Supplemental Information for

Pan-cancer analysis identifies mutations in *SUGP1* that recapitulate mutant
SF3B1 splicing dysregulation

Zhaoqi Liu[a,b,c,1], Jian Zhang[d,1], Yiwei Sun[a,c], Tomin E. Perea-Chamblee[a,b,c], James L. Manley[d,2], Raul Rabadan[a,b,c,2]

[a]Program for Mathematical Genomics
[b]Department of Systems Biology
[c]Department of Biomedical Informatics, Columbia University, New York, NY 10032
[d]Department of Biological Sciences, Columbia University, New York, NY 10027

[1]Z.L. and J.Z. contributed equally to this work.

[2]To whom correspondence may be addressed.

Email: jlm2@columbia.edu or rr2579@cumc.columbia.edu

This PDF file includes:

Figures S1 to S5 and Legends

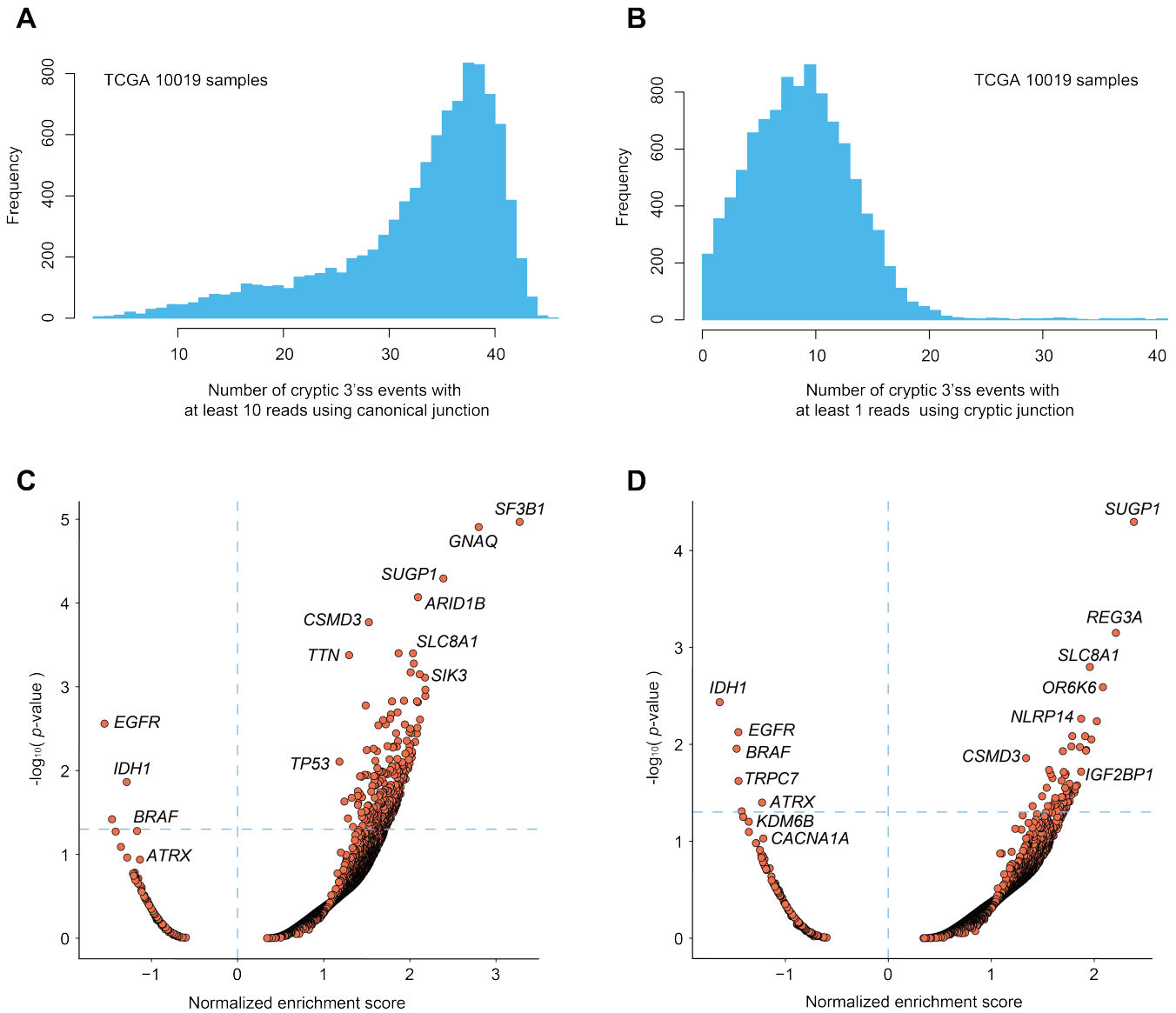## Supplemental Figures

A

B

C

D

Fig. S1. Identification of alternative 3'ss splicing associated genetic variants. (*A*, *B*) Distributions of expression level of the 47 cryptic 3'ss events in 10019 TCGA samples. X-axis denotes the number of cryptic 3'ss events with at least 10 reads using canonical junction (*A*), and the number of cryptic 3'ss events with at least 1 read using cryptic junction (*B*). The Y-axis denotes the frequency (number of TCGA samples). (*C*, *D*) Volcano plot representations of genes correlated with cryptic 3'ss usage. The normalized enrichment score (horizontal axis) and -log10 p-value (vertical axis) are derived from pre-ranked enrichment test. Analysis were performed on all TCGA samples (*C*) and SF3B1 non-mutated samples (*D*).
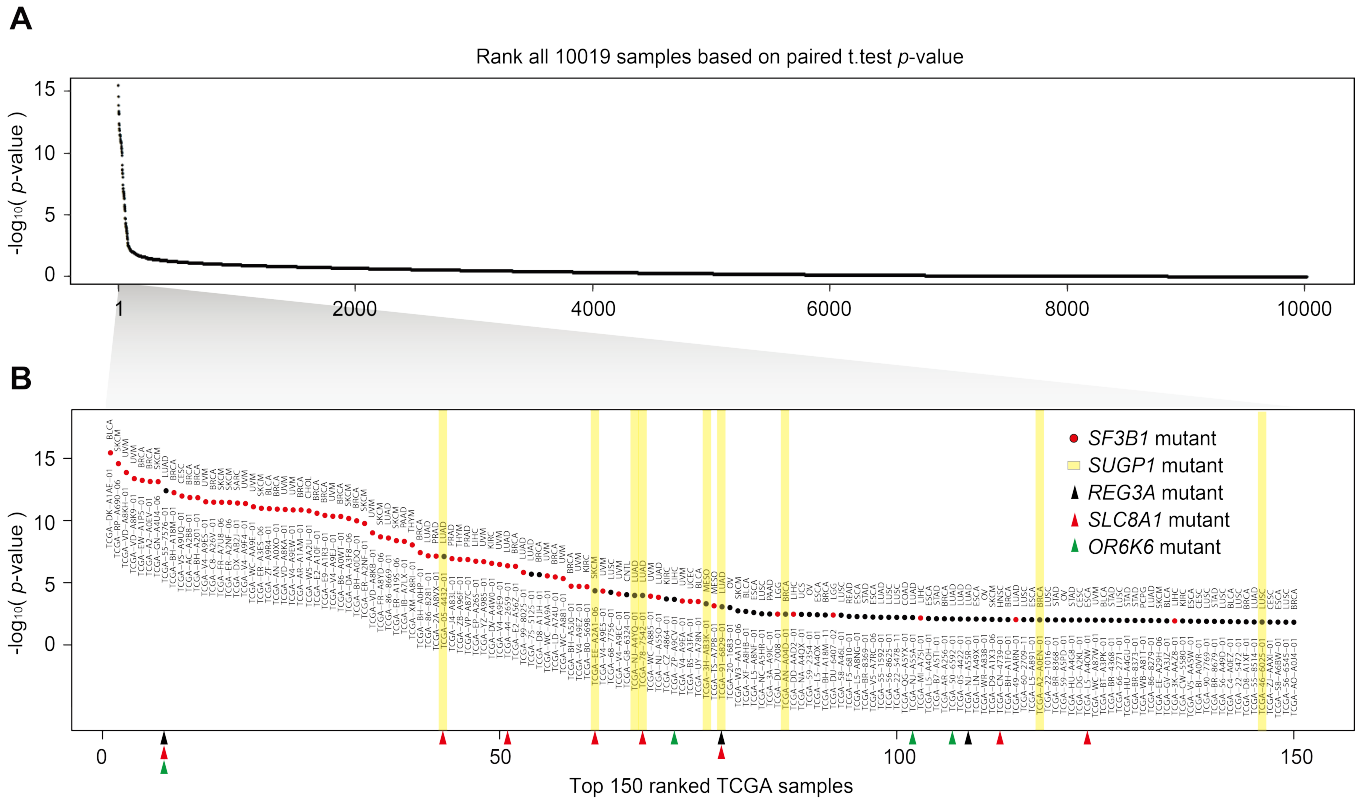
Fig. S2. Identification of genetic variants associated with cryptic 3'ss usage. (*A*) Ranking of all TCGA samples based on paired t.test *p*-value, which indicated the extent of cryptic 3'ss usage. (*B*) Top 150 ranked TCGA samples from (*A*). *SF3B1* mutants are indicated as red dots, *SUGP1* mutants are colored as yellow bars. *REG3A*, *SLC8A1* and *OR6K6* mutants were also shown with arrow of different color.
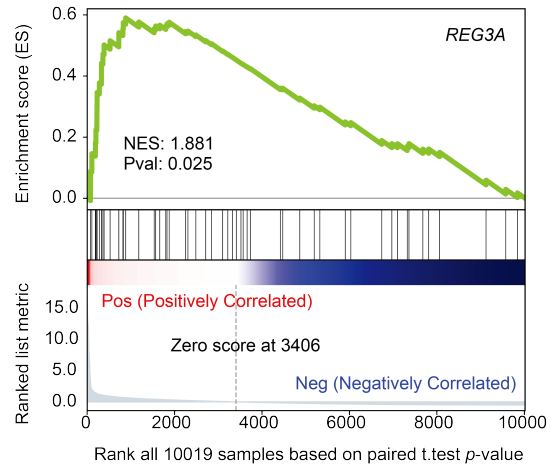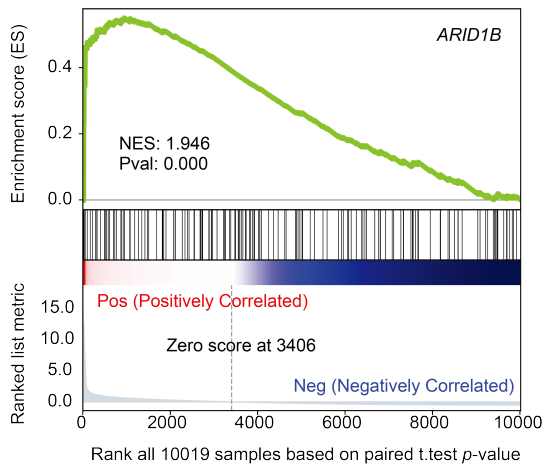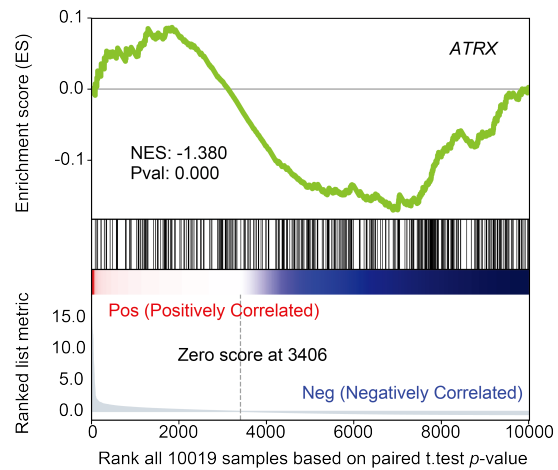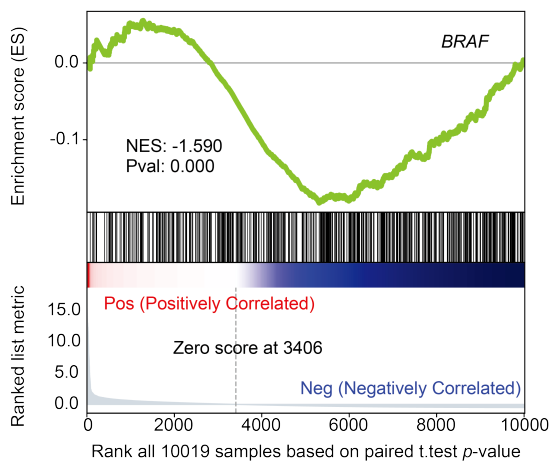
A

NES: 1.946
Pval: 0.000

ARID1B

NES: 1.881
Pval: 0.025

REG3A

B

NES: -1.644
Pval: 0.000

IDH1

NES: -1.586
Pval: 0.000

EGFR

NES: -1.590
Pval: 0.000

BRAF

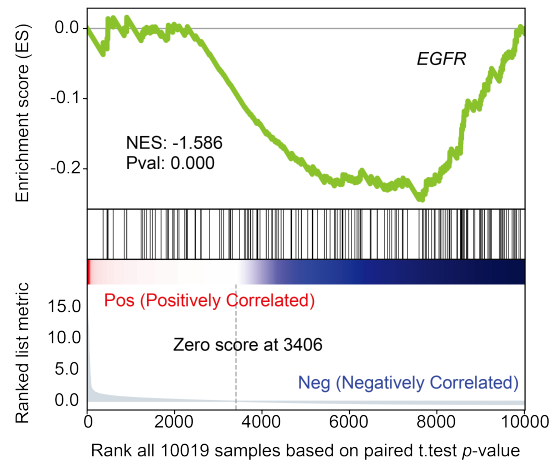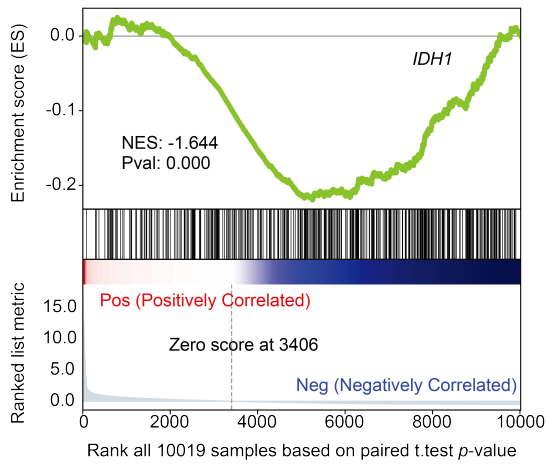NES: -1.380
Pval: 0.000

ATRX

Fig. S3. Pre-ranked enrichment analysis showing top gene mutations that are positively (*A*) or negatively (*B*) associated with alternative 3'ss usage.

Fig. S4. (*A*) Hierarchical clustering and heatmap analysis of the usage of 47 cryptic 3'ss events between top ranked *SF3B1* mutant, *SUGP1* mutant and WT samples by Pearson correlation. Rows and columns represent cryptic 3'ss events and samples, respectively. Values in the matrix represent raw Percent-Spliced-In. (*B*) Principal component analysis of differential 3' splice sites (3'ss) usage between samples from (*A*). The PSI matrix from (*A*) was used as input for this analysis. (*C*) Local PPI network community shows proteins that potentially contribute to 3'ss switch by interacting with SF3B1. Node size is proportional to the -log$_{10}$(*p*-value) from the paired t.test. One step neighbors of SF3B1 are colored as red nodes.
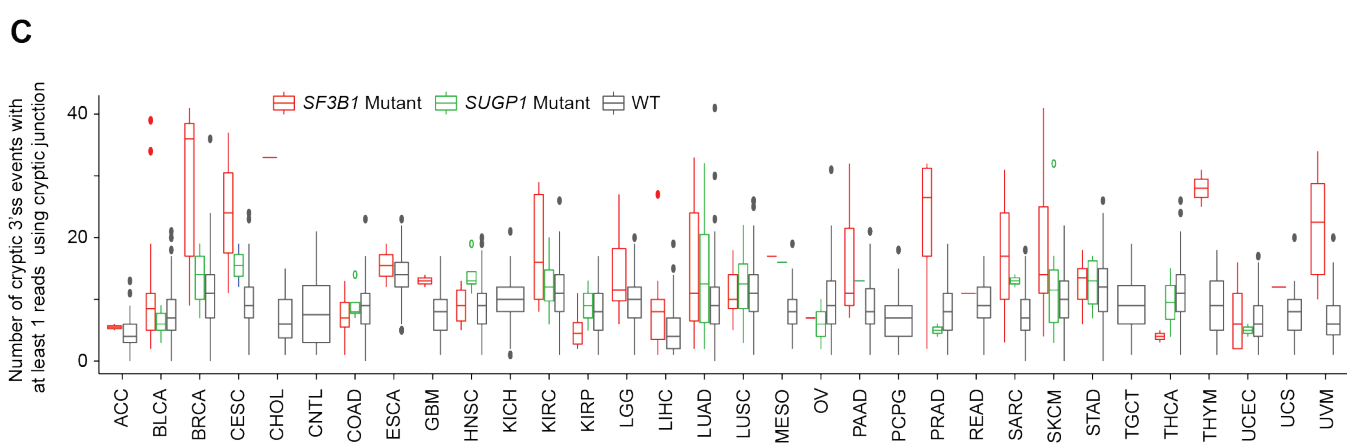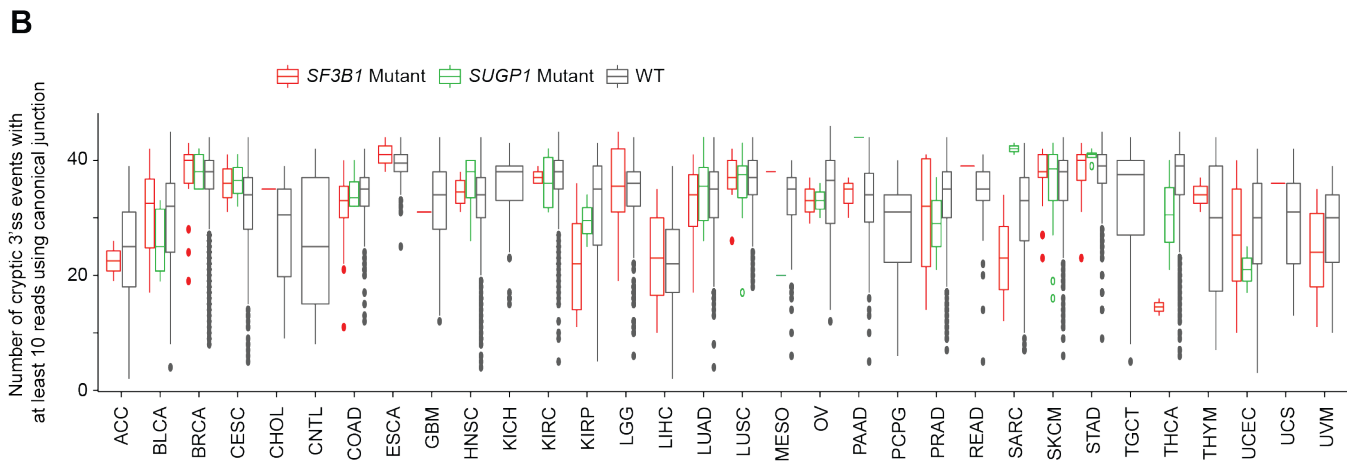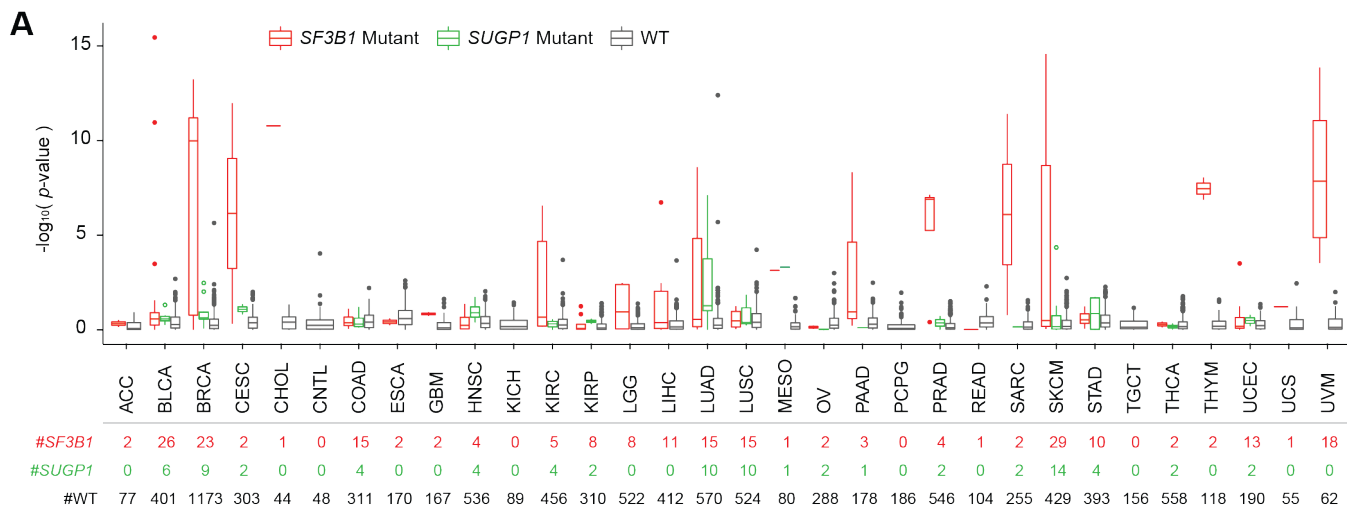
Fig. S5. (*A*) Boxplots show distribution of cryptic 3'ss associations on each TCGA cancer type. *SF3B1* and *SUGP1* mutants were plotted separately. Numbers in the table below indicates the number of mutant samples present in each cancer type. (*B-C*) Boxplots show distributions of expression level of the 47 cryptic 3'ss events on each TCGA cancer types. *SF3B1* and *SUGP1* mutants were plotted separately. Y-axis denotes the number of cryptic 3'ss events with at least 10 reads using canonical junction in (*B*), and the number of cryptic 3'ss events with at least 1 reads using cryptic junction in (*C*).