SUPPLEMENTARY APPENDIX

Differential RNA splicing as a potentially important driver mechanism in multiple myeloma

Michael A. Bauer,[1] Cody Ashby,[1] Christopher Wardell,[1] Eileen M. Boyle,[1] Maria Ortiz,[2] Erin Flynt,[3] Anjan Thakurta,[3] Gareth Morgan[4] and Brian A. Walker[1,5]

[1]University of Arkansas for Medical Sciences, Little Rock, AR, USA; [2]Celgene Institute for Translational Research Europe, Sevilla, Spain; [3]Translational Development and Diagnostics, Celgene Corporation, Summit, NJ, USA; [4]NYU Langone Medical Center, Perlmuter Cancer Center, NYU Langone Health, New York, NY, USA and [5]Division of Hematology Oncology, Indiana University, Indianapolis, IN, USA

 doi:10.3324/haematol.2019.235424

Received: September 13, 2019.

Accepted: February 12, 2020.

Pre-published: February 20, 2020.

Correspondence: *MICHAEL A. BAUER* - mbauer2@uams.edu

Title

Differential RNA splicing as a potentially important driver mechanism in multiple myeloma

Running Title

Alternative splicing in multiple myeloma

Authors:

Michael A Bauer[1], Cody Ashby[1], Christopher Wardell[1], Eileen M. Boyle[1], Maria Ortiz[2], Erin Flynt[3], Anjan Thakurta[3], Gareth Morgan[4] and Brian A Walker[1,5]

Affiliations:

[1] University of Arkansas for Medical Sciences, Little Rock, AR
[2]Celgene Institute for Translational Research Europe, Sevilla, Spain
[3]Translational Development and Diagnostics, Celgene Corporation, Summit, NJ
[4]NYULangone Medical Center, Perlmuter Cancer Center, NYU Langone Health, NYC, NY.
[5]Division of Hematology Oncology, Indiana University, Indianapolis, IN.

Corresponding Author:

Assistant Professor Michael A Bauer, Myeloma Center, University of Arkansas for Medical Sciences, 4301 W Markham Street, Little Rock, AR 72205.

Supplementary Methods

Differential gene and transcript expression analysis

StringTie[1] (v1.3.4), an assembler of RNA-Seq alignments into transcripts, was used to assemble and quantify transcripts aligned with STAR[2], including identifying potential novel transcripts. The R package DESeq2[3] (v1.14.1) was used for expression normalization and gene and transcript level differential expression analysis on both the StringTie and Salmon results.

Pathway analysis

Gene set enrichment analysis was performed using the R package fgsea[4] (v1.6.0). The stat value that DESeq2 provides for each gene tested for differential expression is the ranking value for input into fgsea. The online tool STRING[5] database is a collection of known and predicted protein-protein interactions and builds a network diagram of possible interactions. STRING additionally performs functional enrichment on the network and provides a list of statistically significant (adjusted for multiple testing) over-represented pathways.

Partitioning samples into groups based on novel splice loci

QoRTs[6] QC metrics reports the number of novel splice loci for each sample. A novel splice locus is defined as a previously unannotated splice junction with greater than four reads spanning the junction. After ranking the samples by the number of novel splicing loci the first group set included the top and bottom 20th and middle 60th percent of samples. A second grouping included the top and bottom 5th and middle 90th percent.

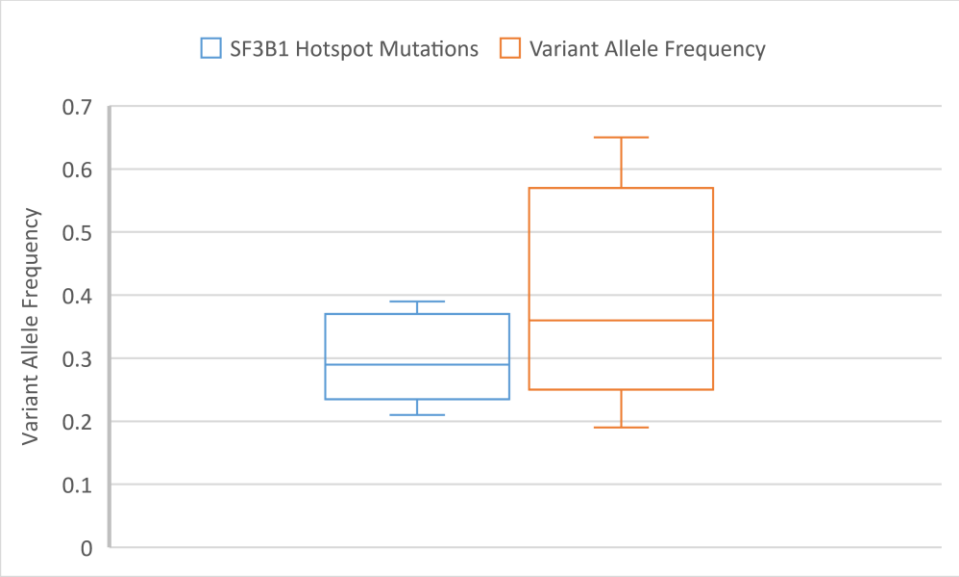
Comparisons of these groups to important clinical and genomic characteristics of MM was evaluated using the chi-square test with significance being *p*-value < 0.05.

Survival analysis statistics

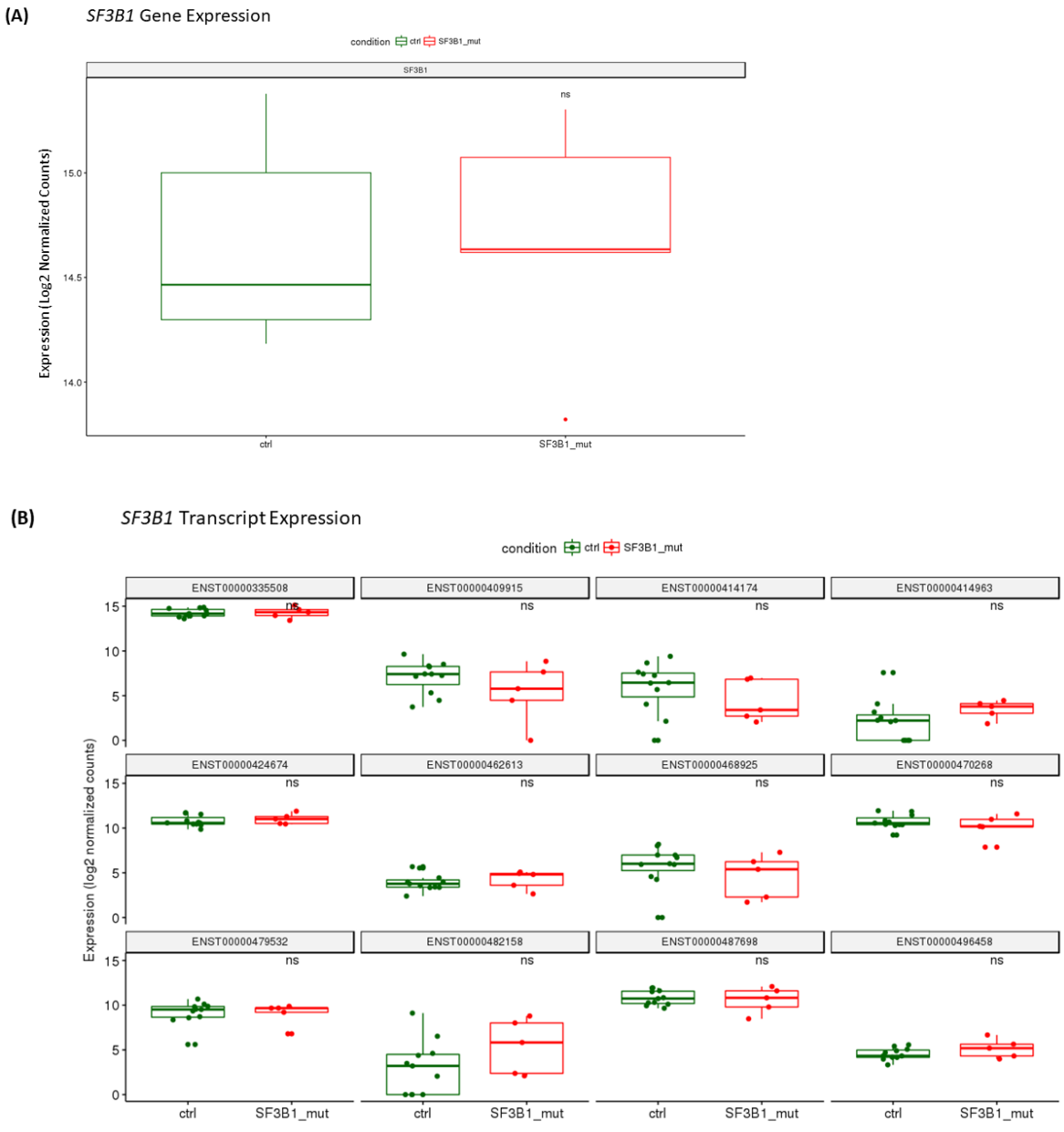
The Kaplan–Meier estimator was used to calculate time-to-event distributions using the R survival[7] (v2.43-3) and survminer[8] (v0.4.3) packages. The logrank test was used to test the null hypothesis that there was no difference between the populations in the probability of an event (progression or death) at any time point. Univariate and multivariate analysis were done using the survivalAnalyisis R package.

Supplementary Figures

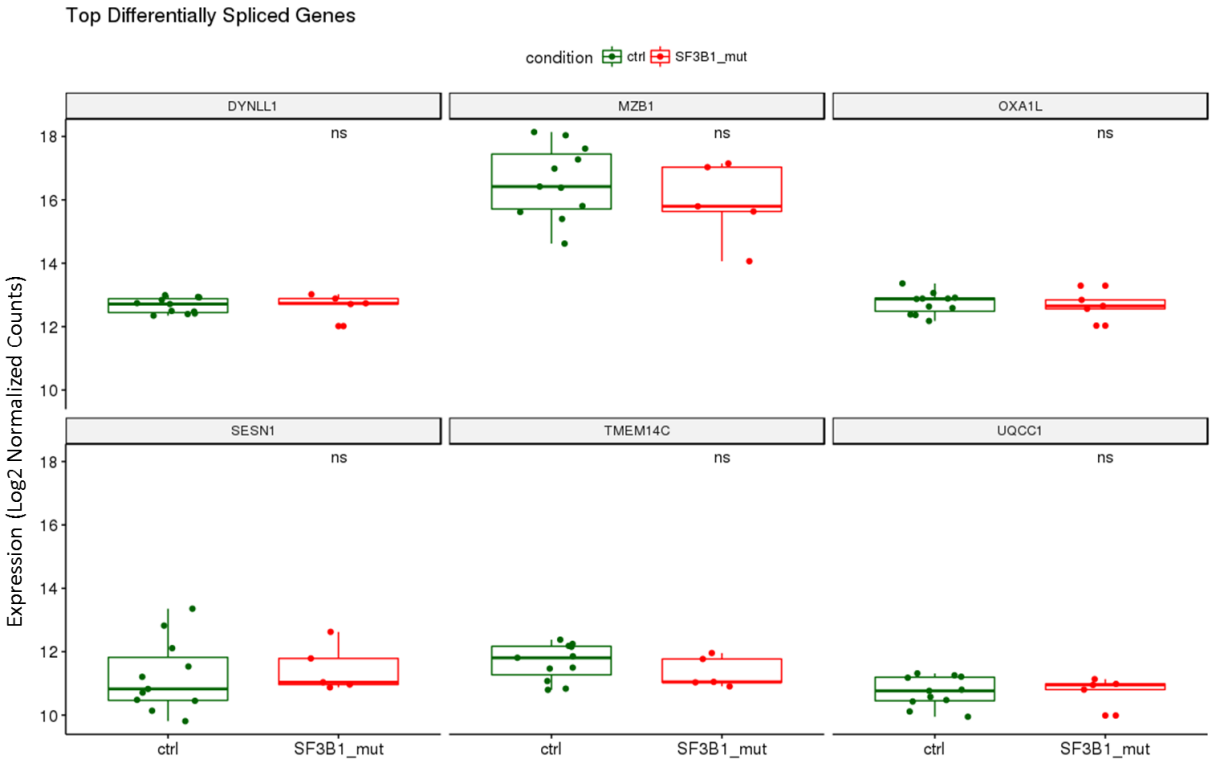
Supplementary Figure 1: Box plot of the variant allele frequencies of the hotspot *SF3B1* mutations versus the other non-silent *SF3B1* mutations. There was not a statistical difference (*p*-value = 0.2629, *n*=12) in VAF. Suggesting that the alternative splicing we see with the *SF3B1* hotspot mutations is not due to a difference in VAF.



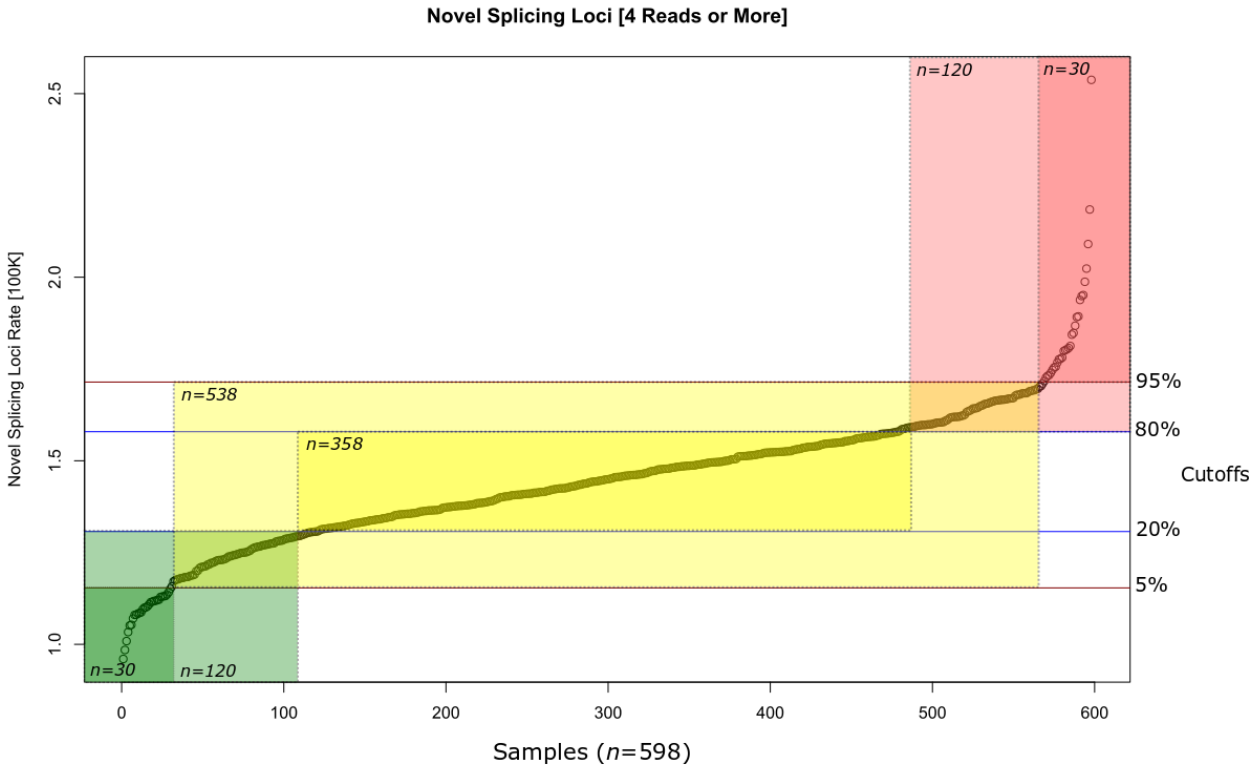
Supplementary Figure 2: *SF3B1* expression. A) *SF3B1* gene expression was not significantly different between the control (*n*=11) and *SF3B1* mutant samples (*n*=5). B) SF3B1 transcript level expression. No significant differences between the control and *SF3B1* mutant samples in any of the transcripts.
(ns: *p*-value>0.05)



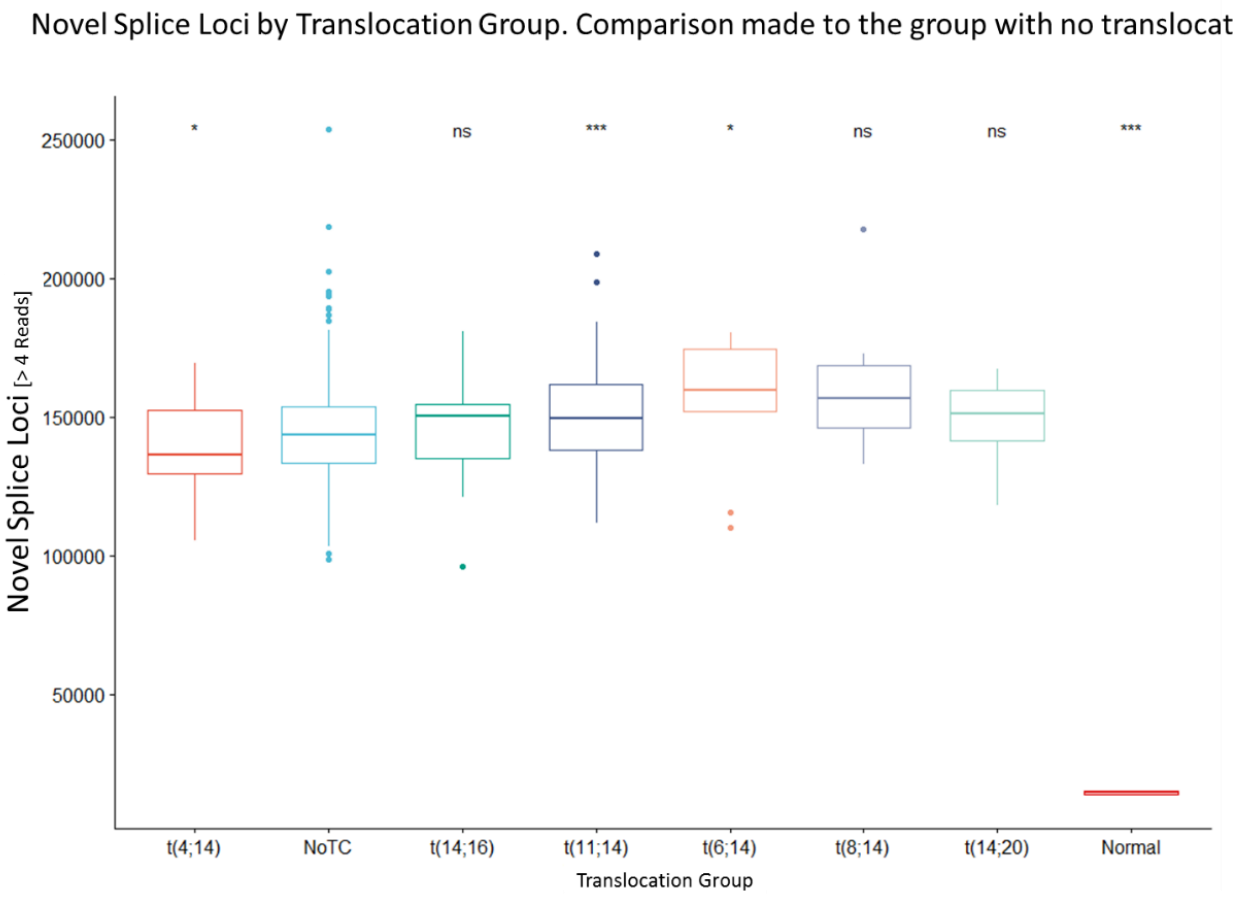
Supplementary Figure 3: Six genes that were identified by all three differential splicing analysis tools, between the control (*n*=5) and *SF3B1* mutated samples (*n*=11). Although they were all found to be significantly differentially spliced they all were not significantly differentially expressed at the gene level.
(ns: *p*-value>0.05)



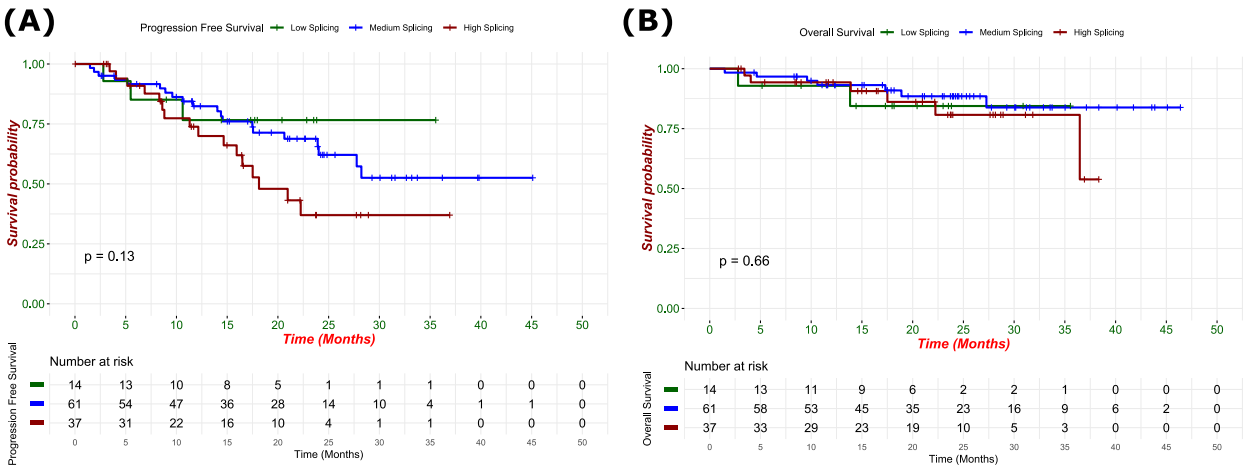
Supplementary Figure 4: Samples ordered by the number of novel splice loci. Novel splice loci are not present in exon annotation and had four or more reads spanning the loci. The samples were split into high, middle, and low groups. First at the 20th and 80th quantile. Additional analysis was done based on groups split at the 5th and 95th quantile.



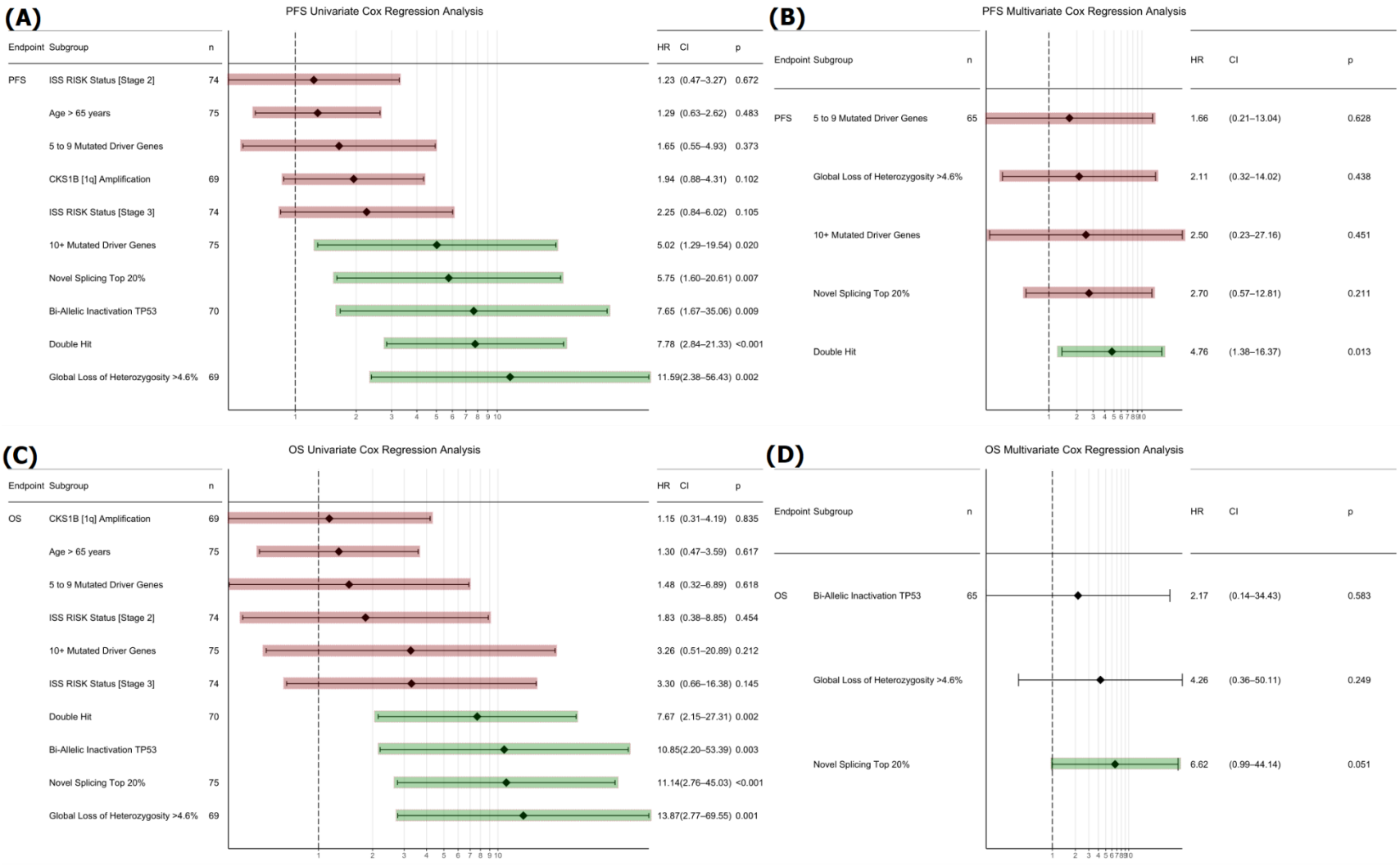
Supplementary Figure 5: Boxplots of the number of novel splice loci that had four or more reads grouped by translocation. A statistical comparison for each group was performed against the group with no translocation (*n*=602). We see that t(4;14) was the only subgroup that had significant lower number of novel splicing. The t(11;14) and t(6;14) were both had significant higher number of novel splicing. A set of pooled normal was included and we can see that all MM groups had significantly increased levels of novel splicing versus the normal samples.
(ns: $p>0.05$, *: $p\leq0.05$, ***: $p\leq0.001$)



Supplementary Figure 6: Kaplan Meier curves in samples with an t(11;14) in regards to high, medium, and low novel splicing patient samples (A) progression free survival (B) overall survival (low: *n*=14, medium: *n*=61, high: *n*=37).



Supplementary Figure 7: Features associated with outcome in the t(4;14) subgroup. Univariate and multivariate Cox regression analysis in regards to PFS and OS. The Univariate analysis included the following covariates: number of driver mutations (0 to 4, 5 to 9 and greater than 10 mutations), ISS stage (I, II and III), novel splice site group (high and medium/low) inactivated *TP53* (normal, one allele and both alleles), double hit (bi-allelic TP53 inactivation or amp 1q on a background of ISS stage III) and global loss of heterozygosity (LOH) (> 4.6%). (A) Bi-allelic inactivation of *TP53*, high novel splicing, double hit, LOH and having more than 10 driver mutations were associated with poorer PFS (*p*-value <= 0.05). (B) Only double hit was found to be significant in the multivariate analysis for PFS with a *p*-value of 0.013 and HR of 4.76 (95% CI 1.38-16.37). (C) Univariate analysis of OS found high novel splicing and bi-allelic inactivation of *TP53* were associated with poorer OS (*p*-value <= 0.05). Multivariate analysis of OS found only high splicing remained close to significant with a *p*-value of 0.051 and hazard ratio of 6.62 (95% CI 0.99-44.14).


(A)
PFS Univariate Cox Regression Analysis
Endpoint Subgroup n HR CI p
PFS ISS RISK Status [Stage 2] 74 1.23 (0.47–3.27) 0.672
Age > 65 years 75 1.29 (0.63–2.62) 0.483
5 to 9 Mutated Driver Genes 1.65 (0.55–4.93) 0.373
CKS1B [1q] Amplification 69 1.94 (0.88–4.31) 0.102
ISS RISK Status [Stage 3] 74 2.25 (0.84–6.02) 0.105
10+ Mutated Driver Genes 75 5.02 (1.29–19.54) 0.020
Novel Splicing Top 20% 5.75 (1.60–20.61) 0.007
Bi-Allelic Inactivation TP53 70 7.65 (1.67–35.06) 0.009
Double Hit 7.78 (2.84–21.33) <0.001
Global Loss of Heterozygosity >4.6% 69 11.59(2.38–56.43) 0.002
1 2 3 4 5 6 7 8 9 10
(B)
PFS Multivariate Cox Regression Analysis
Endpoint Subgroup n HR CI p
PFS 5 to 9 Mutated Driver Genes 65 1.66 (0.21–13.04) 0.628
Global Loss of Heterozygosity >4.6% 2.11 (0.32–14.02) 0.438
10+ Mutated Driver Genes 2.50 (0.23–27.16) 0.451
Novel Splicing Top 20% 2.70 (0.57–12.81) 0.211
Double Hit 4.76 (1.38–16.37) 0.013
1 2 3 4 5 6 7 8 9 10
(C)
OS Univariate Cox Regression Analysis
Endpoint Subgroup n HR CI p
OS CKS1B [1q] Amplification 69 1.15 (0.31–4.19) 0.835
Age > 65 years 75 1.30 (0.47–3.59) 0.617
5 to 9 Mutated Driver Genes 1.48 (0.32–6.89) 0.618
ISS RISK Status [Stage 2] 74 1.83 (0.38–8.85) 0.454
10+ Mutated Driver Genes 75 3.26 (0.51–20.89) 0.212
ISS RISK Status [Stage 3] 74 3.30 (0.66–16.38) 0.145
Double Hit 70 7.67 (2.15–27.31) 0.002
Bi-Allelic Inactivation TP53 10.85(2.20–53.39) 0.003
Novel Splicing Top 20% 75 11.14(2.76–45.03) <0.001
Global Loss of Heterozygosity >4.6% 69 13.87(2.77–69.55) 0.001
1 2 3 4 5 6 7 8 9 10
(D)
OS Multivariate Cox Regression Analysis
Endpoint Subgroup n HR CI p
OS Bi-Allelic Inactivation TP53 65 2.17 (0.14–34.43) 0.583
Global Loss of Heterozygosity >4.6% 4.26 (0.36–50.11) 0.249
Novel Splicing Top 20% 6.62 (0.99–44.14) 0.051
1 2 3 4 5 6 7 8 9 10

References

1. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33(3):290–295.
2. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
4. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv 2016;060012.
5. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45(D1):D362–D368.
6. Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics 2015;16(1):224.
7. Therneau T. A Package for Survival Analysis in S. 2013;1–83.
8. Kassambara A, Kosinski M. survminer: Drawing Survival Curves using “ggplot2.”

Supplemental Table Legend [Excel Sheet]

Supplementary Tables 1 [JunctionSeq AS]: Differential exon/splice junction expression analysis using JunctionSeq. List of features found to be significantly differentially expressed at a p-value < .05

Supplementary Tables 2 [DEXSeq AS]: Differential splicing analysis using DEXSeq. List of differentially spliced genes found to be significantly at a p-value < 0.05

Supplementary Tables 3 [SUPPA2 AS]: Differential splicing analysis using SUPPA2. List of differentially spliced genes found to be significantly at a p-value <0 .05

Supplementary Tables 4 [Combined]: Combined results, all genes found to be significant by all analysis p-value < 0.05. Union and intersection of the different tools.

Supplementary Tables 5 [Leafcutter AS]: Differential splicing analysis using leafcutter. List of differentially spliced genes found to be significantly at a p-value < 0.05

Supplementary Tables 6 [LnScore NovelLoci]: List of significant novel splice junctions and the results of lnScore analysis to predict coding potential of possible novel transcripts.

Supplementary Tables 7 [Salmon DESeq2 Gene]: Differential gene expression analysis using expression values from Salmon and DESeq2. List of genes found to be significantly differentially expressed at a p-value < 0.05

Supplementary Tables 8 [Stringtie DESeq2 TX]: Differential gene expression analysis using expression values from StringTie and analyzed using DESeq2. List of genes found to be significantly differentially expressed at a p-value < 0.05

Supplementary Tables 9 [Salmon DESeq2 TX]: Differential transcripts expression analysis using expression values from StringTie and analyzed using DESeq2. List of transcripts found to be significantly differentially expressed at a p-value < 0.05

Supplementary Tables 10 [SUPPA2 TX]: Differential transcripts expression analysis using expression values from SUPPA2. List of transcripts found to be significantly differentially expressed (included) at a p-value < 0.05