Supplementary Information

Recurrent mutations at codon 625 of the splicing factor *SF3B1* in uveal melanoma

J. William Harbour[1], Elisha D. O. Roberson[2], Hima Anbunathan[2], Michael D. Onken[1], Lori A. Worley[1], Anne M. Bowcock[2]

[1]Department of Ophthalmology & Visual Sciences, and [2]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, 63110, U.S.A.

Correspondence should be addressed to Dr. Bowcock (bowcock@wustl.edu) or Dr. Harbour (harbour@wustl.edu).

Supplementary Methods

Supplementary Tables 1-2

Supplementary Figures 1-3

SUPPLEMENTARY METHODS

Patient materials

This study was approved by the Human Studies Committee at Washington University in St. Louis, and informed consent was obtained from all subjects. Clinical and histopathologic annotations data were collected and de-identified for statistical analysis. Where the numbers for individual factors do not add up to 102, the missing data points were not available. Tumor tissue was obtained at the time of treatment of the primary tumor, and RNA and DNA were prepared as described elsewhere[1]. Genomic DNA samples from primary tumors and liver metastases were prepared with the Wizard Genomic DNA Purification kit (Promega, Madison, WI). DNA from blood was isolated using the Quick Gene DNA whole blood kit S (Fugifilm, Tokyo, Japan). RNA from tumors was isolated using the PicoPure kit (including the optional DNase step). RNA samples were converted to cDNA with the High Capacity cDNA Reverse Transcription kit from Applied Biosystems (Applied Biosystems Inc., Foster City, CA) following the manufacturer's protocol.

Molecular classification

RNA samples were converted to cDNA with the High Capacity cDNA Reverse Transcription kit (Applied Biosystems Inc., Foster City, CA) and then pre-amplified for 14 cycles with pooled probes and TaqMan Pre-Amp Master Mix following manufacturer's protocol. Expression of mRNA for individual genes was quantified using the 7900HT Real-Time PCR System with custom TaqMan® Low-Density Arrays and Gene Expression Master Mix (Applied Biosystems Inc., Foster City, CA). The 15-gene molecular classification assay for assignment of tumors to class 1 or class 2 was performed as described elsewhere[2]. Briefly, molecular class assignments were made by entering the 12 ΔC_t values of each sample into the machine learning algorithm GIST 2.3 Support Vector Machine (SVM) (http://bioinformatics.ubc.ca/svm). SVM was trained using a set of 28 well-characterized uveal melanomas of known molecular class and clinical outcome. SVM creates a hyperplane between the training sample groups (here, class 1 and class 2), then places unknown samples on one or the other side of the hyperplane based upon their gene expression profiles. Confidence is measured using a discriminant score, which is inversely proportional to the proximity of the sample to the hyperplane.

Exome capture and DNA sequencing

Genomic DNA paired-end libraries were created and exons were captured as previously described[3]. Captured genomic DNA was sequenced with either the Illumina Genome Analyzer II (GAIIx) for 76-cycles or the Illumina HiSeq for 101-cycles.

Data were received as SCARF formatted files. Indexed lanes had the detected index sequence and associated qualities concatenated onto the 3' end of read 1. Indexes were identified from the appropriate number of terminal bases, allowing for a single 3' mismatch to account for sequencing error and a single missing 3' base to account for different numbers of index read cycles. 'N' bases were hard clipped from the 3' end of each read. N bases on the 5' end had the quality floored to an appropriate minimum for the quality scale offset rather than hard clipping in order to avoid duplicate marking issues. Low quality 3' bases with a quality score of less than 3 were hard

clipped. Contaminating adapter sequence was identified by ungapped local alignment with adapter sequence, and was hard clipped if terminal 3' read sequence showed greater than 87% identity over more than five bases at the 3' end of reads. DNA-SEQ data was outputted with the phred/Sanger (+33) quality offset. All output files were appropriate transformed from SCARF to FASTQ format.

FASTQ files were aligned to a repeat masked human genome (hg19, repeats 'N' masked) with the BWA software (v5.9rc1)[4]. Output SAM files were cleaned, converted to BAM files, coordinate sorted and optical duplicates were marked with Picard tools (v1.53+; CleanSam, SamFormatConverter, SortSam, MarkDuplicates). Potential variant sites were called with samtools pileup (v0.1.8)[5], and filtered to retain those with a SNP quality of at least 20. All variants were stored in a relational database (MySQL v5.0.75). Known polymorphisms (dbSNP 129, dbSNP132 common, thousand genomes 2010 data release) and control variants in eight previously sequenced HapMap individuals[6] were identified. Variants were excluded if they were present in DNA from patient blood, or if they were known SNPs from dbSNP129 SNPs present in dbSNP132, 1000 genomes variants (Nov. 2010 release), or eight HapMap individuals. Remaining data were filtered for a variant quality score ≥ 175 by samtools pileup, a read depth ≥ 30, and no more than 65% of reads with the allele of the reference sequence. The filtered variant list was annotated for coding changes using SIFT (http://sift.jcvi.org/)[7]. Splice site mutations were annotated using a MySQL table of all RefSeq gene splice donor and splice acceptor locations. Mutations were considered deleterious if they disrupted a splice site, were predicted by SIFT to be damaging, created a termination read-through or created a premature termination codon. Several thousand variants were identified in this manner, but the vast majority of these were found in only one tumor sample. Since our main interest was in identifying potential driver mutations in uveal melanoma, we filtered these variants to those in genes predicted to be "cancer drivers" in Sanger (Cosmic: http://www.sanger.ac.uk/genetics/CGP/Census/), and then validated them by Sanger sequencing of DNA from tumor and matched blood samples. Only two genes were confirmed to harbor predicted deleterious somatic mutations in three or more tumors: *GNAQ* (hg19 chr9:g.80409488T>G, p.Q209P; hg19 chr9:g.80409488T>A, p.Q209L) and *SF3B1* (hg19 chr2:g.198267484G>A, p.R625C).

Mutation Validation

Oligonucleotide primers were designed from intronic sequences to amplify exons 12 to 15 of *SF3B1* by PCR. Genomic DNA of tumor and blood from the same patient were subjected to PCR amplification and re-sequencing as described elsewhere[3]. Oligonucleotide primer sequences are available upon request.

Microarray expression profiling

We analyzed the gene expression profiles performed using Illumina Ref8 Bead Arrays on five *SF3B1*-mutant and six *SF3B1*-wildtype class 1 tumors as previously described[8]. Data were subjected to cubic spline normalization and background subtraction using BeadStation software. The data were analyzed for differentially expressed transcripts using Significance Analysis of Microarrays (http://www-stat.stanford.edu/~tibs/SAM/) and a false discovery rate (FDR) ≤ 5%.

RNA-seq alignment
Raw data processing from SCARF file to FASTQ file was performed as above for exome capture. FASTQ files were aligned human hg19 genome (N-repeat masked) using tophat with a RefSeq GTF file to guide alignment (v1.4.1, bowtie v0.12.7; options: --solexa1.3-quals, --coverage-search, --microexon-search, --mate-inner-dist 25, --mate-std-dev 75, --GTF, --transcriptome-index).

Splice donor or acceptor retention
Rather than testing for retention of the entire intron, we focused on splice donor and splice acceptor retention. All splice donor and acceptor locations from RefSeq were loaded into a MySQL table (v5.0.75). Any donor or acceptor that overlapped with an alternate exon or UTR was removed, to prevent detecting retention differences that could represent differential isoform expression. The remaining sites were used to create a BED file. A count of how many reads from each sample's RNA-SEQ BAM file overlapped each feature of the BED file was calculated using the bed tools coverageBed program in conjunction with samtools (samtools v0.1.18 [r982:295], coverageBed v2.15.0; pseudocode: samtools view -uf 0x3 CurrSampleRNAseq.bam | coverageBed -split stdin -b SpliceSites.bed > currSampleSpliceCounts.txt). Counts for each splice site were merged into a single matrix in the R programming language (R v2.15.1, RStudio v0.96.330). The matrix was filtered to retain only sites with at least one read in any of the RNA-SEQ samples. DESeq (DESeq v1.8.3, Biobase v2.16.0, locfit v1.5.8, BioGenerics v0.2.0) was used to determine differential splice donor or splice acceptor retention between *SF3B1* mutants (n=3) and *SF3B1* wild-type (n=5). The matrix was converted to a DESeq count dataset (newCountDataSet) with the mutant / non-mutant status indicated in a vector of factors. Size factors and dispersions were estimated for the counts (estimateSizeFactors, estimateDispersions). Significance of counts between groups was determined using the nbinomTest function. After false-discovery rate correction, no splice donor or acceptor demonstrated evidence for statistically significant retention with the *SF3B1* mutation. While this does not rule out an effect on specific transcripts, we did not find evidence for global intron splicing retention.

Array comparative genomic hybridization
Genome-wide DNA copy number data obtained by array comparative genomic hybridization (aCGH) were previously described[8] and were re-analyzed here with respect to *SF3B1* mutation status. A significant DNA copy number deviation was defined as a log2 average raw ratio of > |0.5| across at least three consecutive probes. *SF3B1*-wildtype and –mutant samples were compared for significant differences in copy number gains and losses using Partek Genomics Suite software (v6.6).

Statistical analysis
Fisher's exact test was used to assess the significance of association between two categorical variables. Student's two tailed t-test was used to assess the significance of association between two continuous variables. Kaplan-Meier analysis was used to assess time-dependent association between clinical, pathologic and genetic variables and clinical outcome. All statistical analyses were performed with MedCalc software (v12.3.0.0).

References

1. Onken, M.D. *et al.* Loss of heterozygosity of chromosome 3 detected with single nucleotide polymorphisms is superior to monosomy 3 for predicting metastasis in uveal melanoma. *Clin Cancer Res* 13, 2923-7 (2007).
2. Onken, M.D., Worley, L.A., Tuscan, M.D. & Harbour, J.W. An accurate, clinically feasible multi-gene expression assay for predicting metastasis in uveal melanoma. *J Mol Diagn* 12, 461-8 (2010).
3. Harbour, J.W. *et al.* Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science* 330, 1410-3 (2010).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-60 (2009).
5. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987-93 (2011).
6. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272-276 (2009).
7. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31, 3812-4 (2003).
8. Worley, L.A. *et al.* Transcriptomic versus chromosomal prognostic markers and clinical outcome in uveal melanoma *Clin Cancer Res* 13, 1466-71 (2007).

SUPPLEMENTARY TABLES

Supplementary Table 1. Summary of primary uveal melanomas harboring SF3B1 mutations.

Sample	*SF3B1* nucleotide change	Predicted protein alteration	Age at diagnosis	Sex	Ciliary body involvement	Tumor diameter	Tumor thickness	Extraocular tumor extension	Epithelioid cell type	Treatment for primary tumor	Follow-up (months)	Death from metastasis
MM010	CGT>TGT	R625C	42	Male	No	17	10	No	No	Enucleation	142	Yes
MM028	CGT>CAT	R625H	48	Male	Yes	20	12	No	No	Enucleation	105	No
MM032	CGT>GGT	R625G	63	Female	Yes	22	11	No	No	Enucleation	104	No
MM049	CGT>CAT	R625H	65	Female	Yes	20	12	No	No	Enucleation	100	No
MM061	CGT>CAT	R625H	66	Male	No	11	4	Yes	No	Enucleation	95	Yes
MM064	CGT>CAT	R625H	62	Female	No	18	6	No	No	Enucleation	39	No
MM065	CGT>TGT	R625C	45	Female	Yes	24	7	No	No	Enucleation	32	Yes
MM066	CGT>TGT	R625C	47	Male	Yes	22	9	No	No	Enucleation	76	No
MM101	CGT>CAT	R625H	66	Female	No	15	6	No	No	Enucleation	11	No
MM131	CGT>CTT	R625L	60	Male	Yes	13	15	Yes	No	Enucleation	0	No
MM133	CGT>TGT	R625C	55	Female	No	20	15	No	Yes	Enucleation	8	No
MM134	CGT>TGT	R625C	57	Female	Yes	19	9	No	No	Enucleation	36	Yes
MM167	CGT>CAT	R625H	65	Female	Yes	N/A	N/A	Yes	No	Enucleation	3	No
MM170	CGT>CAT	R625H	76	Male	No	9	2	Yes	No	Enucleation	1	No
MM178	CGT>CAT	R625H	66	Male	No	N/A	N/A	Yes	No	Enucleation	2	No
NB119	CGT>CAT	R625H	69	Female	Yes	16	6	No	No	Brachytherapy	27	No
NB122	CGT>CAT	R625H	48	Female	No	14	5	No	No	Brachytherapy	51	No
NB125	CGT>CAT	R625H	16	Female	No	16	8	No	No	Brachytherapy	29	No
NB126	CGT>CAT	R625H	35	Female	No	17	8	No	No	Brachytherapy	33	No

N/A, not available

Supplementary Table 2. Differentially expressed genes in *SF3B1*-mutant versus *SF3B1*-wildtype class 1 uveal melanomas[1].

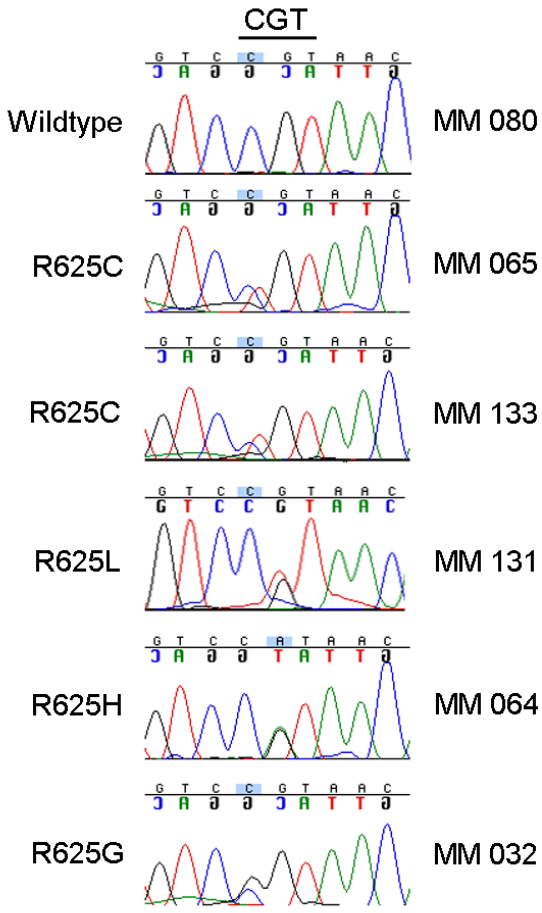
Gene Symbol	Gene Name	Gene Ontology Categories[1]	Fold Change
SERINC3	Serine incorporator 3	Induction of apoptosis	8.1
PIH1D1	PIH1 domain containing 1	Box C/D snoRNP assembly	1.4
ANKHD1	Ankyrin repeat and KH domain containing 1	RNA binding	2.4
NDUFB8	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 8	Mitochondrial electron transport	-1.7
ELP2	Elongation protein 2 homolog	Transcription elongation from RNA polymerase II promoter	-2.5
CCL28	Chemokine (C-C motif) ligand 28	Chemotaxis	-5.0
SEPT2	Septin 2	GTP binding, cell division, neuron projection regulation	-2.0
MUT	Methylmalonyl CoA mutase	Cellular lipid metabolic process, post-embryonic development	-2.0
NRD1	Nardilysin (N-arginine dibasic convertase)	Cell migration, cell proliferation	-2.0
DOCK7	Dedicator of cytokinesis 7	GTP binding, neuron projection regulation	-2.0

[1]Data collected on Illumina Ref8 Bead Arrays and analyzed using Significance Analysis of Microarrays (SAM) and a false discovery rate of <5%.

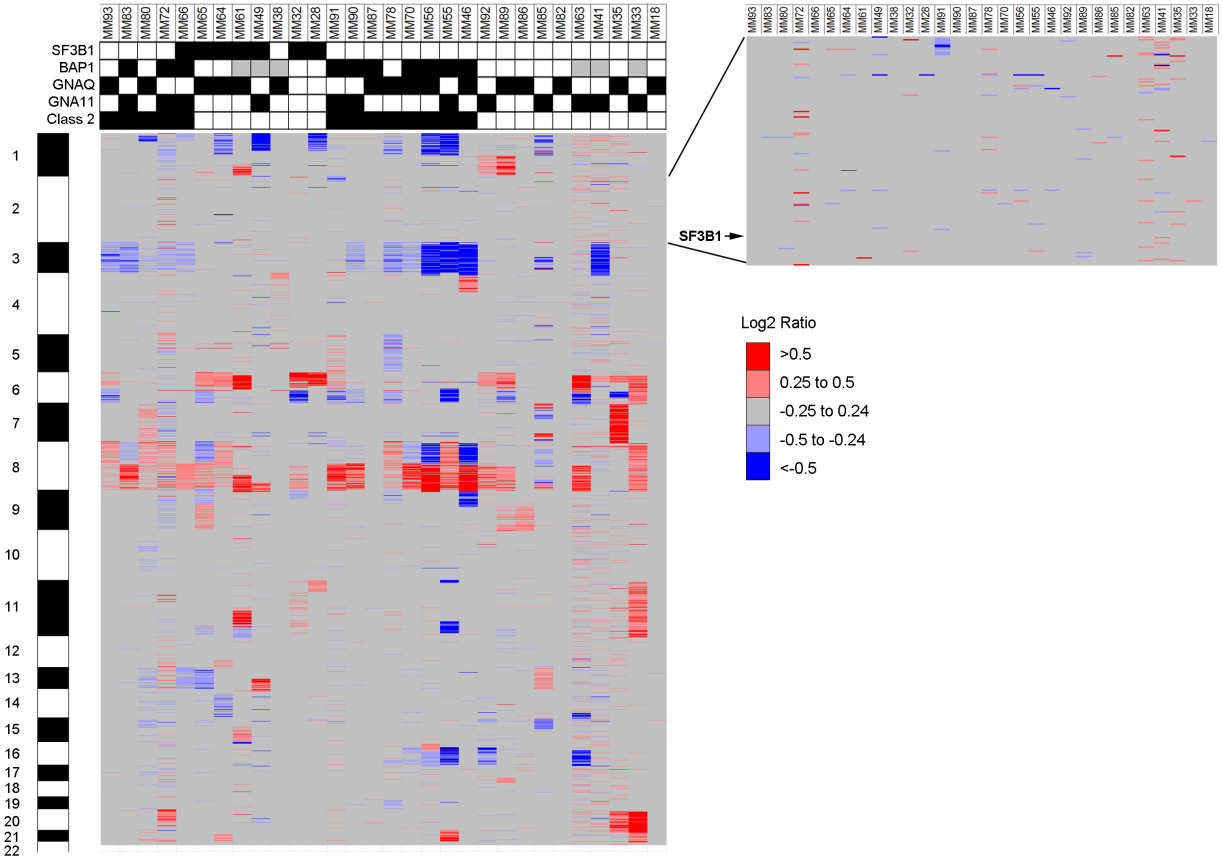
[2]Based on UniProt-GOA

SUPPLEMENTARY FIGURES

Supplementary Figure 1. Representative Sanger sequence traces of *SF3B1*-wildtype and *SF3B1*-mutant tumors.



Supplementary Figure 2. Whole genome chromosome copy number analysis using array comparative genomic hybridization (aCGH) in 30 uveal melanomas. Tumor samples included 7 *SF3B1*-mutant and 23 *SF3B1*–wildtype tumors. For the header, black squares indicate class 2 transcriptomic signature or presence of indicated mutation, white indicates class 1 signature or absence of indicated mutation, and grey indicates data not available. For the heatmap, grey indicates 2N DNA content, red indicates greater than 2N DNA content, and blue indicates less than 2N DNA content. No copy number gains or losses were found to be associated with *SF3B1* mutation status, and no gains or losses were noted at the *SF3B1* locus at 2q33.1 (expanded section).



Supplementary Figure 3. Significance Analysis of Microarrays (SAM) plot comparing the gene expression profiles of five *SF3B1*-mutant and six *SF3B1*-wildtype class 1 uveal melanomas. Red dots indicate genes that are up-regulated and green dots indicate genes that are down-regulated in *SF3B1*-mutant tumors at a false discovery rate <5%.