differentially spliced genes and SRSF2 binding targets.

m^5C MeRIP-seq analysis

Gene expression was evaluated on the basis of HTSeq counts for input samples.[70] m^5C sites were identified from IP samples with the m6aViewer peak-calling tool,[71] using the input to estimate background noise. Reported m^5C sites are the ones showing significant enrichment over input in all siCtrl replicates, present in genes with an expression level of at least 1 TPM and having a sufficient coverage of input (more than 20). Differential sites were defined as sites showing differential p values smaller than 0.05 and absolute fold-changes higher than 1.5 consistently for all replicates. For visual representations of local enrichment profiles, HPB normalized coverage profiles were generated with the bamTobw tool (https://github.com/YangLab/bamTobw)[78] and uploaded into the IGV tool.[64] The sites were annotated according to the "PAR-CLIP, m^5C MeRIP-seq and RNA-BisSeq annotation" section.

PAR-CLIP-seq data analysis

The raw sequencing data were first analyzed using FastQC.[57] Subsequently, reads were stripped of adaptor sequences using cutadapt[58] with parameters: cutadapt -a AGATCGGAAGAG, and cutadapt -m 15 -u 4 -a AAAAAAAAAA (only for libraries prepared by the SMARTer smRNA-seq kit), and then low-quality bases were removed with Trimmomatic.[59] Processed reads exceeding 15 nt in length were defined as clean reads. The resulting fastq data were again analyzed using FastQC to ensure no further processing was needed. Bowtie[79] was applied to map clean reads against the hg19 genome, with up to two mismatches allowed. PARalyzer software[63] was used to define the cluster of SRSF2 binding sites with default parameters. The results were further filtered by ReadCount $\geq$ 40. A "bedtools merged"-based in-house script was used to identify binding sites observed in all replicate experiments and only the binding sites common to two replicates were used for downstream analysis. The binding sites were annotated according to the "PAR-CLIP and m^5C MeRIP-seq annotation" section.

To perform motif analysis, intersectBed[62] was used to associate SRSF2 peaks with transcripts in the RefSeq transcriptome. The strand of each peak was attributed to its associated transcript. The peaks were then extended to 250 bp on both sides of the center. The corresponding sequence of each extended peak was extracted with "bedtools getfasta"[62] in a stranded way. The SRSF2 binding motifs were analyzed with both the "Centrimo" and the "DREME" tool (http://meme-suite.org/).[65] The DREME search window was set between 5 and 8.

To obtain visual representations of local enrichment profiles, the coverage profiles were HPB (Hits Per Billion-mapped-bases) normalized with the bamTobw tool (https://github.com/YangLab/bamTobw)[78] and then visualized with the Integrative Genomics Viewer.[64]

Functional enrichment analysis

Gene functional annotation enrichment and pathway analysis were performed with DAVID online tool(DAVID, https://david.ncifcrf.gov)[66,67] or Metascape.[80] When analyzed with the David tool, only the GO terms for biological process categories and KEGG pathways are shown.

Overlaps between SRSF1-3 binding sites

The SRSF1- and SRSF3-binding sites identified by PAR-CLIP-seq in HeLa cells was downloaded from the GEO database under accession number GEO: GSE71096.[52] The downloaded binding sites were further filtered by ReadCount $\geq$ 40. A "bedtools merged"-based in-house script was used to identify binding sites observed in all replicate experiments, and only the binding sites common to two replicates were used for downstream analysis. Binding sites of the SR proteins were then overlapped using the same script.

Similarities between SR protein motifs and RNA probe

To check the similarity of the SR protein binding motifs to the RNA probe used in biotin pull-down experiments, the classic Needleman-Wunsch algorithm was applied (https://www.bioinformatics'.org/sms2/pairwise_align_dna.html)[81] with default

parameters except for internal gaps set at -6 . The percentage of motifs aligning with the RNA probe sequence was presented as a bar plot.

RNA-BisSeq analysis of the CMML cohort

Adaptors and low-quality bases in the raw sequencing reads were removed with Cutadapt[58] and Trimmomatic,[59] respectively. Clean reads with lengths greater than 18 nt were mapped to the bisulfite-converted rRNA and tRNA sequences described in ''RNA-seq and m^5C MeRIP-seq preprocessing,'' using meRanGh from meRanTK[72] with parameters: -fmo -mmr 0.01. The unmapped reads were mapped against hg19 bisulfite-converted genome using the same tool and parameters. Only samples with more than 99.5% C-to-T conversion rate in both CMML RNAs and luciferase spike-in RNAs were used for further analysis. The m^5C sites were called with meRanCall from meRanTK with the following parameters: -mBQ 20 -mr 0 -fdr 0.05. The high-confidence m^5C sites with a coverage depth $\geq$ 30, methylation level $\geq$ 0.1, and methylated cytosine depth $\geq$ 5 were associated with transcripts by means of intersectBed from bedtools.[62] They were associated with transcript regions as described in the ''PAR-CLIP, m^5C MeRIP-seq and RNA-BisSeq annotation'' section. For the mRNA transcripts of each sample, m^5C levels were then averaged or set at 0 when no site was present. Transcripts where m^5C was absent in more than half of the samples were excluded. NSUN2 levels were evaluated with HTSeq counts in meRanGh mapped reads and used to classify patients as ''NSUN2-high'' or ''-low'', with the median as cut-off. m^5C levels were averaged for transcripts in each category and the m^5C levels of the transcripts methylated in NSUN2-high patients (average $m^5C > 0.1$) were assessed in both categories. For gene set enrichment analysis (GSEA),[82] a t-test was first computed for each gene between the ''NSUN2-high'' and ''NSUN2-low'' categories. The p value was then converted into significance ($-\log_{10}$(p value)) and multiplied by –1 if the m^5C level was lower in the NSUN2-low group. Transcripts ranked according to this score were then submitted to GSEA against the Hallmark dataset.

Integrated analysis of SRSF2 targets and m^5C sites

The m^5C sites identified by RNA-BisSeq in HeLa mRNA under accession number GEO: GSE93751 (platform Illumina HiSeq 2500)[9] were downloaded from the GEO database. To seek evidence of SRSF2 binding to m^5C at the transcriptome-wide level, the distribution of SRSF2 binding sites around the m^5C sites from published RNA-BisSeq and in-house m^5C MeRIP-seq data were computed independently. The regions covering 3 kb upstream and downstream of each m^5C site, transcript-wise (i.e., introns excluded), were divided into 100 bins and the SRSF2-binding sites located within 3 kb of an m^5C site were identified with bedtools intersect.[62] Then the count of SRSF2-binding sites was computed for each bin. As a control for the binding sites, the same analysis was performed with positions randomly selected along transcripts (keeping only the longest isoform of a gene, introns included). Moreover, SRSF2-binding transcripts identified in PAR-CLIP experiments were intersected with m^5C-containing transcripts (from MeRIP-seq and published RNA-BisSeq, respectively) by means of a ''bedtools merged''-based in-house script. Percentages of m^5C sites associated with SRSF2-bound and -unbound transcripts were computed globally and after stratifying the m^5C sites according to their stoichiometry: low (0%–33%), medium (34%–67%), or high (> 67%). The t-test was used to compare percentages of m^5C sites associated with SRSF2-bound transcripts.

Distribution of SRSF2 targets, m^5C, and DS events

To investigate the relationship between SRSF2 RNA binding, m^5C modification, and RNA splicing, the distributions of SRSF2 binding sites and m^5C sites surrounding splicing events were computed. To ensure an exonic position of the splicing event, the splicing position was defined as slightly upstream of the actual event, at the center of the flanking exon (exon located between ''flankingES'' and ''flankingEE'' for A5SS and A3SS events ; ''upstreamES'' and ''upstreamEE'' for MXE, RI and SE events in positive strand transcripts, and ''downstreamEE'' and ''downstreamES'' for MXE, RI, and SE events in negative strand transcripts). Then, the exonic regions covering 3 kb upstream and downstream of each splicing position, transcript-wise (i.e., introns excluded), were divided into 100 bins. Next, the SRSF2 binding sites (identified by PAR-CLIP seq in control HeLa or K562 cells), randomly selected control positions or m^5C sites located within 3 kb of the splicing event were identified with intersectBed[62] and finally the corresponding counts were plotted for each bin.

Translation efficiency analysis

Polysome profiling sequencing data in HeLa cells were downloaded from the GEO database under accession number GEO: GSE117299.[53] Translation efficiency is defined as the ratio of polysome/monosome reads. The polysome profiling sequencing data in K562 cells were downloaded from https://academic.oup.com/narcancer/article/4/2/zcac015/6576546#supplementary-data.[54] Translation efficiency is defined as the number of normalized polysome reads divided by the number of normalized RNA sequence reads. We then stratified the SRSF2 binding targets obtained by PAR-CLIP seq into loss and gain upon NSUN2 knockdown or SRSF2 mutant. Finally, the polysome profiling data were used to compare the translation efficiency of the altered SRSF2 binding transcripts with a Wilcoxon test.

NSUN2/NSUN6 expression level analysis

The expression profiles from Franzini et al. (GEO: GSE135902, CMML and age-matched old control samples)[34] and Pronier et al. (GEO: GSE188624, GSE165305)[55] were selected and downloaded from the GEO database for analysis. The raw expression data

of the unpublished collaborative CMML cohort were converted to transcripts per million (TPM) with the R function convertCounts in the package ''DGEobj.utils''. Expression levels of *NSUN2* and *NSUN6* were extracted and shown in boxplots. *P* values were calculated with the Wilcoxon test. GEPIA2 (http://gepia2.cancer-pku.cn/)[83] was used to analyze the expression of *NSUN2*/*NSUN6* in LAML patient samples and healthy controls from the TCGA (https://www.cancer.gov/tcga) and GTEx (https://gtexportal.org/) databases, respectively.

Overall survival analysis

Gene expression profiles of peripheral blood or bone marrow mononuclear cells from Bamopoulos et al., containing 246 AML samples (platform Illumina HiSeq 1500), were downloaded from the GEO database under accession number GEO: GSE146173[35] and analyzed to assess the prognostic impacts of $SRSF2^{P95H}$ and NSUN2. The raw gene counts, P95H mutation information, and overall survival information were extracted. Patients without the P95H mutation were labeled as the "WT group''. Gene counts were converted to counts per million (CPM) normalized with EdgeR's trimmed mean of M values (TMM) by means of R function convertCounts in the package ''DGEobj.utils''. Kaplan–Meier survival analysis based on overall survival and SRSF2 mutation was performed, and the p value was calculated with the log-rank test. Subsequently, patients were subdivided into ''WT *NSUN2*-high'', ''WT *NSUN2*-low'', ''P95H *NSUN2*-high'', and ''P95H *NSUN2*-low'' groups based on their median *NSUN2* expression level and P95H mutation status. The prognostic values of these four combinations were also estimated and visualized using the Kaplan–Meier method. Additionally, the association between SRSF2 mutation, *NSUN2* expression and survival was evaluated by means of single Cox proportional hazards models. The hazard ratios (HRs) and 95% confidence intervals (CIs) were calculated. Identical analysis was performed to assess the prognostic impacts of $SRSF2^{P95H}$ and NSUN6. For analysis of the Beat AML cohort, we downloaded the "Beat AML cohort clinical summary" table and "RPKM gene count" table from https://www.nature.com/articles/s41586-018-0623-z#Sec37.[36] These two tables contain gene expression and clinical characteristics information for 451 AML patient samples, respectively. The high-confidence SRSF2 P95H mutation information on AML patients refers tothe ''Genotype of patients from the AML cohorts'' table downloaded from https://www.nature.com/articles/s41586-019-1618-0#Sec29.[27] The RPKM count data for 451 patients were converted to $\log_2$TPM and then the patients were divided into four groups according to the median *NSUN2* or *NSUN6* expression levels. Patients whose life status was unknown and whose cause of death was marked as ''death-treatment,'' ''death-unknown,'' or ''death-other'' were excluded. For survival analysis of a total of 325 patients, the *NSUN2*/*NSUN6*-high and -low groupings were the same as those used in the gene expression analysis. All these survival analyses were performed using the ''survival'' package in R.

Leukemia-related gene expression analysis

We referred to the LGL database (the database of leukemia gene literature) for leukemia-associated genes in this study.[56] Differences in the expression levels of serval leukemia-associated genes in the ''WT *NSUN2*-high'', ''WT *NSUN2*-low'', ''P95H *NSUN2*-high'', and ''P95H *NSUN2*-low'' groups from the aforementioned two AML cohorts were analyzed and shown by boxplots. P values were calculated using the Wilcoxon test to compare gene expression between ''WT *NSUN2*-high'' and ''P95H *NSUN2*-low'' groups. The same analysis was performed on patients grouped according to the *NSUN6* expression levels.

QUANTIFICATION AND STATISTICAL ANALYSIS

For all of the experiments shown, n represents the number of replicates or patients and is indicated in the figure legends. Bioinformatics-associated statistical analyses were performed with the R package for statistical computing. For experimental quantification, ImageJ software was used for protein and RNA signal quantification. All statistics were evaluated by unpaired two-tailed Student's t test with GraphPad Prism 9 software, unless otherwise specified in the Figure legend or STAR Methods. Data and graphs are presented as mean ± SEM. The statistical significance criterion was p value < 0.05.

Molecular Cell, Volume *83*

Supplemental information

SRSF2 plays an unexpected role as reader of m^5C on mRNA, linking epitranscriptomics to cancer

Hai-Li Ma, Martin Bizet, Christelle Soares Da Costa, Frédéric Murisier, Eric James de Bony, Meng-Ke Wang, Akihide Yoshimi, Kuan-Ting Lin, Kristin M. Riching, Xing Wang, John I. Beckman, Shailee Arya, Nathalie Droin, Emilie Calonne, Bouchra Hassabi, Qing-Yang Zhang, Ang Li, Pascale Putmans, Lionel Malbec, Céline Hubert, Jie Lan, Frédérique Mies, Ying Yang, Eric Solary, Danette L. Daniels, Yogesh K. Gupta, Rachel Deplus, Omar Abdel-Wahab, Yun-Gui Yang, and François Fuks

Supplemental Information

Contents

Supplemental Figures, Titles, and Legends

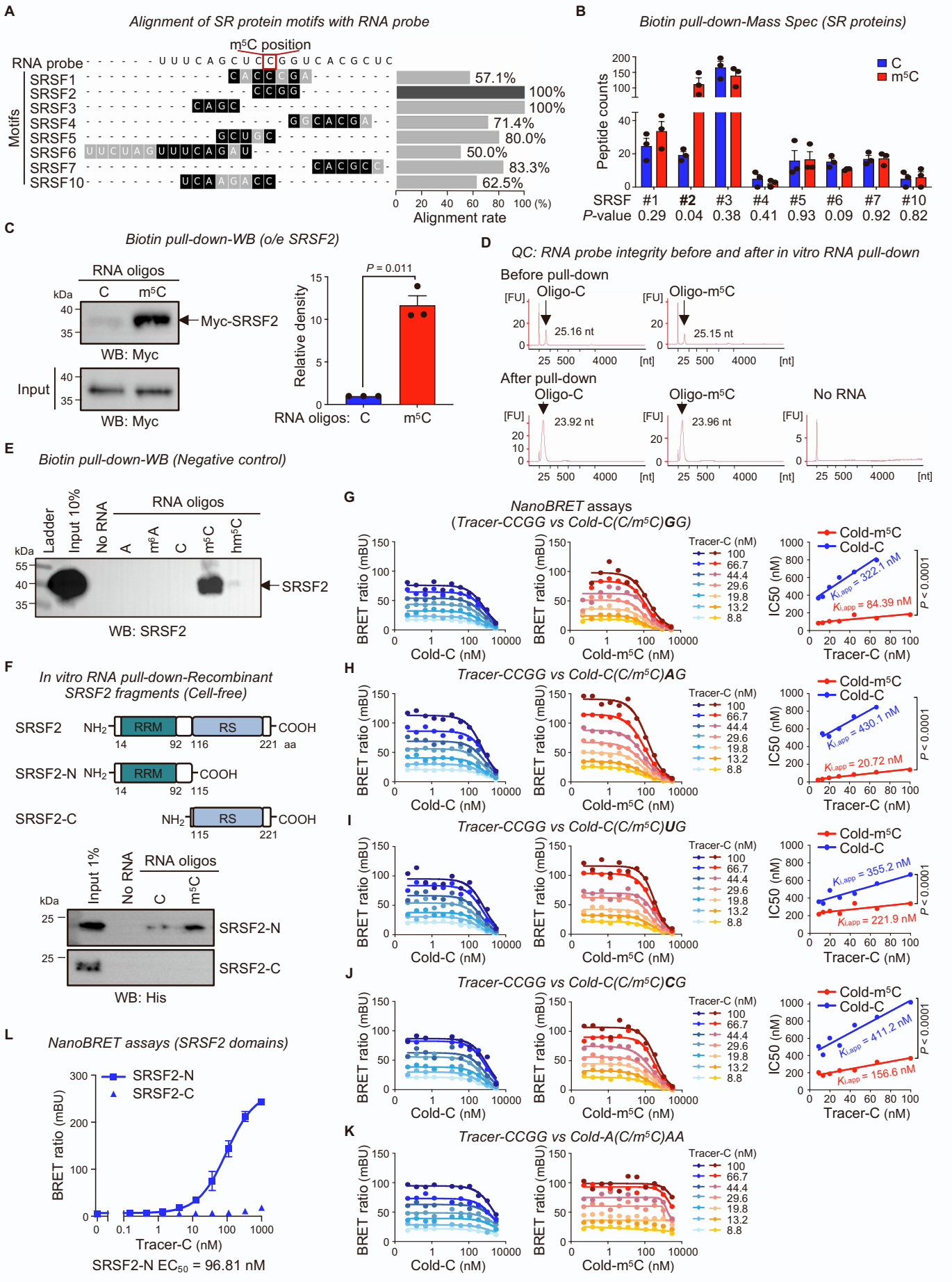


Figure S1 (related to Figure 1), Ma *et al.*

Figure S1. Among the SR proteins, only SRSF2 binds preferentially to the m^{5}C probe, and this binding is through its N-terminal domain (related to Figure 1)

(A) SR proteins detected by mass spectrometry all had a binding motif with at least 50% sequence identity to the RNA probe. Left: SR protein-binding motifs were aligned with RNA probe sequences (top) by means of the Needleman-Wunsch algorithm. Well-aligned bases are shaded in black. Right: bar chart showing the percentage of SR protein-binding motifs aligning with the RNA probe used in the biotin pull-down experiments.
(B) Among the SR-family proteins, only SRSF2 preferentially binds m^{5}C-modified RNA. Bar charts representing peptide counts of SR-family proteins, obtained by biotin pull-down followed by mass spectrometry (n = 3, mean ± SEM, paired two-tailed Student's t test).
(C) Biotin pull-down followed by western blotting shows that overexpressed Myc-tagged SRSF2 binds to oligo-m^{5}C with higher affinity than to oligo-C (n = 3, mean ± SEM, paired two-tailed Student's t test). A representative western blot (left) and quantification graph (right) are shown.
(D) Unmodified and m^{5}C-modified probes show the same stability in biotin pull-down experiments. Representative bioanalyzer electropherograms of RNA probes before (upper panel) and after (lower panel) *in vitro* RNA pull-down (n = 2 independent experiments). Peaks at 25 nt are gel migration markers. Normalized fragment sizes are indicated near the peaks. X-axis, fragment size in nucleotides, y-axis, arbitrary fluorescence units (FU).
(E) SRSF2 binds specifically to an m^{5}C-marked RNA oligo, as opposed to other modified RNA oligos. Biotinylated RNA pull-down of different RNA oligos in HeLa cells, followed by western blotting (n = 2 independent experiments).
(F) The N-terminus of SRSF2 is essential for m^{5}C binding. *In vitro* RNA pull-down was performed with the recombinant N-terminal or C-terminal domain of SRSF2 and unmodified or m^{5}C-modified RNA oligos (n = 2 independent experiments).
(G-K) SRSF2 binds to four different Cm^{5}CNG-containing RNA probes with higher affinity than to unmodified ones, whereas it binds weakly to non-C(C/m^{5}C)NG-containing probes. Concentration-dependent attenuation of BRET from nanoluciferase-SRSF2 fusions upon titration with cold-C (left) or cold-m^{5}C (middle) in the presence of a dilution series of CCGG-containing tracer RNA. Linearized Cheng–Prusoff analysis allows determination of the apparent dissociation constant ($K_{i,app}$) (right). Cold RNAs containing the sequences C(C/m^{5}C)GG (G), C(C/m^{5}C)AG (H), C(C/m^{5}C)UG (I), C(C/m^{5}C)CG (J), and A(C/m^{5}C)AA (K) were tested. Non-C(C/m^{5}C)NG-containing probes, A(C/m^{5}C)AA, were used as negative controls (K). P values were calculated by two-tailed F test. IC_{50}, half-maximal inhibitory concentration.
(L) NanoBRET binding assays with varying concentrations of RNA tracer-C were performed on permeabilized cells transiently transfected with constructs encoding Nanoluciferase fused to the N- or C-terminal portion of SRSF2. Individual data points are the mean ± SEM of three independent experiments. EC_{50}, half-maximal effective concentration.

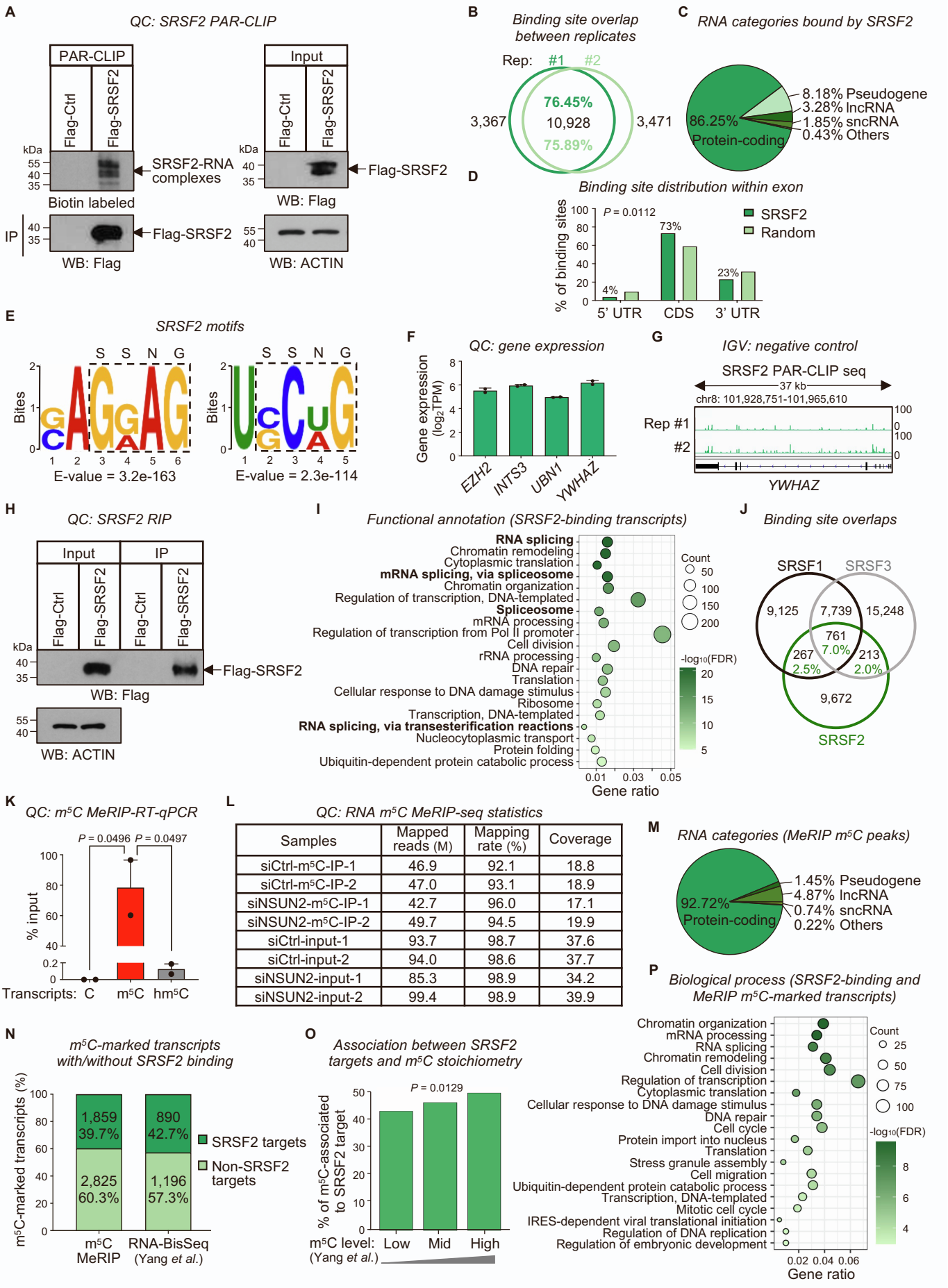


Figure S2 (related to Figure 2), Ma *et al.*

Figure S2. Quality control and analysis of SRSF2 PAR-CLIP-seq and RNA m^{5}C MeRIP-seq data (related to Figure 2)

(A) SRSF2 PAR-CLIP followed by a 3' end biotin-labeling assay of RNA pulled down by Flag-SRSF2 from HeLa cell extracts (upper left). Western blot showing the Flag-SRSF2 IP efficiency (lower left), Flag-Ctrl being used as a negative control. Overexpression of Flag-SRSF2 and of Flag-Ctrl are shown in the right panel; ACTIN was used as a loading control. The data shown are representative of two independent experiments. Ctrl, control.
(B) Venn diagram showing the overlap between biological replicates for SRSF2 PAR-CLIP-seq. Percentages were calculated by dividing the number of binding sites in the overlap by the number identified in each replicate (color matches the sample color).
(C) SRSF2 preferentially binds protein-coding transcripts. Percentages of various RNA species containing SRSF2-binding sites detected by PAR-CLIP-seq. lncRNA, long noncoding RNA; sncRNA, small noncoding RNA.
(D) Non-random distribution of SRSF2-binding sites. Bar chart showing the distribution of SRSF2-binding sites (dark green) and randomly selected control positions (light green) across the 5' UTR, CDS, and 3' UTR regions of protein-coding genes. The p value was determined with a chi-square test.
(E) Other SSNG-containing motifs overrepresented at SRSF2-binding sites. The E-value is the enrichment p value (Fisher's exact test) times the number of candidate motifs tested.
(F) The expression levels of SRSF2-binding targets (e.g., *enhancer of zeste homolog 2 (EZH2), integrator complex subunit 3 (INTS3), ubinuclein 1 (UBN1)*) and non-binding targets (*YWHAZ*) are comparable. Bar plot showing transcript levels (log_2TPM) in HeLa cells, as determined by RNA-seq (n = 2, mean ± SEM).
(G) IGV tracks displaying PAR-CLIP-seq read coverage at the *YWHAZ* locus (as a negative control).
(H) Western blot showing the efficiency of Flag-SRSF2 IP for the RIP-qPCR experiment in HeLa cells (representative of three independent experiments). Flag-Ctrl was used as a negative control; ACTIN was used as a loading control.
(I) Biological processes and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways over-represented among transcripts bound by SRSF2. The color represents the Benjamini–Hochberg adjusted p values obtained with DAVID. Spot size indicates the number of genes involved.
(J) Very little overlap between SRSF2- and SRSF1- or SRSF3-binding sites. Venn diagram showing the overlap between SRSF2-, SRSF1-, and SRSF3-binding sites identified by PAR-CLIP-seq. Percentages were calculated by dividing the number of binding sites in the overlap by the number identified for each protein. The p value for the overlaps of the SRSF2-binding sites with the SRSF1- and SRSF3-binding sites was equal to 1, as calculated with the hypergeometric test.
(K) The m^{5}C antibody we used is specific to m^{5}C-containing RNA. Bar plot showing MeRIP-RT-qPCR results for *in vitro* transcribed RNA transcripts containing unmethylated, methylated, or hydroxymethylated cytosines (n = 2, mean ± SEM, unpaired two-tailed Student's t test).
(L) RNA m^{5}C MeRIP-seq statistics. The number of mapped reads, mapping rate and coverage are shown in the table.
(M) Most of the identified m^{5}C peaks are within protein-coding transcripts. Percentages of various RNA species containing m^{5}C peaks detected by MeRIP-seq.
(N) Approximately 40% of m^{5}C-modified RNAs are bound by SRSF2. Stacked bar chart depicting the percentage and number of m^{5}C-modified RNAs identified by m^{5}C MeRIP-seq and RNA-BisSeq bound (dark green) or not bound (light green) by SRSF2.
(O) The proportion of SRSF2-target-associated m^{5}C sites is highest among high-stoichiometry sites. The m^{5}C sites were binned according to their m^{5}C levels: low (0-33%), medium (34-67%) or high (> 67%). The p value was determined with the unpaired two-tailed Student's t test.
(P) The top 20 over-represented biological processes of SRSF2-bound m^{5}C-marked (MeRIP-seq) transcripts. The color represents the adjusted p value. Spot size indicates the number of genes concerned.

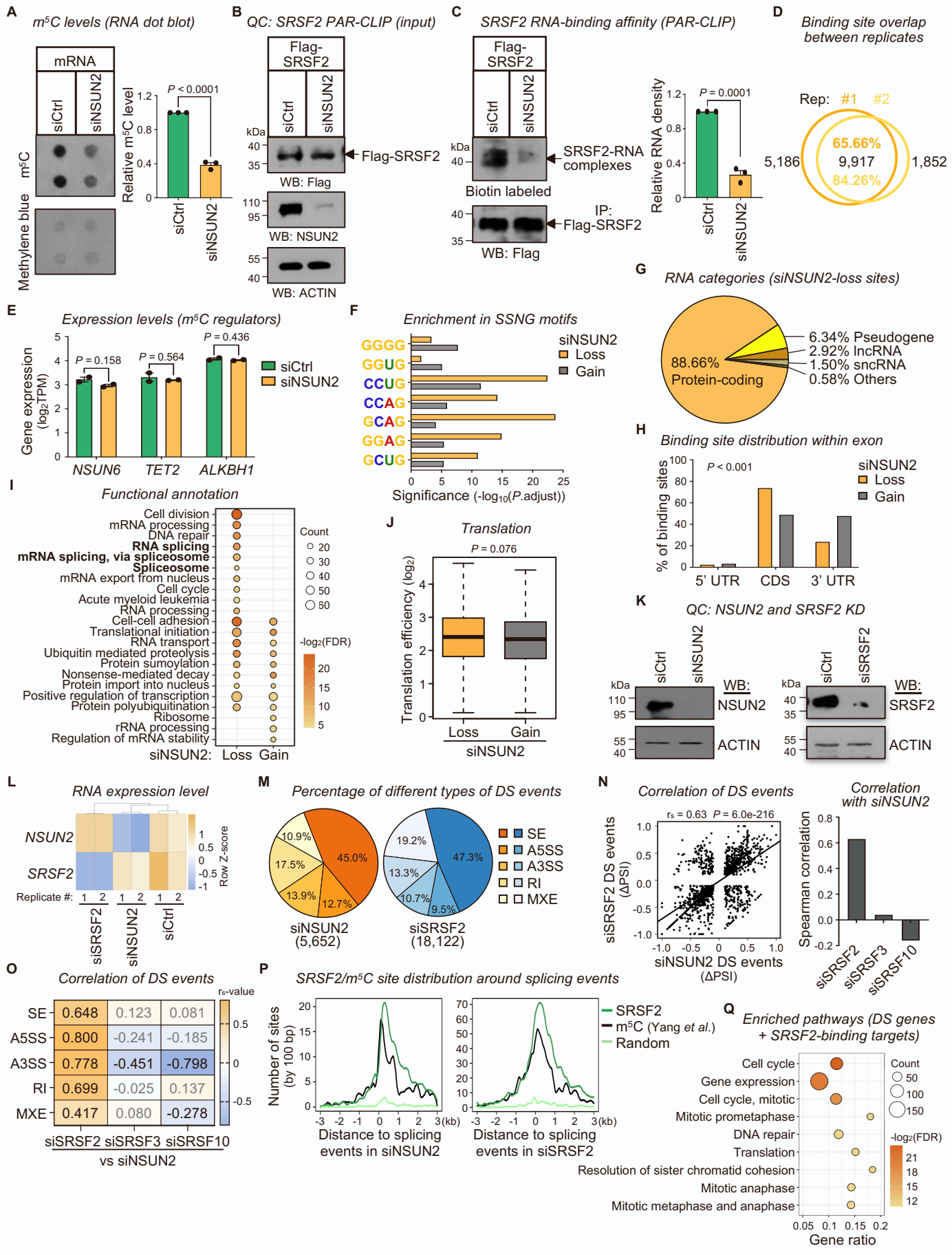


Figure S3 (related to Figure 3), Ma *et al.*

Figure S3. Quality control and effects of NSUN2 knockdown on m^5C levels, SRSF2 RNA binding, translation and splicing in HeLa cells (related to Figure 3)

(A) Overall decreased m^5C upon NSUN2 knockdown. Dot blotting with anti-m^5C antibody, applied to mRNA from control and NSUN2 KD HeLa cells. The results are mean ± SEM (n = 3, unpaired two-tailed Student's t test) with a representative blot shown.
(B) Western blot showing the efficiency of Flag-SRSF2 overexpression and of NSUN2 knockdown. ACTIN was used as a loading control (n = 3).
(C) SRSF2 RNA-binding affinity decreased upon NSUN2 knockdown. PAR-CLIP followed by 3' end biotin-labeling assay (representative pictures are shown on the left panel) and quantification (right) of RNA pulled down by Flag-SRSF2 in control and NSUN2 KD HeLa cells (n = 3, mean ± SEM, unpaired two-tailed Student's t test).
(D) Venn diagram showing the between-replicate overlap of SRSF2-binding sites identified by PAR-CLIP-seq in NSUN2 KD HeLa cells. Percentages were calculated by dividing the number of binding sites in the overlap region by the number identified in each replicate (color matches the sample color).
(E) Expression levels of m^5C-machinery genes did not change significantly upon NSUN2 knockdown. Bar plot showing the expression (log_2TPM) of mRNA m^5C writer NSUN6 and erasers (TET2, ALKBH1) from control (green) and NSUN2 knockdown (yellow) HeLa cells identified by RNA-seq (n = 2, mean ± SEM, unpaired two-tailed Student's t test).
(F) Redirection of SRSF2 towards non-C-containing binding sites upon NSUN2 knockdown. Bar graph showing the enrichment scores ($-log_{10}(P.adjust)$) of SSNG (S = C/G, N = C/G/A/U) motifs at siNSUN2-gain and -loss SRSF2-binding sites. Only SSNG motifs with adjusted p value < 0.05 are displayed. Motif enrichment p values were analyzed according to Fisher's exact test, adjusted for multiple tests with a Bonferroni correction.
(G) The majority of SRSF2-binding targets lost after NSUN2 knockdown are protein-coding transcripts. Pie charts showing the percentages of various RNA species containing siNSUN2-loss SRSF2-binding sites.
(H) Redistribution of SRSF2-binding sites upon NSUN2 knockdown. Bar chart showing the proportion of SRSF2-binding sites lost (orange) or gained (grey) after NSUN2 KD in protein-coding transcripts. The p value was determined with a chi-square test.
(I) Functional annotation enrichment analysis of SRSF2-binding transcripts lost or gained after NSUN2 KD. The color represents the adjusted p values. The scale of the spots indicates the number of genes involved.
(J) Altered binding profiles observed in NSUN2-depleted cells do not affect translation. Box plot showing the translation efficiency of SRSF2-binding targets lost (orange) or gained (grey) after NSUN2 KD in HeLa cells (Wilcoxon test). The center line in the box represents the median value, the box and whiskers represent the interquartile range (IQR) and 1.5 times the IQR, respectively.
(K) Representative western blot showing the knockdown efficiency of NSUN2 and SRSF2. ACTIN was used as a loading control (n = 2 independent experiments).
(L) Heatmap showing the expression levels of NSUN2 and SRSF2 in siCtrl, siNSUN2, and siSRSF2 cells estimated from RNA-seq (n = 2 biological replicates). High expression is indicated in orange and low expression is in blue.
(M) Pie-chart displaying the percentages of each class of alternative splicing events identified in SRSF2- or NSUN2-depleted HeLa cells compared with the control samples.
(N) Scatter plot demonstrating strong positive correlation between differential percent spliced in (ΔPSI) for identified DS events following NSUN2 and SRSF2 depletion. Correlations between

siNSUN2 and siSRSF3 or siSRSF10 were used as negative controls. The centerline indicated the regression line. Correlation coefficient (r_s) and p values were calculated by Spearman's correlation analysis.

(O) Heatmap showing the correlation between ΔPSI for the five types of DS events identified in siNSUN2, siSRSF2, siSRSF3 and siSRSF10 HeLa cells. Positive correlation is indicated in orange and negative correlation is indicated in blue. Correlation coefficient values are labeled on the heatmap, with black values representing p values < 0.05 and gray values representing p values > 0.05. The r_s-values and p values were calculated by Spearman's correlation analysis.

(P) SRSF2-binding sites and m^5C sites localize preferentially around NSUN2- and SRSF2-associated splicing events. Frequency plots showing the distribution of SRSF2-binding sites (dark green), m^5C sites (black), and randomly selected control positions (light green) relative to the splicing events identified in NSUN2-depleted (left) and SRSF2-depleted (right) cells.

(Q) Pathway enrichment analysis of differentially spliced SRSF2-binding genes. The color represents the Benjamini–Hochberg adjusted p values. The scale of the spots indicates the number of genes involved.

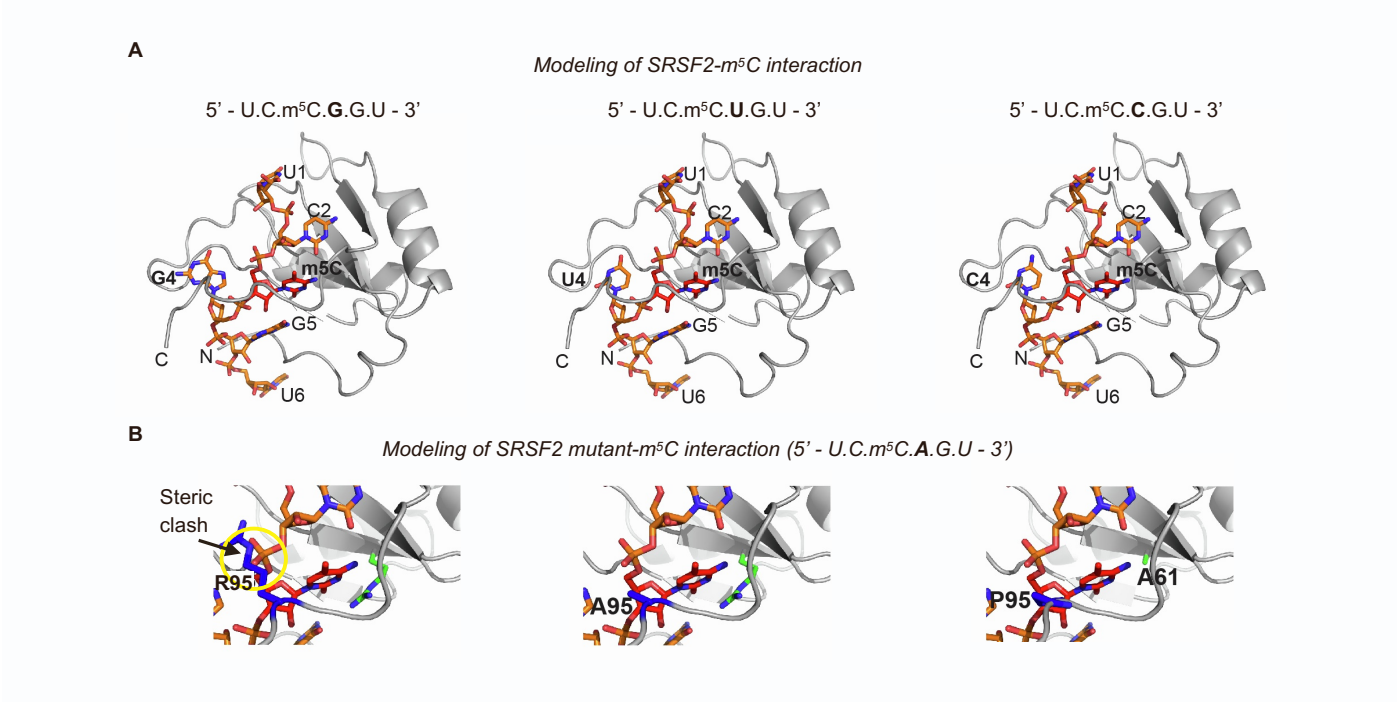


Figure S4 (related to Figure 4), Ma *et al.*

Figure S4. Modeling of wild-type and mutant SRSF2 binding to m^{5}C-modified RNAs (related to Figure 4)

(A) Modeling of SRSF2 RRM domain binding to the m^{5}C-modified RNAs UCm^{5}CGGU (left panel), UCm^{5}CUGU (middle panel), and UCm^{5}CCGU (right panel).
(B) Close-up view of the m^{5}C binding pocket of the P95R mutant (modeled arginine, left panel), the P95A mutant (modeled alanine, middle panel), and the R61A mutant (modeled alanine, right panel). Protein is presented as a gray cartoon, RNA is presented as orange sticks, and important protein side chains involved in m^{5}C interaction are represented as sticks.

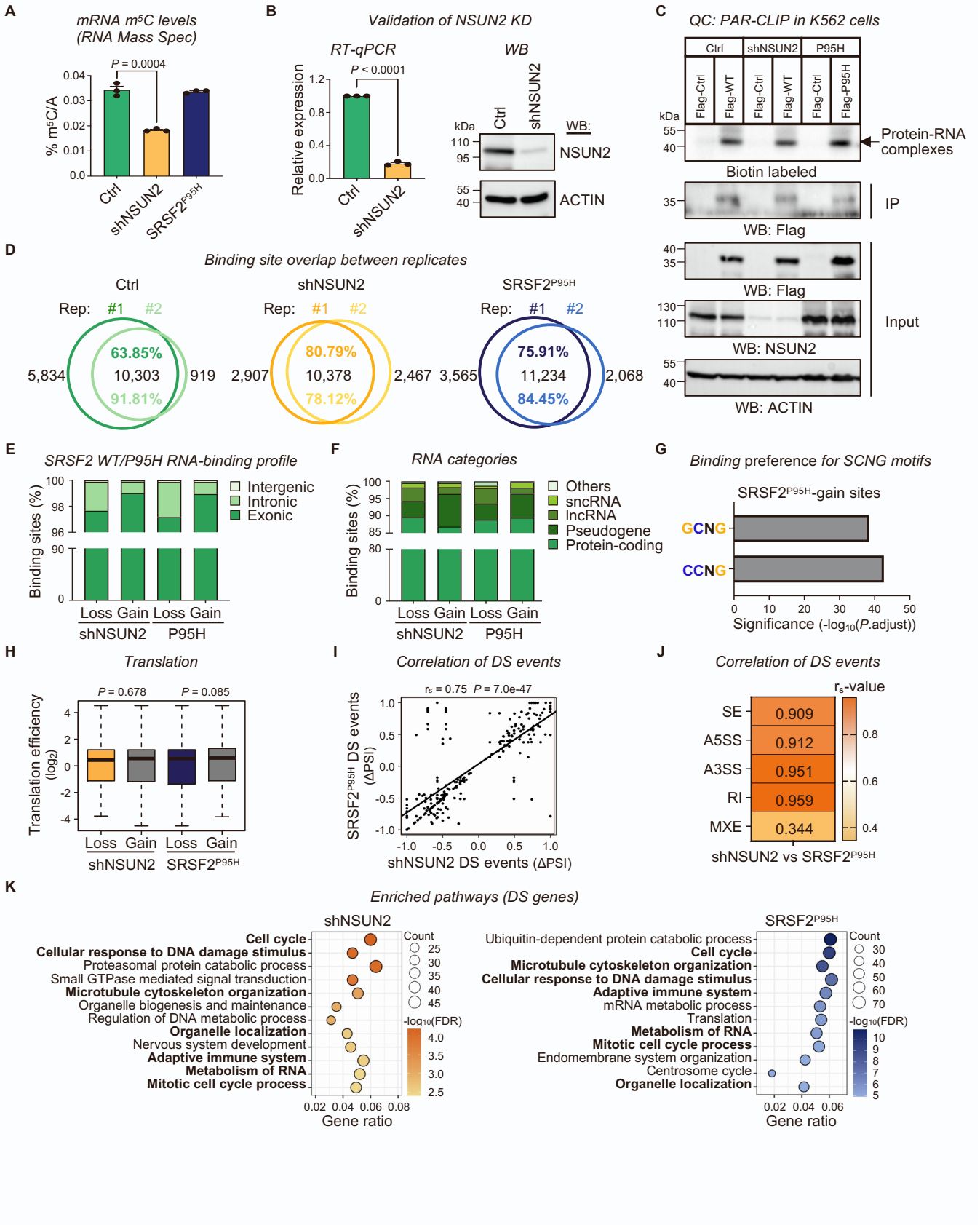


Figure S5 (related to Figure 5), Ma *et al.*

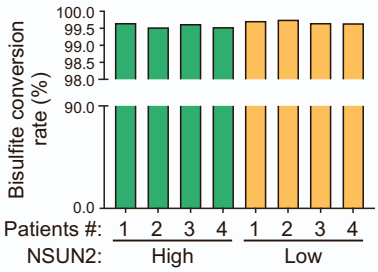
Figure S5. Involvement of m^5C regulatory transcripts in leukemia (related to Figure 5)

(A) LC-MS/MS results showed that NSUN2 KD reduces m^5C levels, while the $SRSF2^{P95H}$ mutation does not affect them. Bar plot showing mRNA m^5C levels in control, NSUN2 KD, and $SRSF2^{P95H}$ mutant K562 cells (n = 3, mean ± SEM, unpaired two-tailed Student's t test).
(B) RT-qPCR and western blot of control and NSUN2 KD K562 cells showing NSUN2 knockdown efficiency (n = 3, mean ± SEM, unpaired two-tailed Student's t test). ACTIN was used as a loading control.
(C) PAR-CLIP followed by a 3' end biotin-labeling assay of RNA pulled down by Flag antibody in K562 cells. Western blot showing Flag IP efficiency and Flag-SRSF2 WT or P95H overexpression efficiency. Flag-Ctrl as negative control and ACTIN as a loading control. The data shown are representative of two independent experiments.
(D) Venn diagrams showing good overlap of binding sites between two biological replicates identified in control (green), NSUN2 KD (orange) and $SRSF2^{P95H}$ (blue) K562 cells.
(E) Most binding sites are located within exonic regions. Stacked bar chart displaying the percentages of SRSF2-binding sites aligned to exonic, intronic, and intergenic regions.
(F) The majority of binding targets are protein-coding transcripts. Stacked bar chart showing the percentages of various RNA species bound by SRSF2 based on PAR-CLIP-seq analysis.
(G) $SRSF2^{P95H}$ mutant alters the affinity of SRSF2 for binding to SCNG motifs. Bar graph showing the enrichment scores of the SCNG (S = C/G, N = C/G/A/U) motifs at $SRSF2^{P95H}$-gain sites. Motif enrichment p values were analyzed with Fisher's exact test, adjusted for multiple tests with a Bonferroni correction.
(H) Altered binding profiles observed in NSUN2-depleted or $SRSF2^{P95H}$ mutant cells do not affect translation. Box plot showing the translation efficiency of SRSF2-binding targets lost or gained after NSUN2 KD or SRSF2 mutant in K562 cells (Wilcoxon test). The center line in the box represents the median value, the box and whiskers represent the interquartile range (IQR) and 1.5 times the IQR, respectively.
(I) Scatter plot showing a strong positive correlation between ΔPSI values of DS events identified in shNSUN2 and $SRSF2^{P95H}$ cells. The center line indicated the regression line. The correlation coefficient (r_s) and p value were calculated by Spearman's correlation analysis.
(J) Heatmap showing strong positive correlations between ΔPSI values for five types of DS events identified in NSUN2-depleted and SRSF2 mutant K562 cells. Correlation coefficient values are labeled on the heatmap. The r_s-values and p values were calculated by Spearman's correlation analysis.
(K) NSUN2 depletion and SRSF2 mutation mediated splicing alterations co-impact many downstream biological pathways. Bubble plot displaying the top twelve enriched pathways in DS genes identified in NSUN2-depleted (left panel) or SRSF2 mutant (right panel) cells, colored by the enrichment FDR values. The size of the spots indicates the number of genes involved.

A

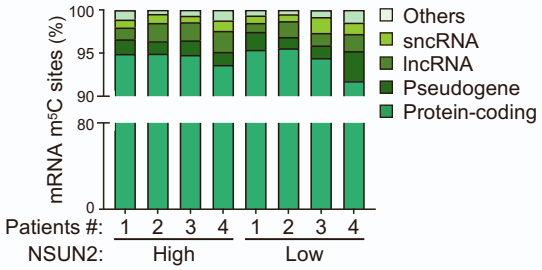
QC: RNA-BisSeq statistics

Patients	Mapped reads (M)	Mapping rate (%)	Coverage
NSUN2-High-1	48.4	73.22%	38.8
NSUN2-High-2	35.5	69.66%	28.5
NSUN2-High-3	47.8	68.75%	38.4
NSUN2-High-4	58.3	73.77%	46.8
NSUN2-Low-1	48.3	75.71%	38.8
NSUN2-Low-2	46.8	75.21%	37.5
NSUN2-Low-3	54.4	73.71%	43.7
NSUN2-Low-4	55.3	73.60%	44.3



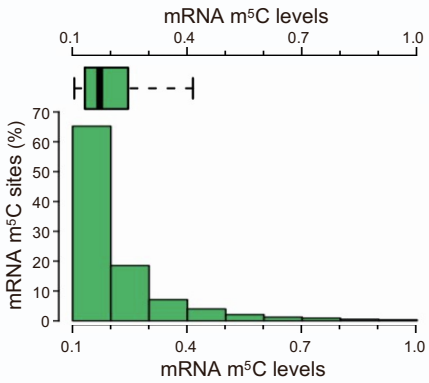
B

RNA categories (RNA-BisSeq)



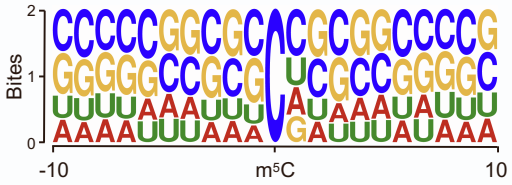
C

RNA-BisSeq m5C levels (all patients)



D

Sequence frequency logo



E

RNA-BisSeq m5C site distribution

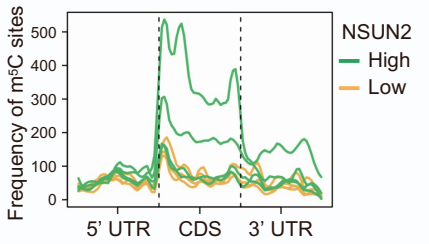


Figure S6 (related to Figure 6), Ma *et al.*

Figure S6. Overview of RNA m^5C patterns in CMML monocytes (related to Figure 6)

(A) RNA m^5C MeRIP-seq statistics. The number of mapped reads, mapping rate, and coverage are shown in the table (left panel). The bisulfite conversion rate (total analyzed C to T conversions) is presented in the bar graph (right panel).
(B) Most m^5C sites mapped to protein-coding transcripts. Stacked bar graph showing the percentage of m^5C sites identified by RNA-BisSeq mapping to various RNA species in each patient.
(C) Histogram and box plot showing mRNA m^5C levels in all patients. The median m^5C level was 16.7%, over 60% of m^5C sites having levels between 10% and 20%.
(D) m^5C sites are embedded in environments with high CG content. Representative sequence frequency logo showing the sequence context of m^5C sites.
(E) Frequency plots displaying the distribution of m^5C sites identified by RNA-BisSeq along mRNA transcripts in each patient.

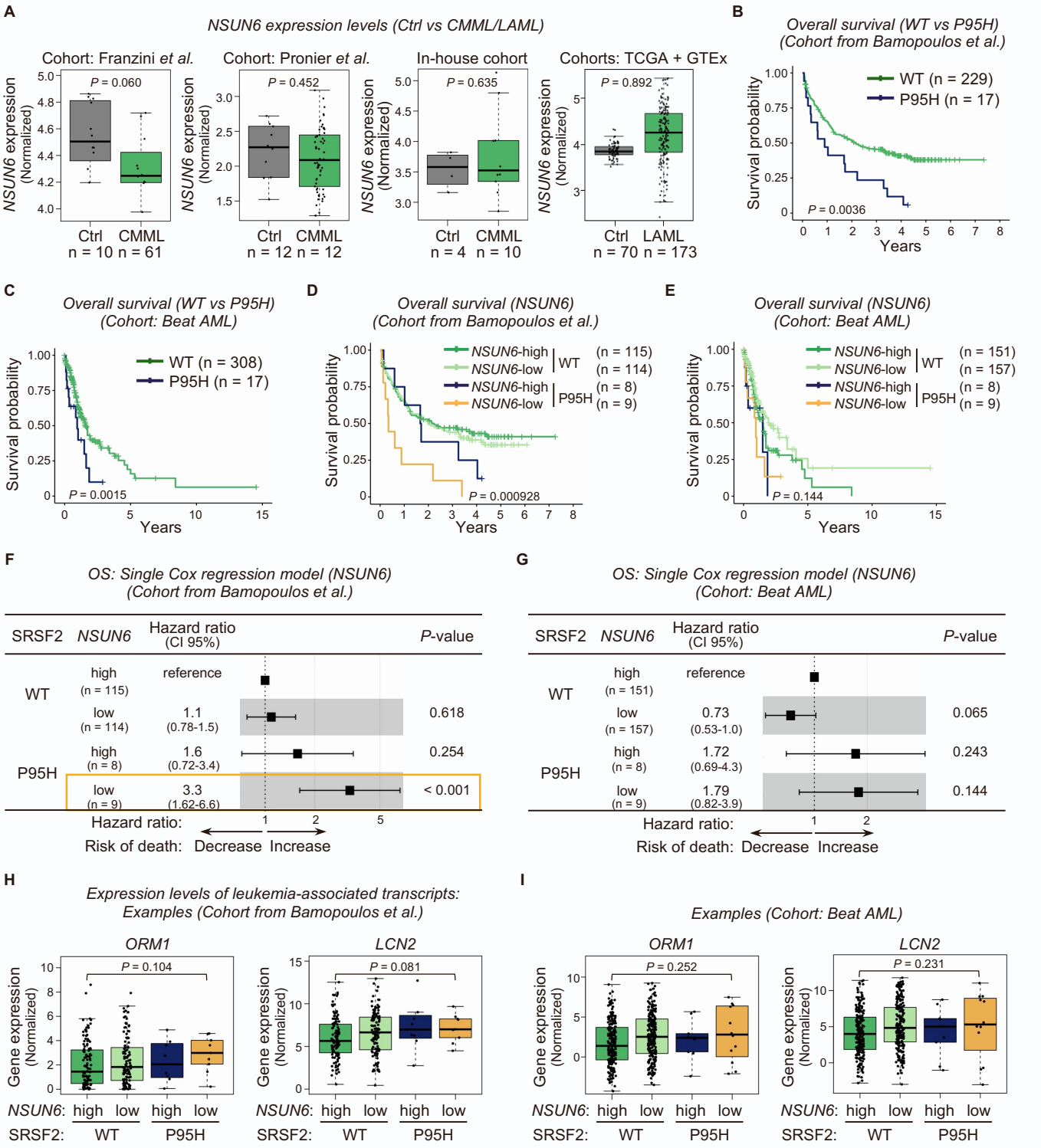


Figure S7 (related to Figure 7), Ma *et al.*

Figure S7. Survival and oncogene expression analysis of AML patients (related to Figure 7)

(A) Boxplots overlaid with dotplots showing *that NSUN6* expression levels in CMML and LAML patients are not significantly different from those recorded for healthy controls. The p value comparing TCGA and GTEx data was computed by GEPIA2. All other p values were calculated with the Wilcoxon test.
(B)-(C) SRSF2^{P95H} is associated with poor prognosis. Overall survival curves for AML patients with WT or mutant SRSF2 from Bamopoulos et al. (B) and from Beat AML (C).
(D)-(E) Kaplan-Meier graphs for AML patients, grouped by SRSF2 mutation status and *NSUN6* expression levels. P values for (B)-(E) were determined with the log-rank test.
(F)-(G) Forest plot depicting overall survival hazard ratios (HRs) with 95% confidence intervals (CIs) for AML patients. Squares represent the hazard ratios and the horizontal bars extend from the lower to the upper limit of the 95% confidence interval of the hazard ratio.
(H-I) High oncogene expression was not observed in AML patients with SRSF2^{P95H} and low *NSUN6* expression. Boxplots overlaid with dotplots showing expression levels of the leukemia-associated oncogenes *ORM1* and *LCN2* in four patient groups. The center line in the box represents the median value, the box and whiskers represent the IQR and 1.5 times the IQR, respectively. P values were calculated with the Wilcoxon test.