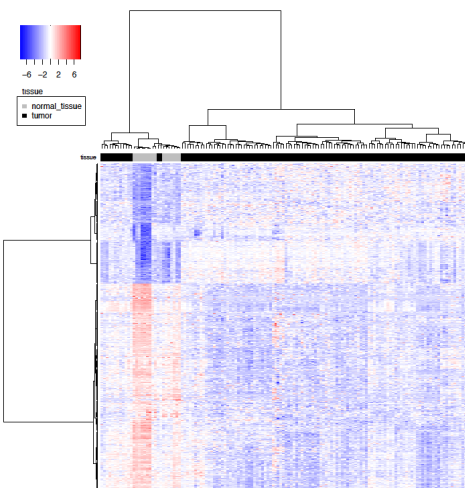


Supplementary Figure 1. A flowchart of fusion filtering

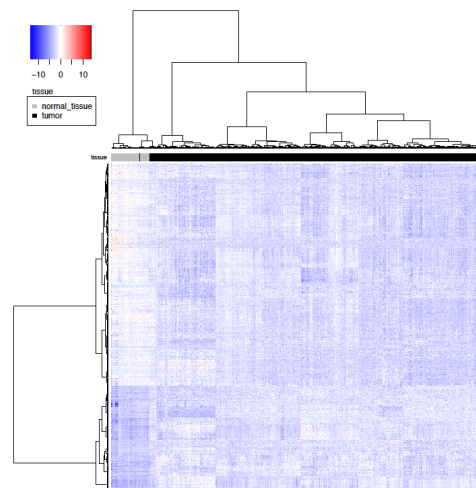
To obtain a bona-fide fusion, fusion filtering was performed based on the number of discordant read pairs, perfect match junction spanning reads, gene homology, transcript allele fraction, and partner gene variety. By using RNA seq data obtained from normal samples, tumor-specific fusions were extracted.

BLCA



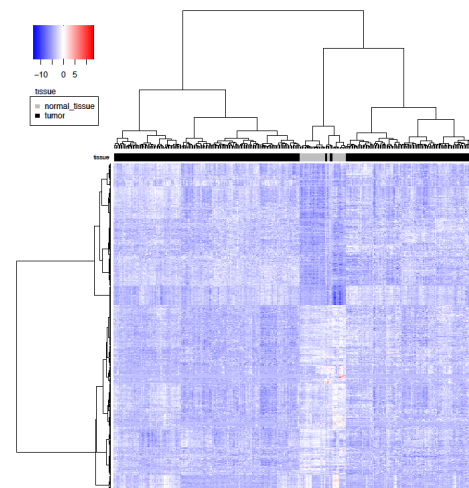
2268 genes

BRCA



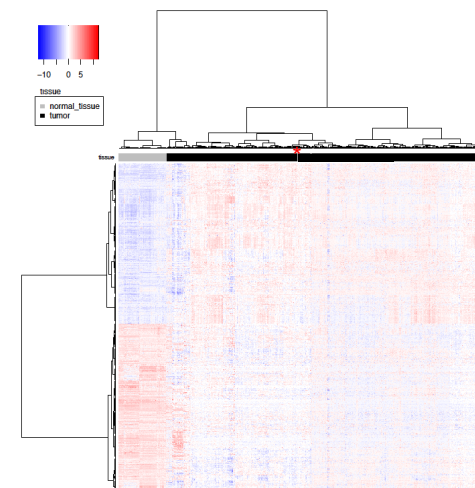
4572 genes

HNSC



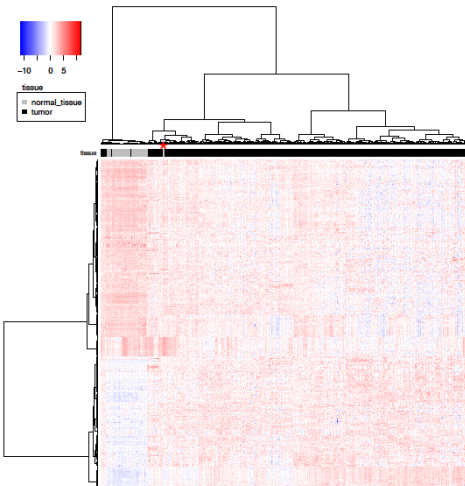
3472 genes

KIRC



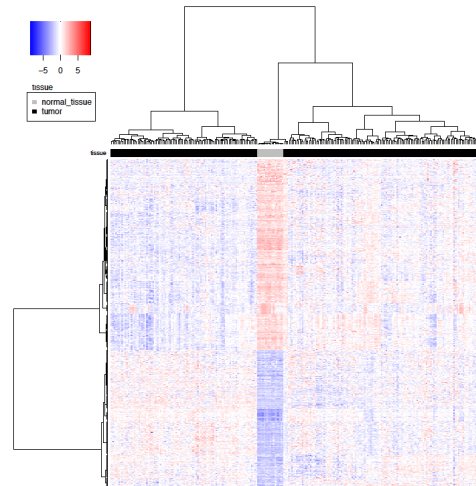
3787 genes

LUAD



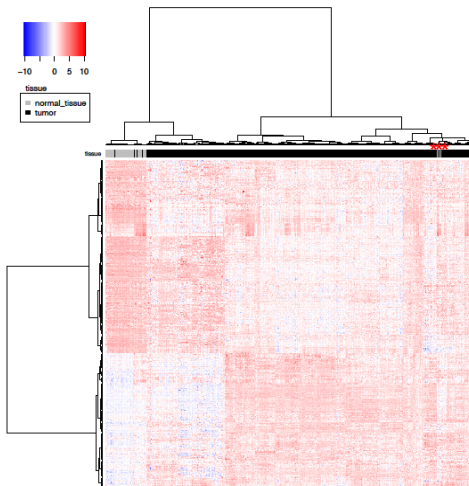
4602 genes

LUSC



5353 genes

THCA



2933 genes

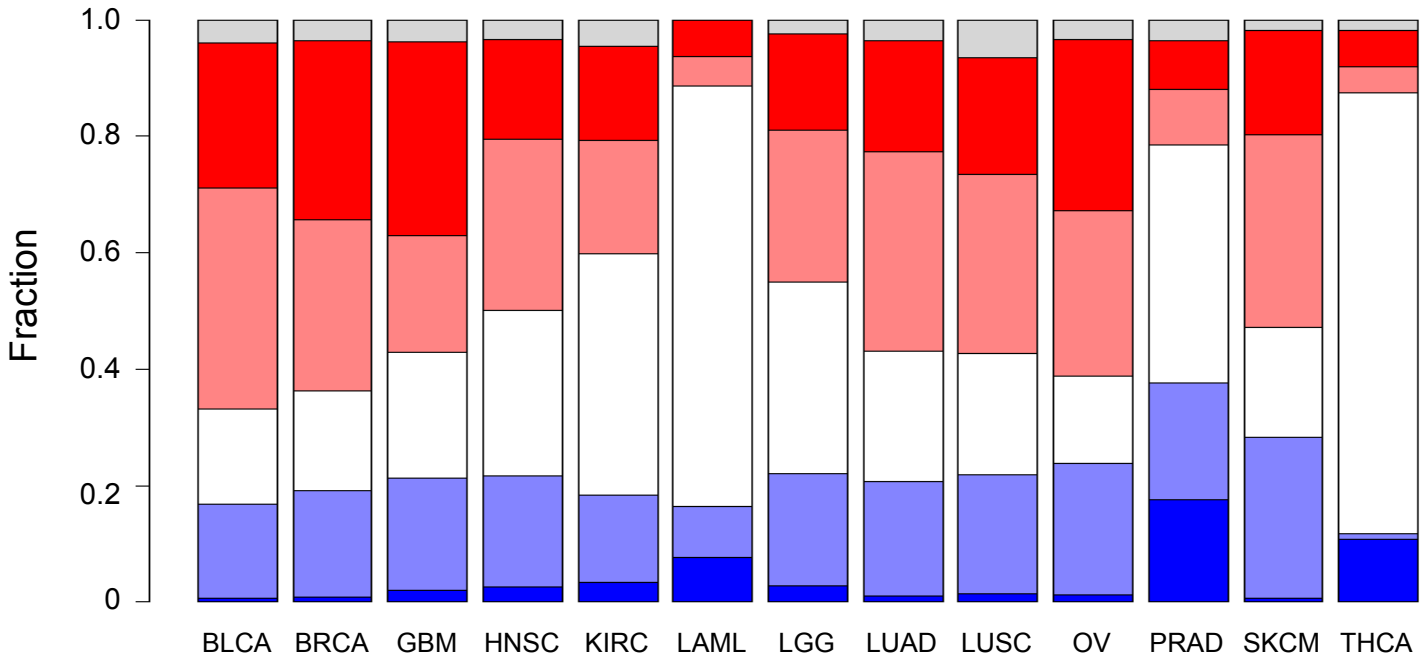
* Five normal samples located in tumor cluster

Supplementary Figure 2. Detection of normal samples with a high likelihood of tumor cell contamination.

Supervised clustering was performed by using different expressed gene between tumor and normal samples. Five normal samples (red asterisk) were clustered into tumor cluster.

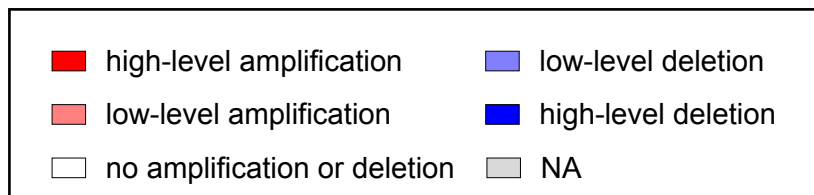
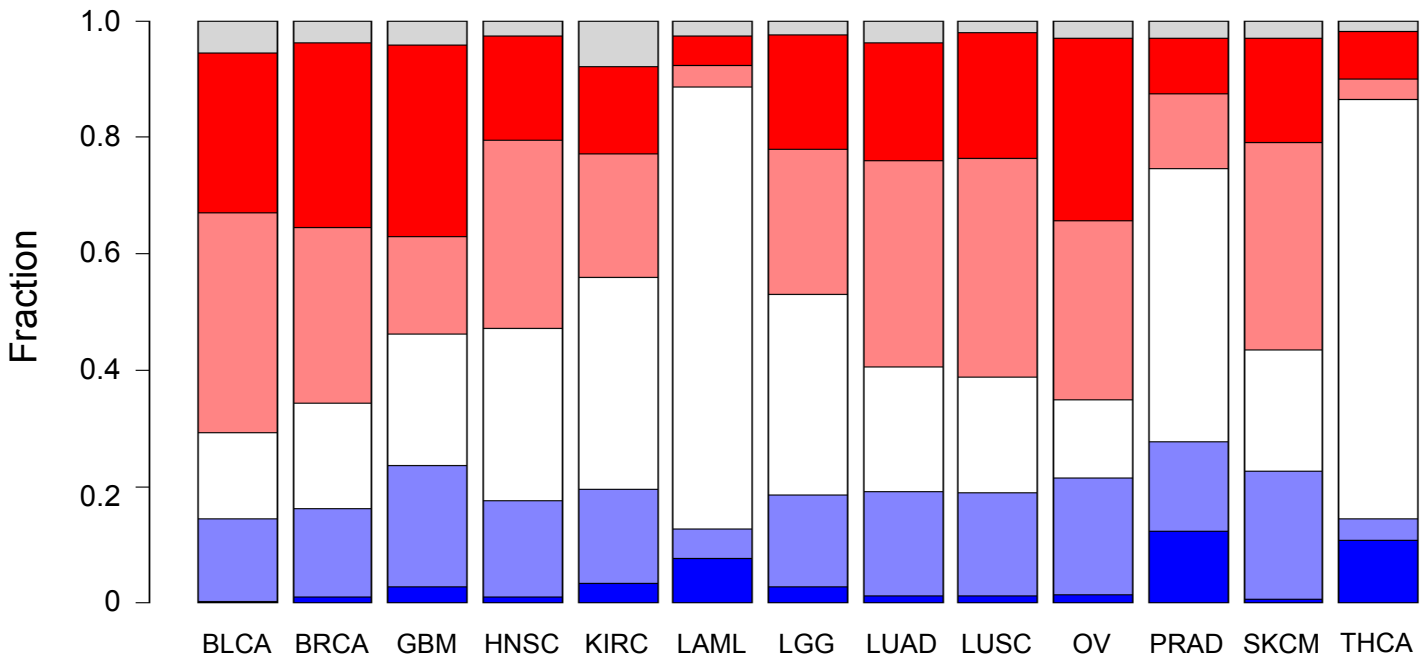
A

Gene A



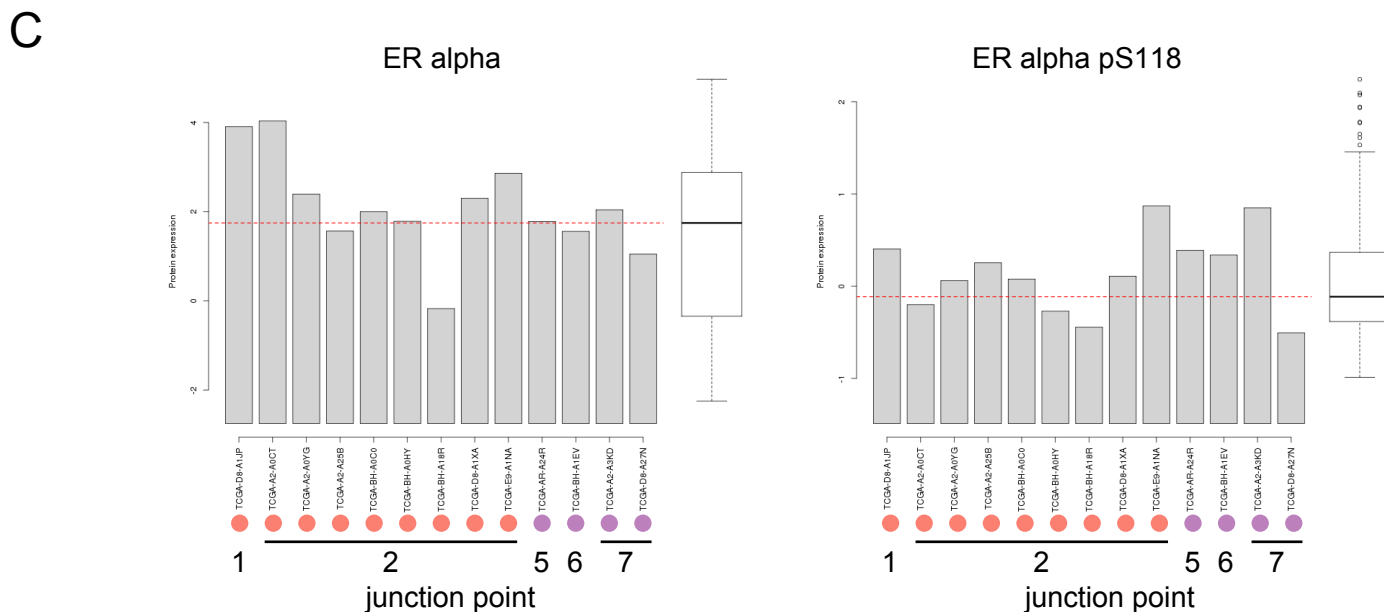
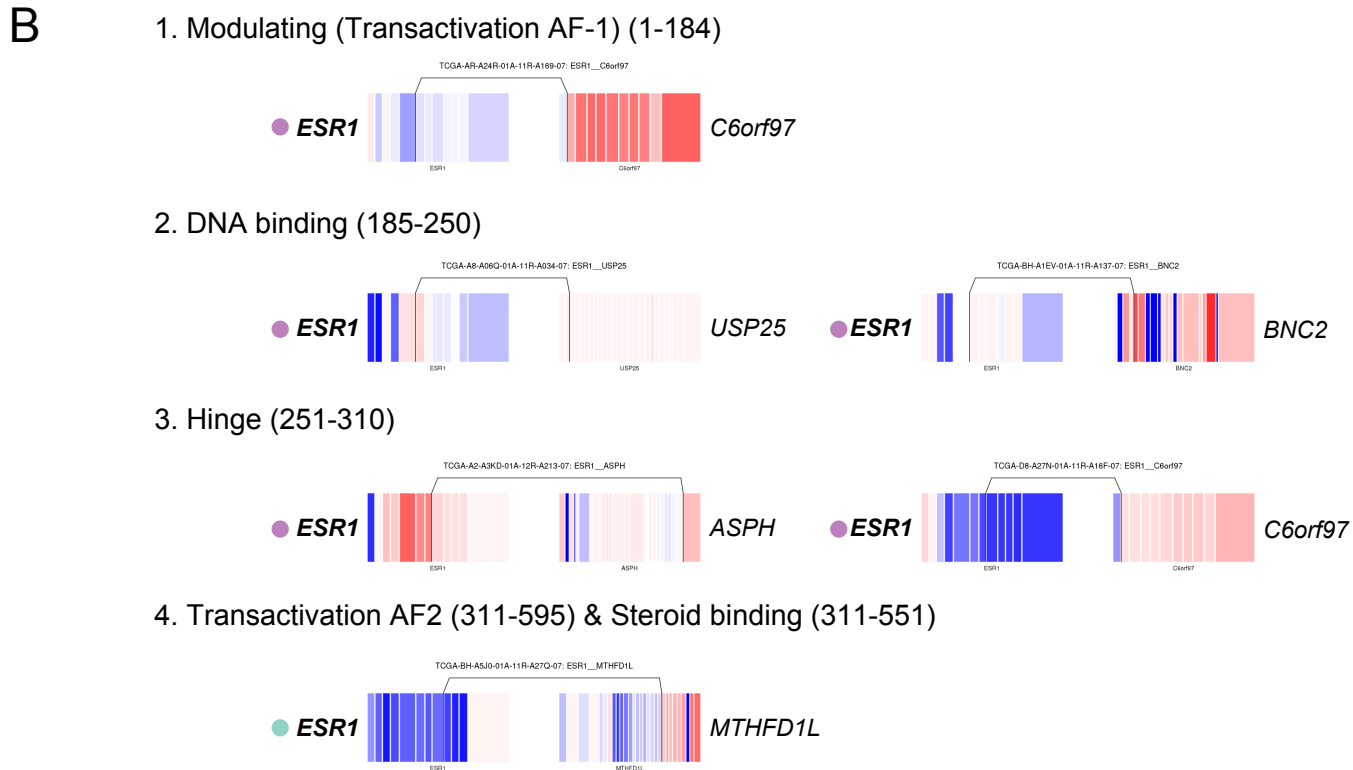
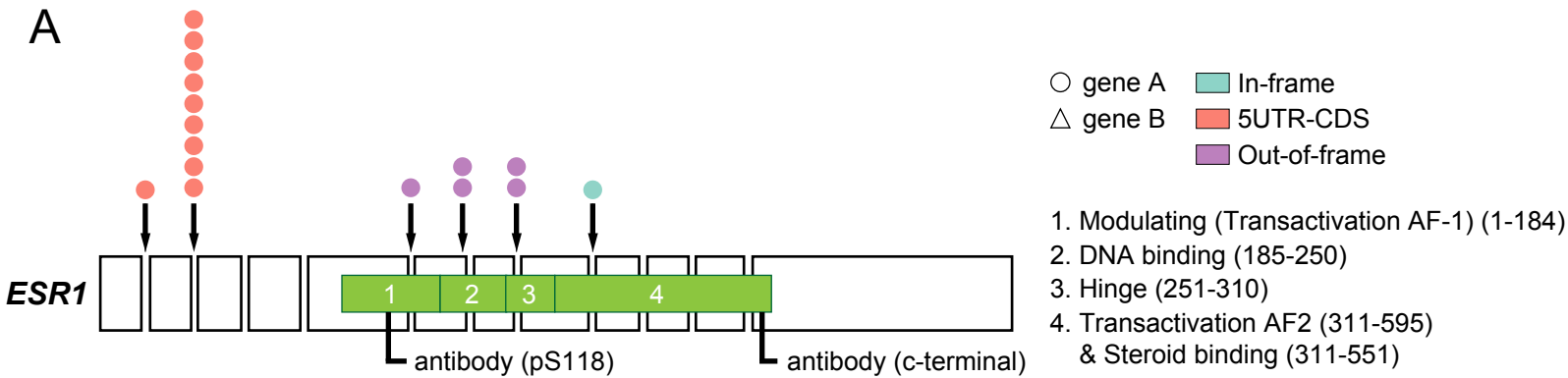
B

Gene B



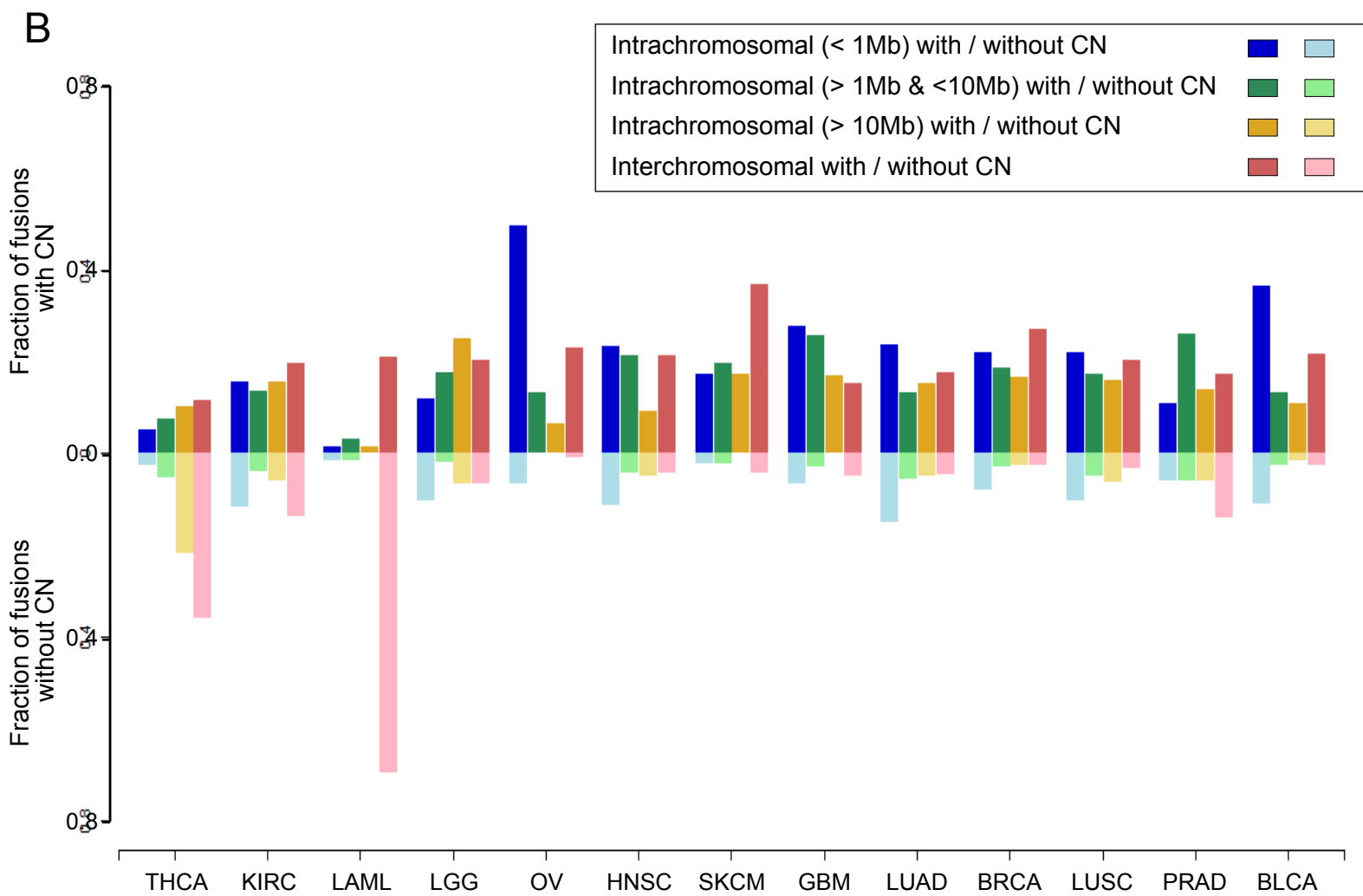
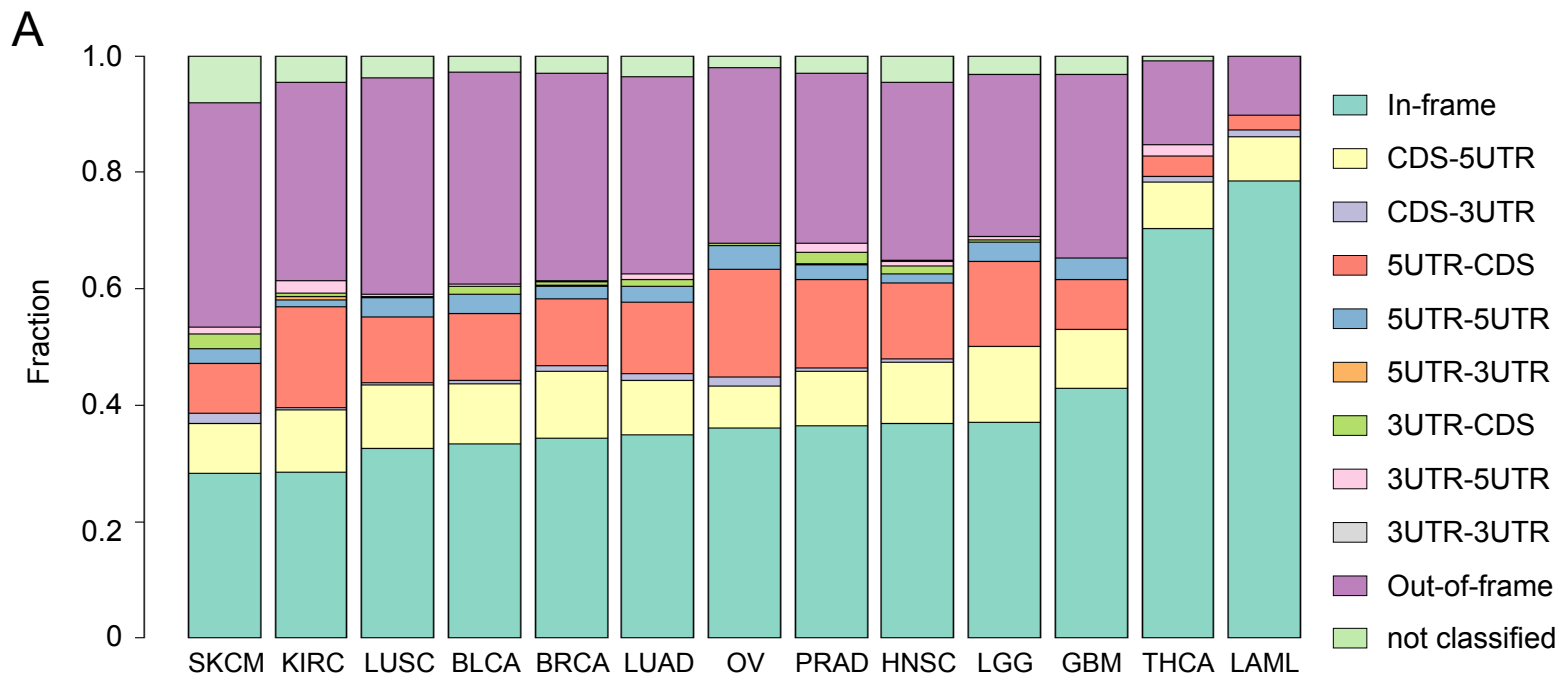
Supplementary Figure 3 Copy number status of each gene for 13 tumor types

Bar plots represent the fraction of five types of copy number status for each tumor type. Gene-level copy number alteration was identified by applying GISTIC 2.0 to TCGA level 3 copy number data for each tumor type. A few genes constituting of fusions were not available in GISTIC 2.0 due to incompatibility of gene symbols between fusion and copy number data and these genes were annotated as "NA" .



Supplementary Figure 4. The details of *ESR1* fusions in breast cancer.

(A) Dot plots with *ESR1* transcript structure demonstrate the frequency of *ESR1* fusions and junction points for each fusion. (B) Exon expression plots show the Z-normalized exon expression level for six *ESR1* fusions located on functional domain (C) Bar plots with box-whisker plots represent RPPA protein expression level for ER alpha (left) and ER alpha pS18 in the overlapped breast cancer samples between fusion and RPPA data sets. Box-Whisker plot represent the mean of protein expression in *ESR1* fusion negative samples.



Supplementary Figure 5. The distribution of fusion transcripts inferred as “in-frame”

(A) Bar plots show the fraction of in-frame fusions. Tumor types were sorted by the fraction of in-frame fusion. (B) Each bar represents the fraction of eight types of in-frame fusions. Tumor types were sorted by the fraction of samples in which at least a single fusion transcript was detected per tumor type.

BLCA (n = 94)

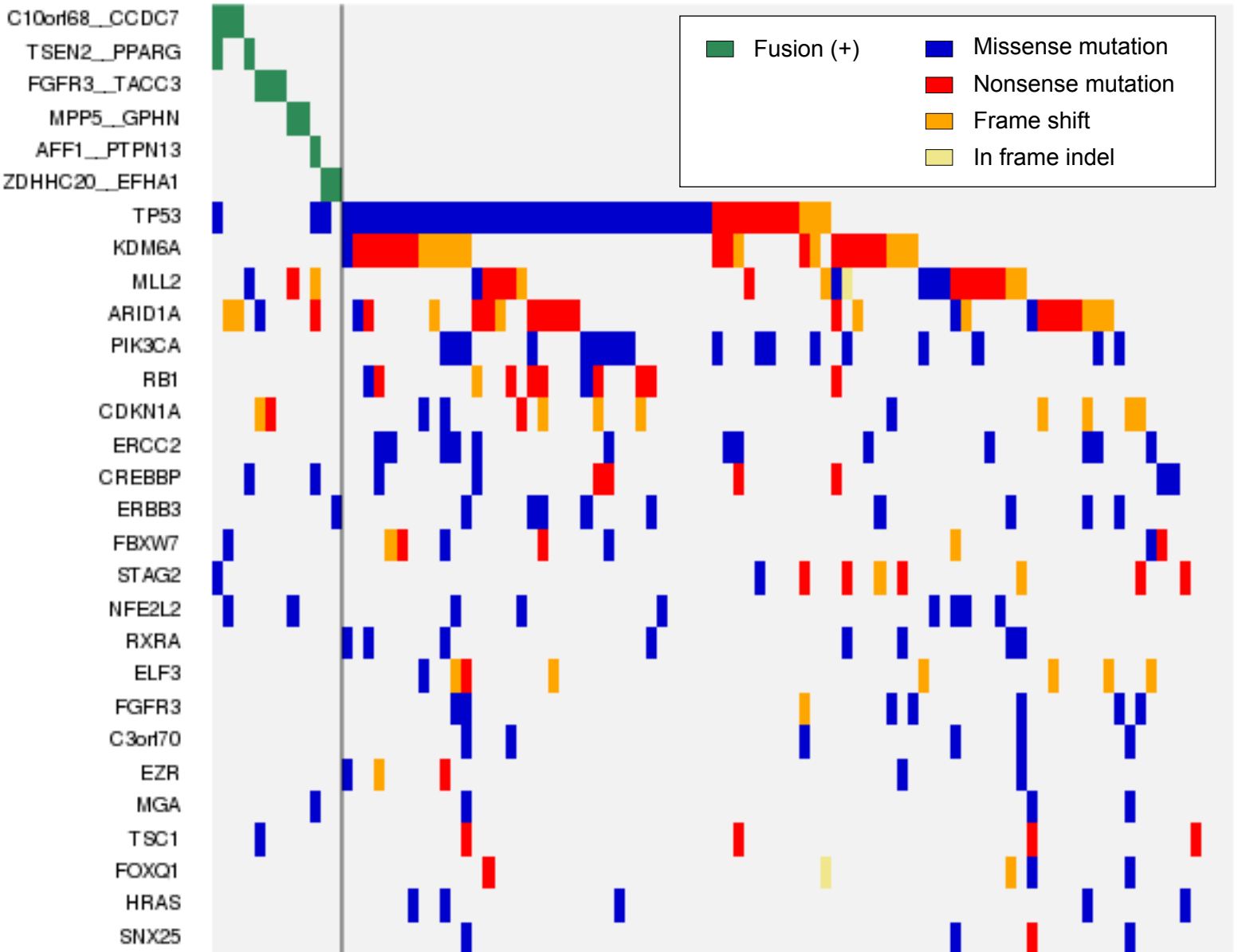
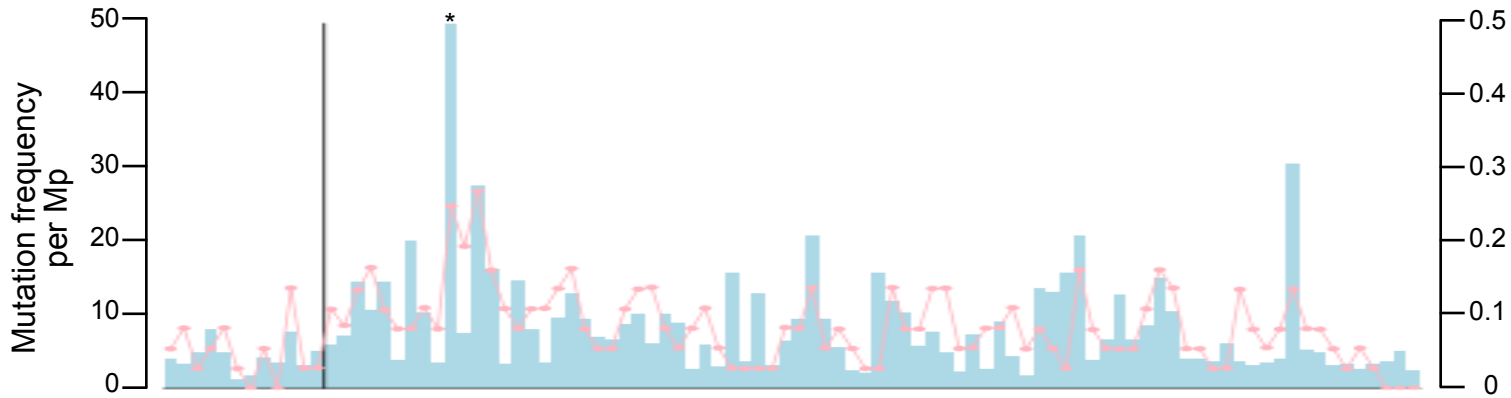
* (n>50)

mutation (left axis)

significant mutation (right axis)

Welch's t-test, p = 3.6e-05

Welch's t-test, p = 0.0067



Supplementary Figure 6. Relationship between fusion and mutation events for each tumor type.

The frequencies of somatic mutations (lightblue) and significant mutation (pink) were shown as bar. To compare the frequency between samples with and without recurrent fusions ($n \geq 2$), Welch's t test was performed. A heatmap shows recurrent fusions (green) and significant mutation events in each tumor types. Prostate cancer and melanoma samples with recurrent fusions had no available mutation data.

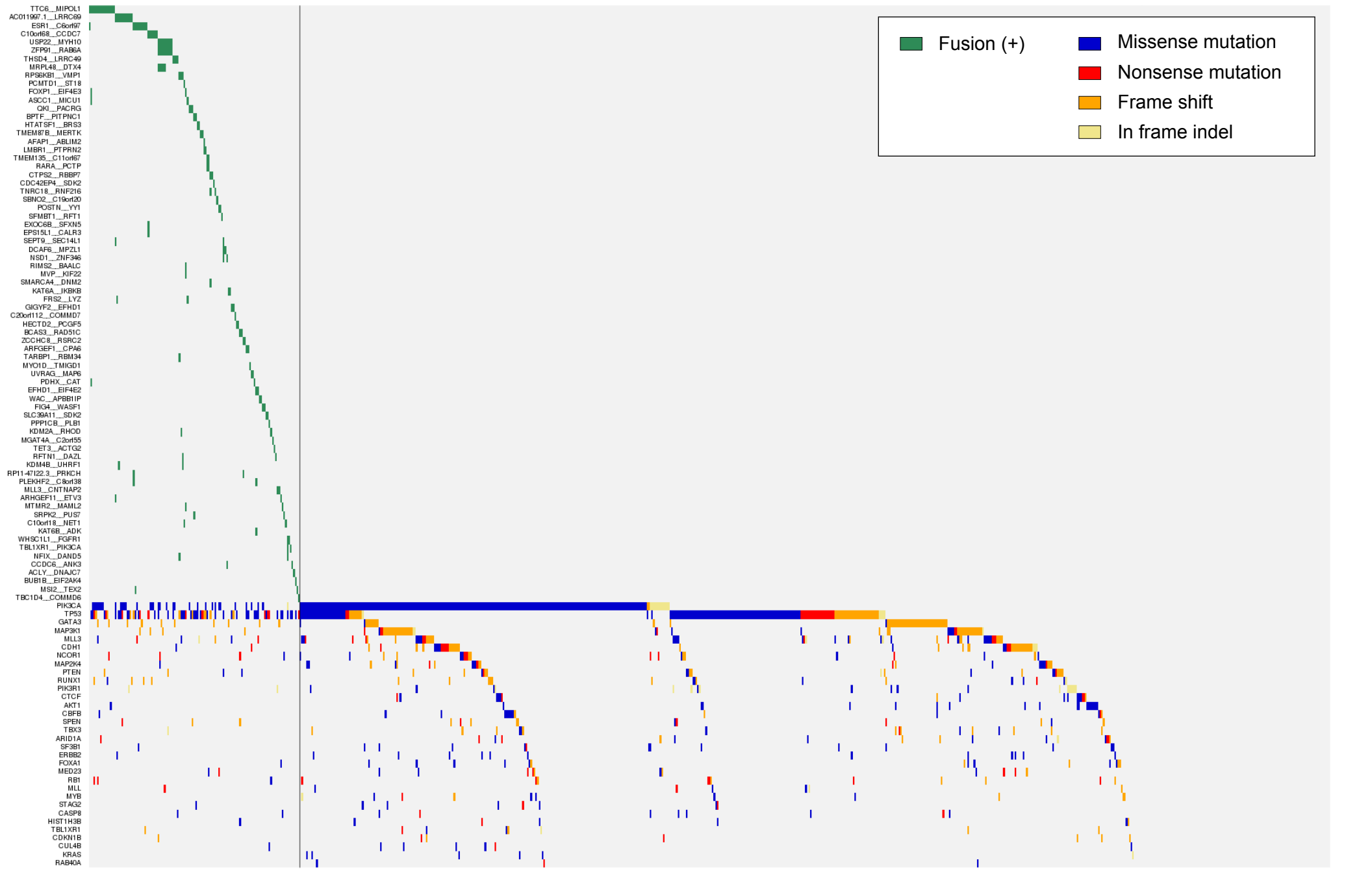
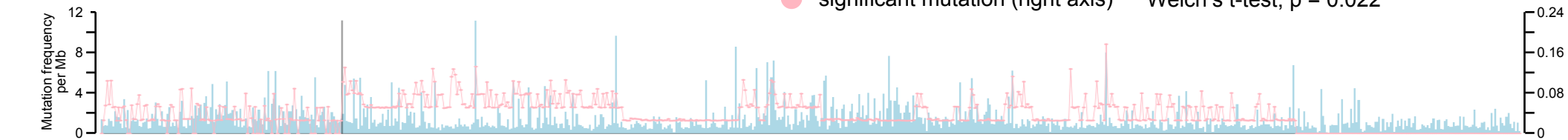
Supplementary Fig. 6 BRCA (n=754)

mutation (left axis)

Welch's t-test, $p = 0.0035$

significant mutation (right axis)

Welch's t-test, $p = 0.022$



Supplementary Fig. 6

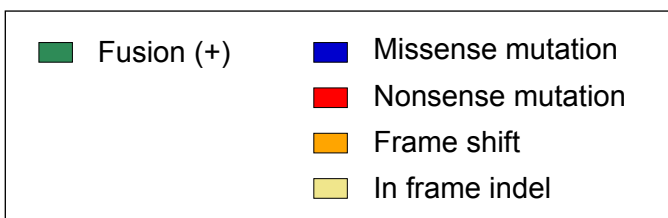
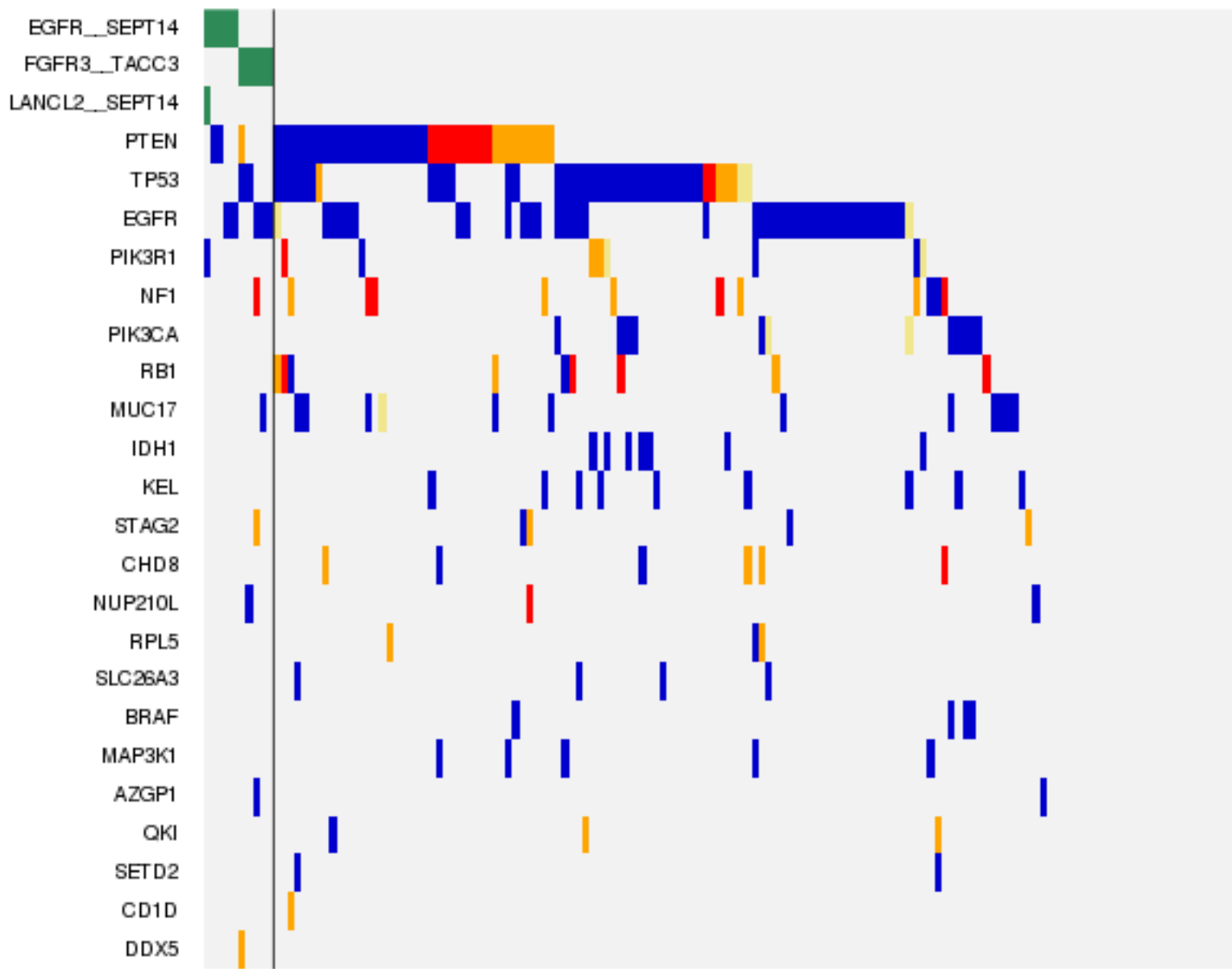
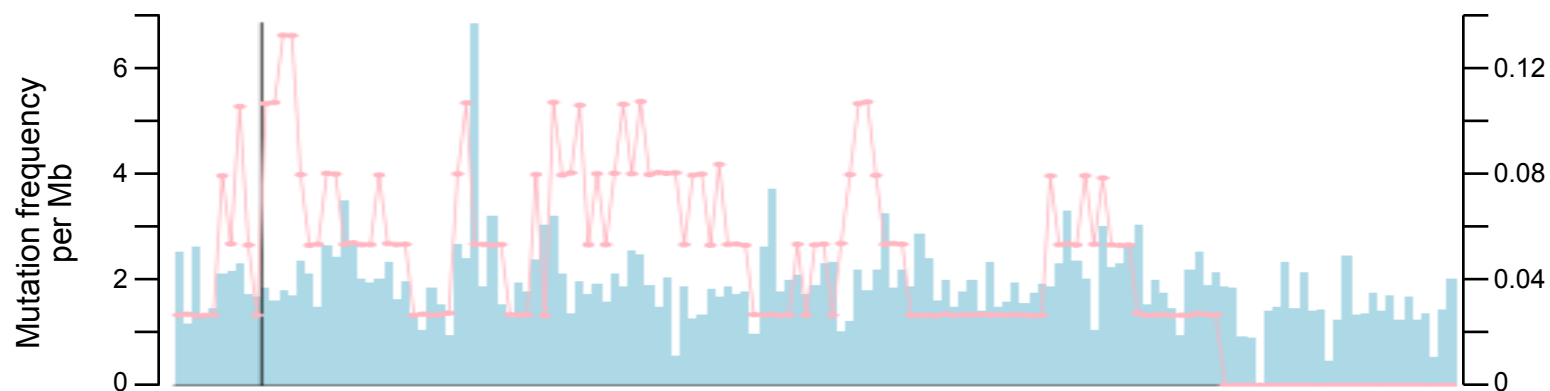
GBM (n=147)

mutation (left axis)

Welch's t-test, $p = 0.97$

significant mutation (right axis)

Welch's t-test, $p = 0.89$



Supplementary Fig. 6

HNSC (n = 296)

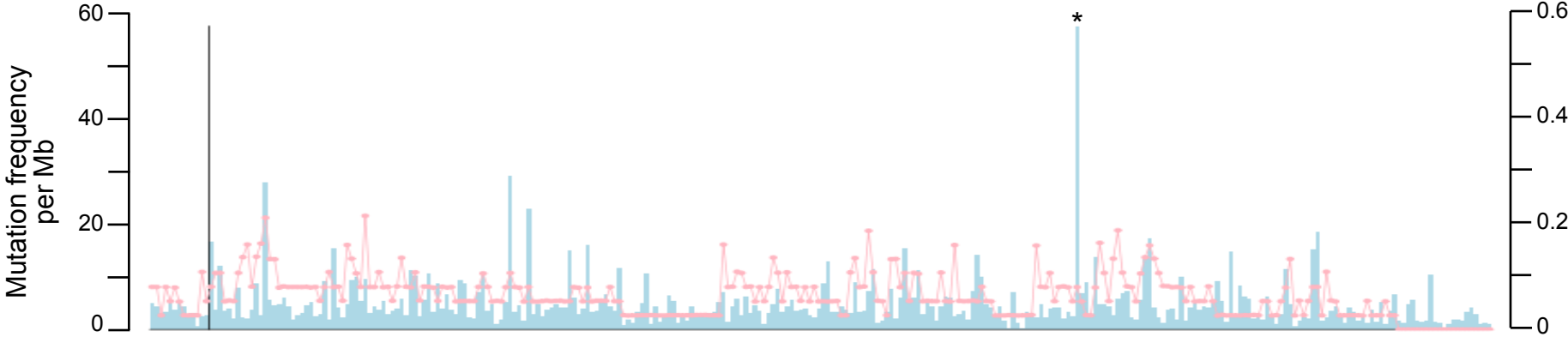
mutation (left axis)

Welch's t-test, p = 0.0023

*(n>60)

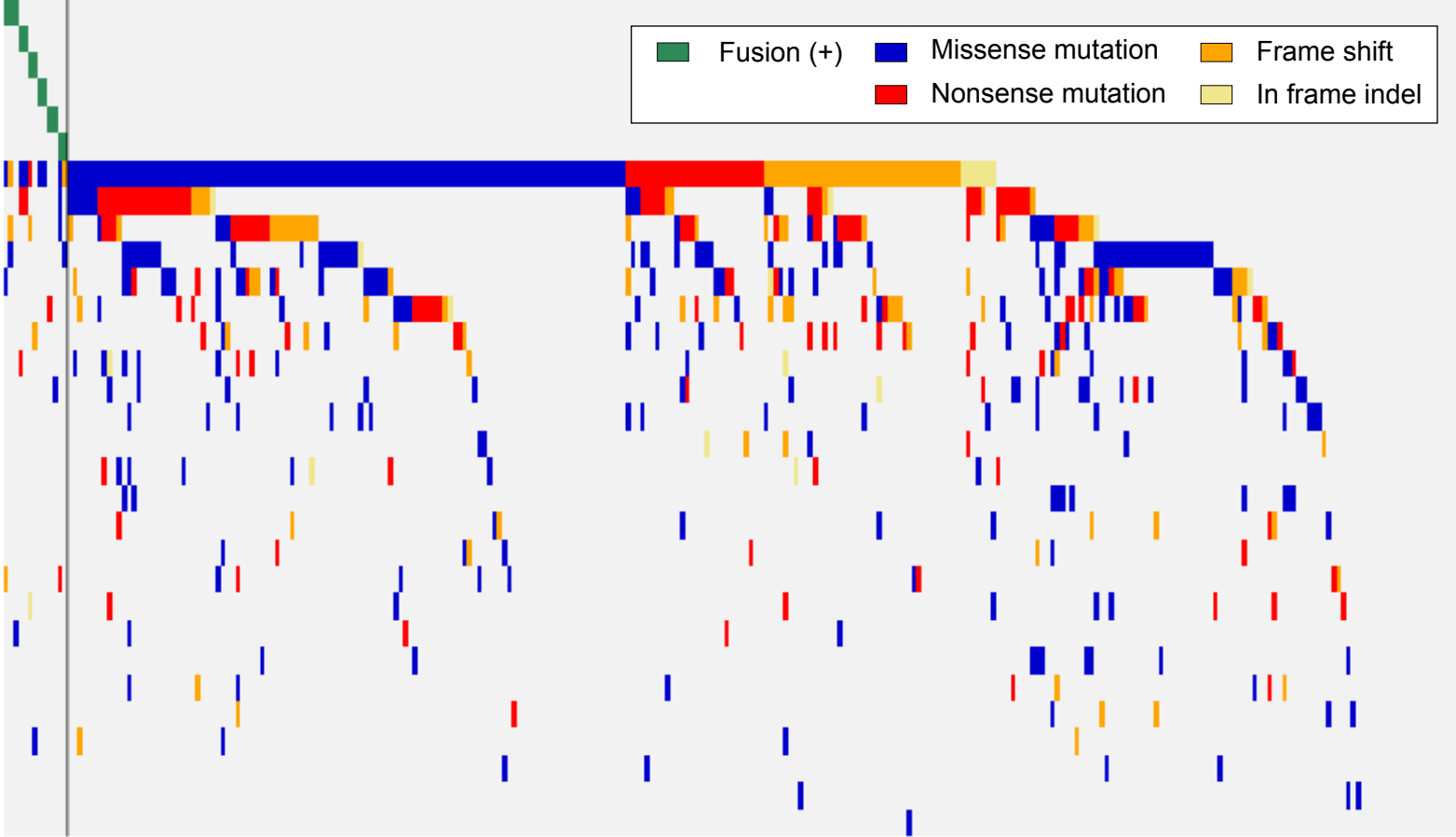
significant mutation (right axis)

Welch's t-test, p = 0.030



- RPS6KB1__VMP1
- AC008745.1__EHD2
- RNASE10__CD38
- PANX1__HEPHL1
- FGFR3__TACC3
- SH3PXD2A__OBFC1
- TP53
- CDKN2A
- FAT1
- PIK3CA
- NOTCH1
- MLL2
- NSD1
- CASP8
- EP300
- NFE2L2
- EPHA2
- RASA1
- HRAS
- ZNF750
- CTCF
- HLA-A
- TGFBR2
- PTEN
- RAC1
- HLA-B
- B2M
- MAP4K3
- RHOA
- IPO7
- OTUD7A

Fusion (+)	Missense mutation	Frame shift
	Nonsense mutation	In frame indel



Supplementary Fig. 6

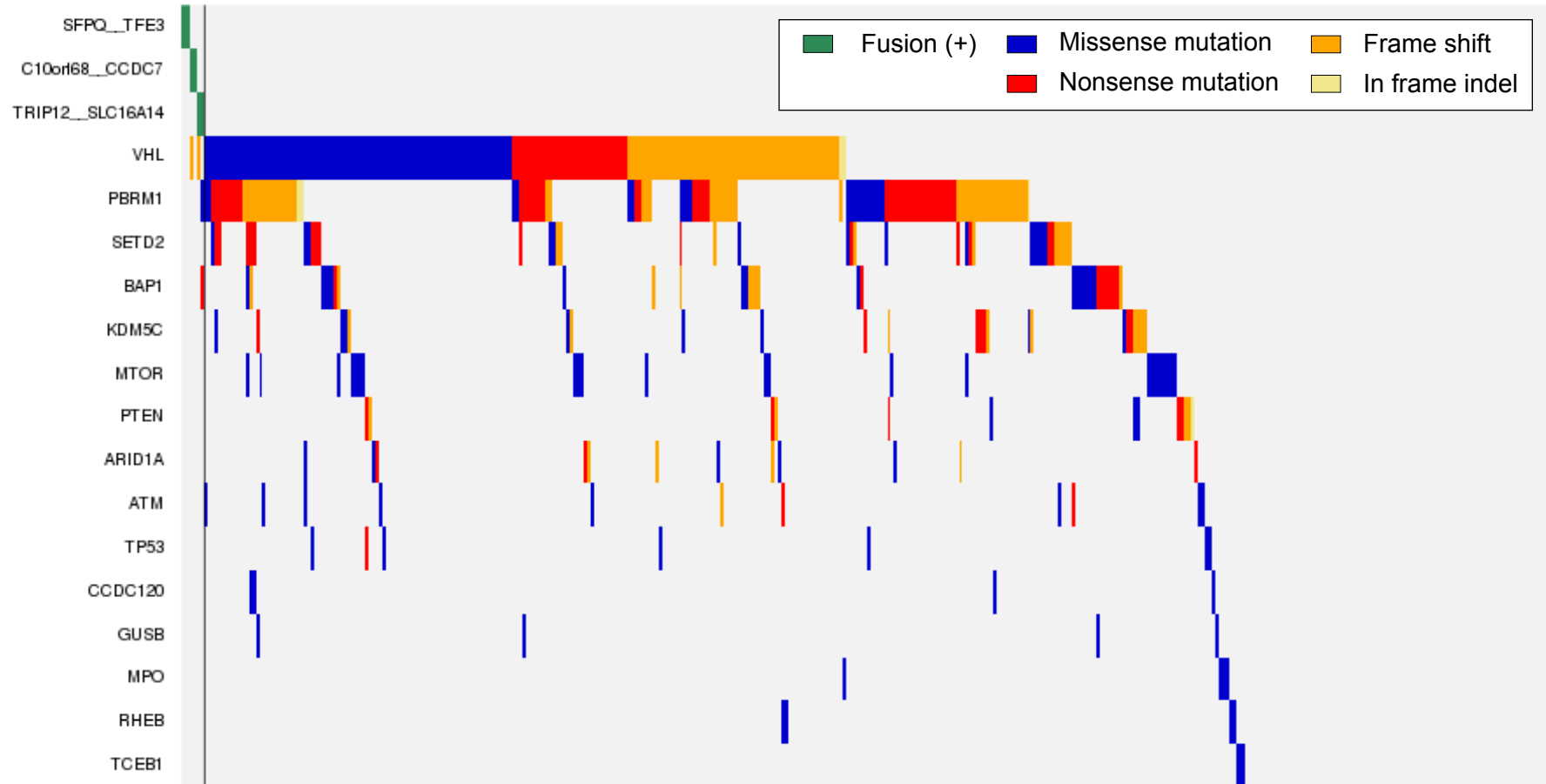
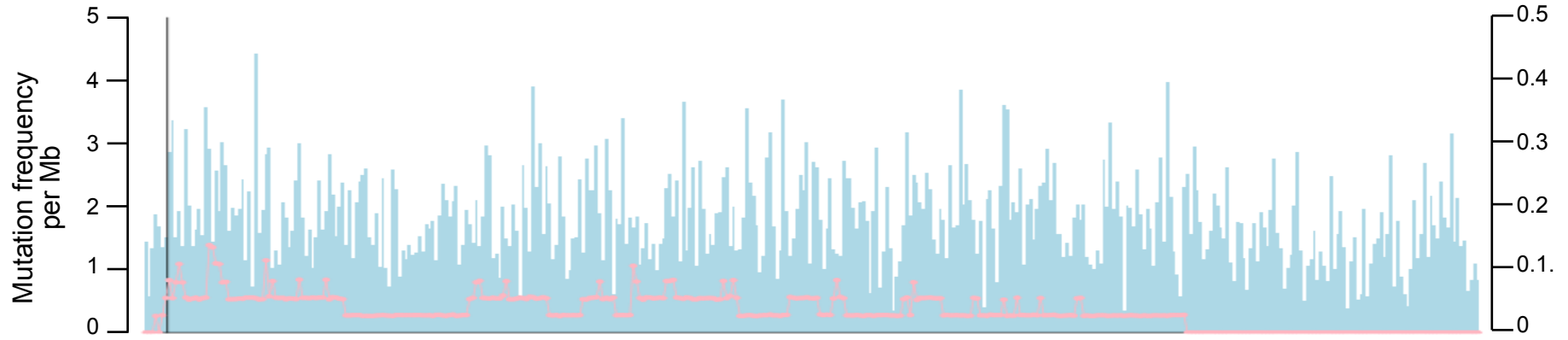
KIRC (n = 400)

mutation (left axis)

Welch's t-test, p = 0.039

significant mutation (right axis)

Welch's t-test, p = 0.063



Supplementary Fig. 6

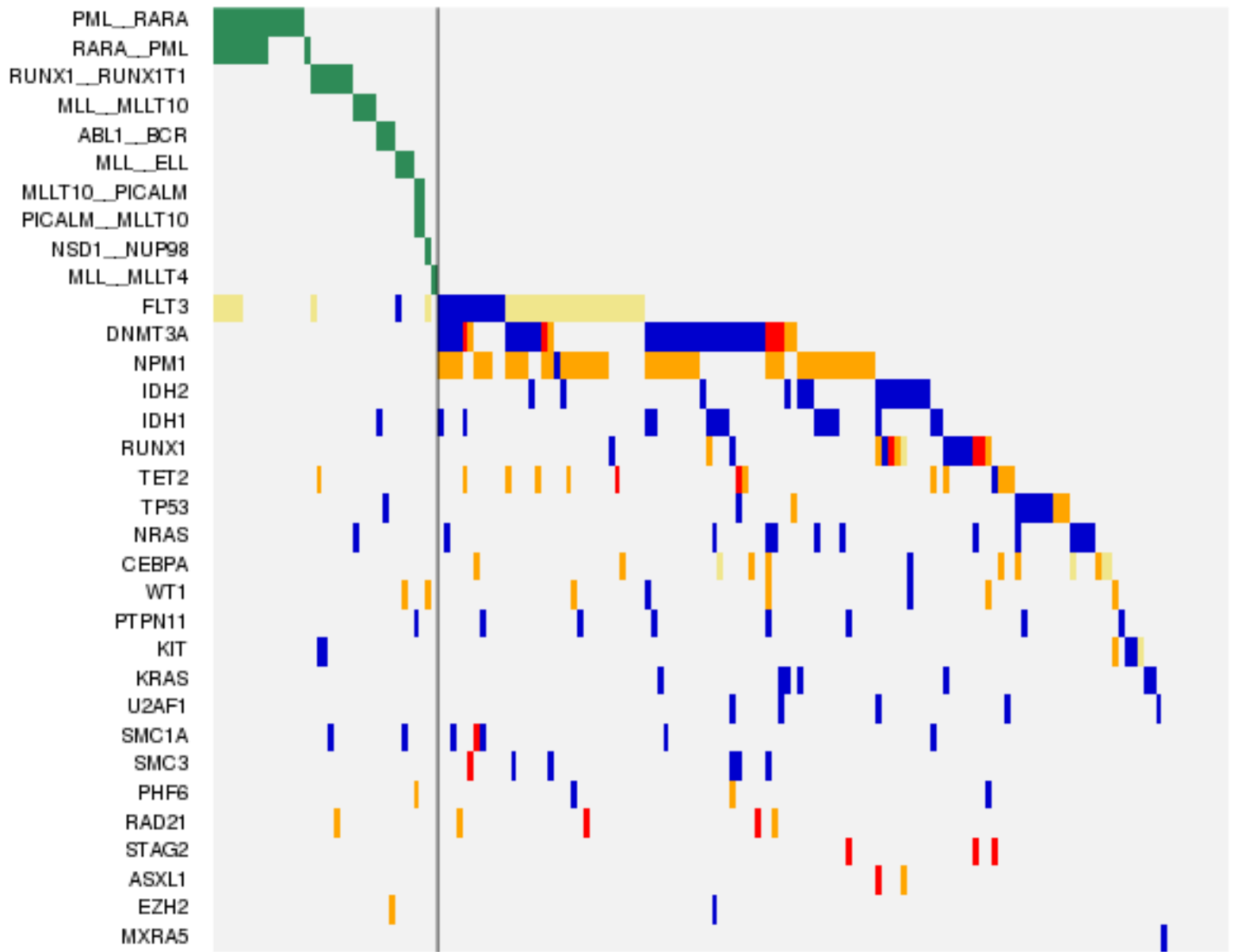
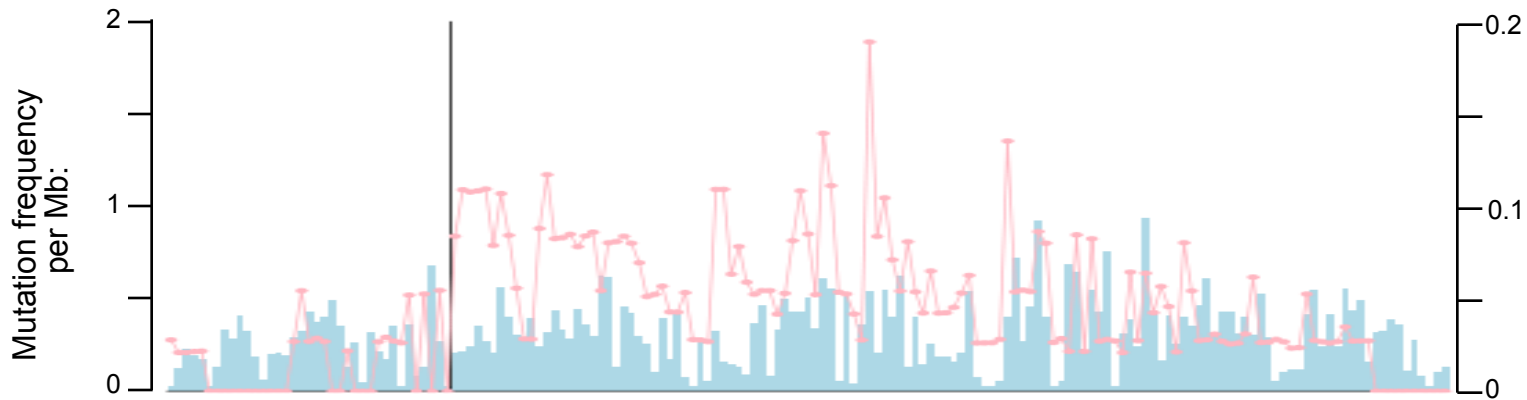
LAML (n = 167)

mutation (left axis)

Welch's t-test, p = 0.0048

significant mutation (right axis)

Welch's t-test, p = 4.8e-15



Supplementary Fig. 6

LUAD (n = 171)

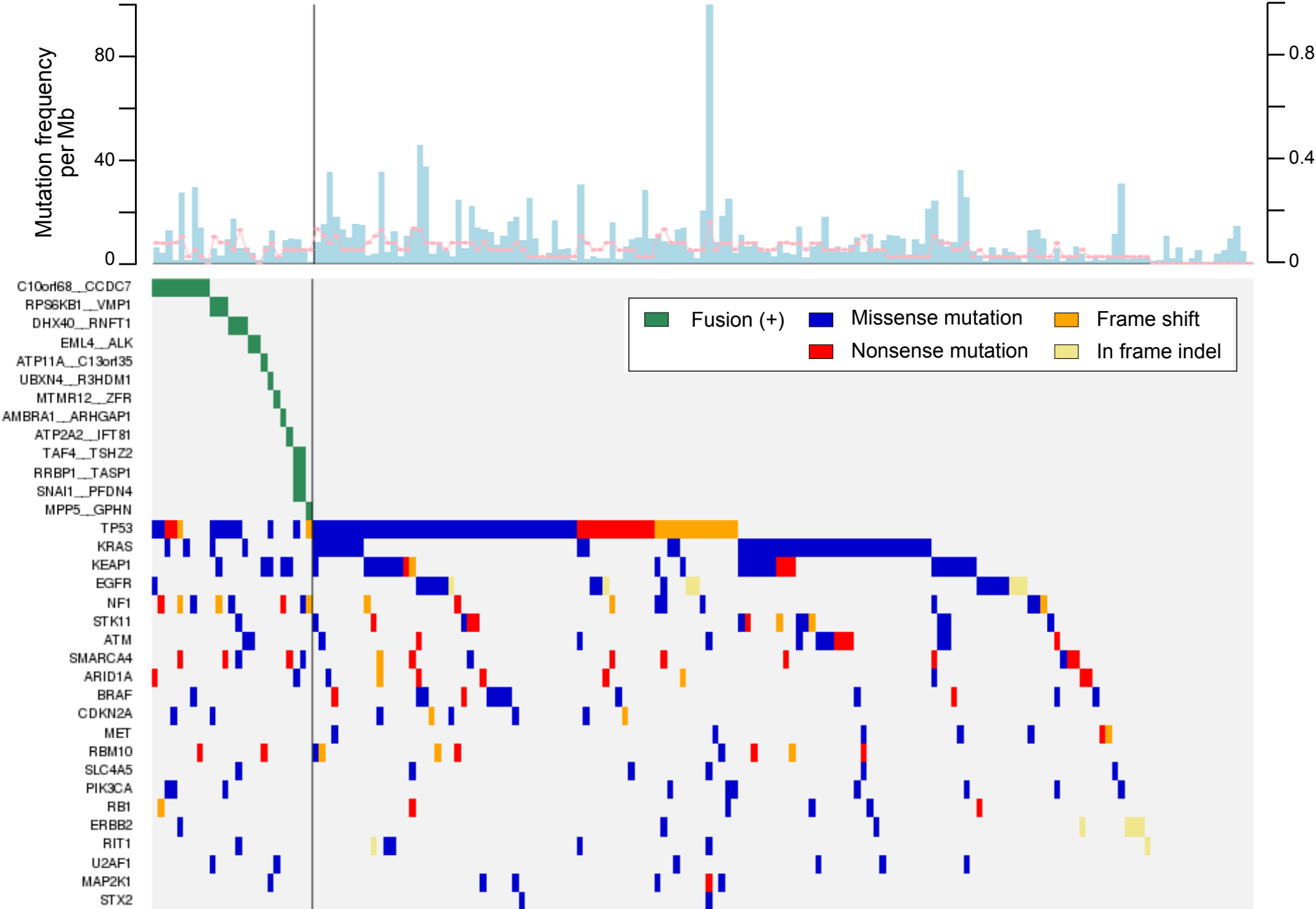
*(n>100)

mutation (left axis)

significant mutation (right axis)

Welch's t-test, p = 0.23

Welch's t-test, p = 0.32

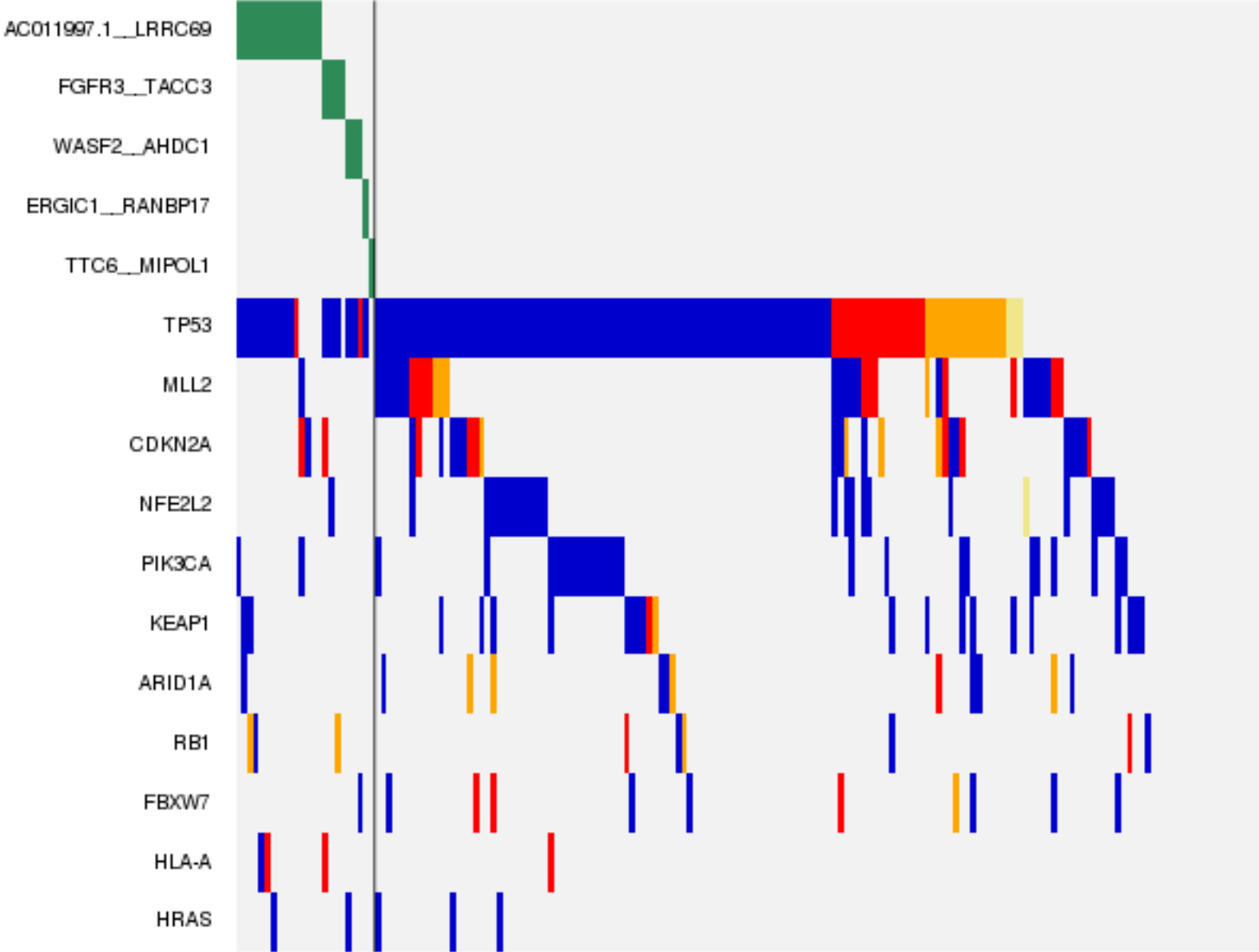
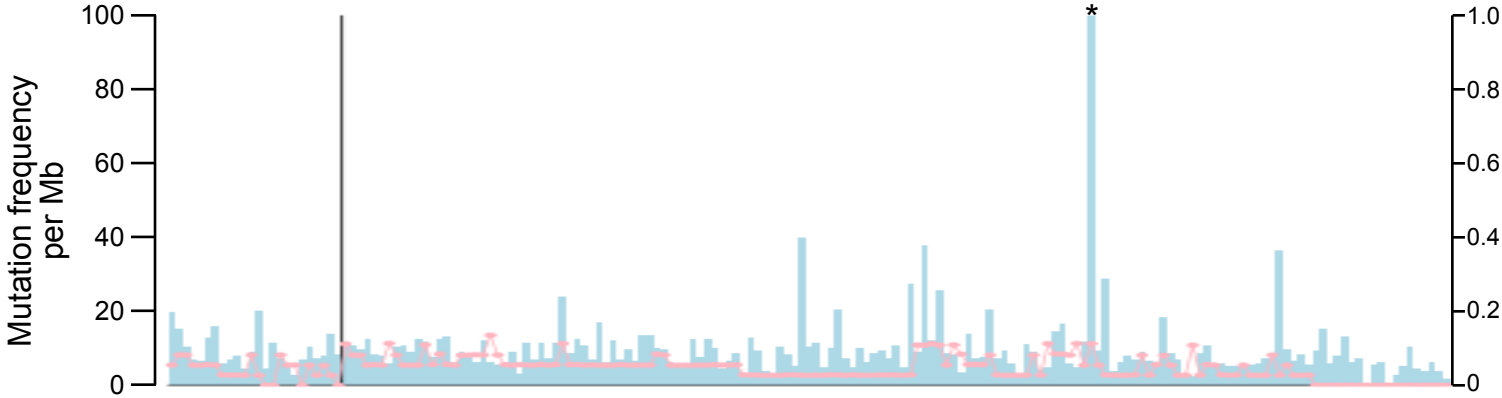


Supplementary Fig .6

LUSC (n = 177)
*(n>100)

mutation (left axis)
significant mutation (right axis)

Welch's t-test, p = 0.72
Welch's t-test, p = 0.32



Supplementary Fig. 6

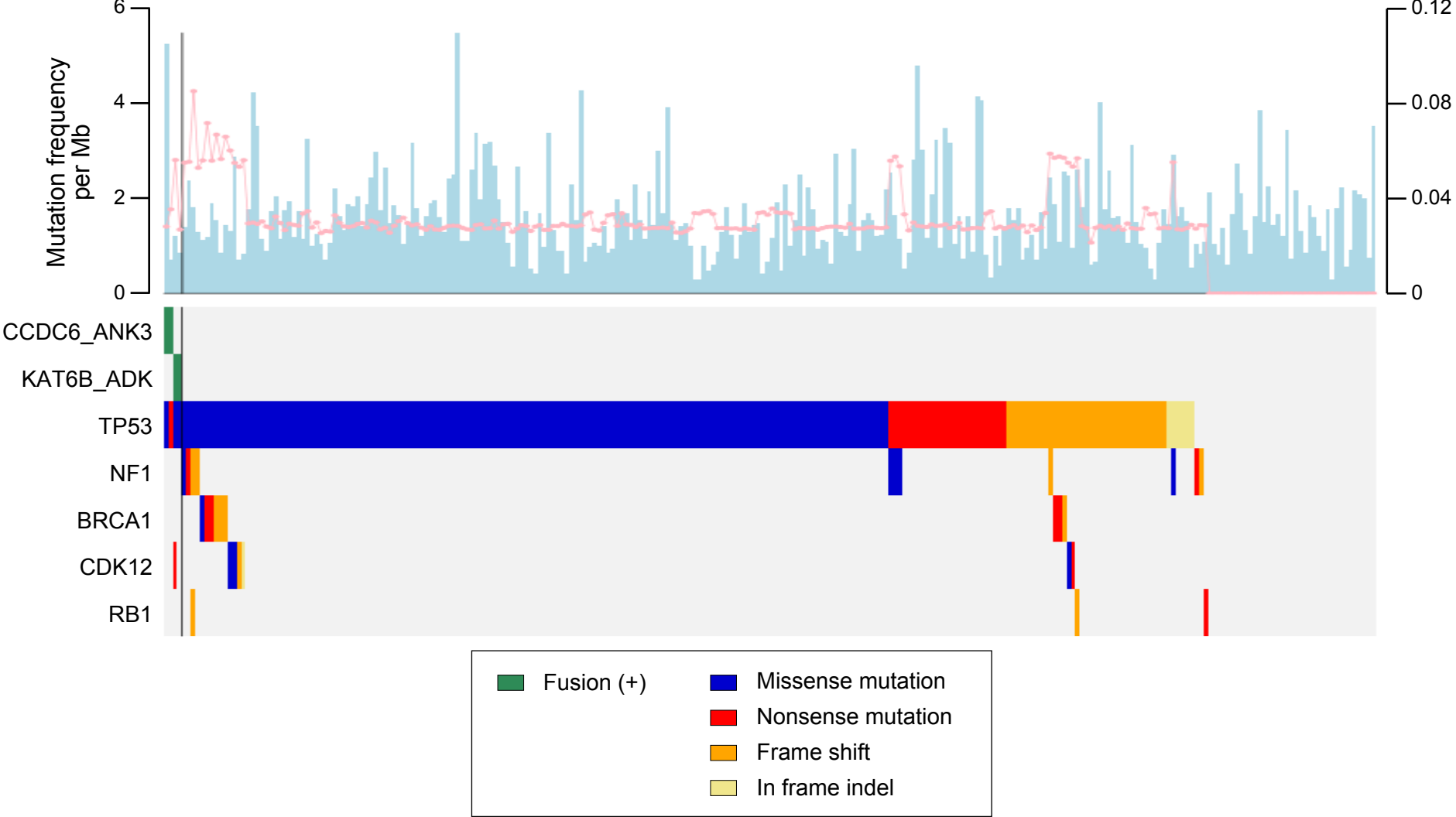
OV (n = 266)

mutation (left axis)

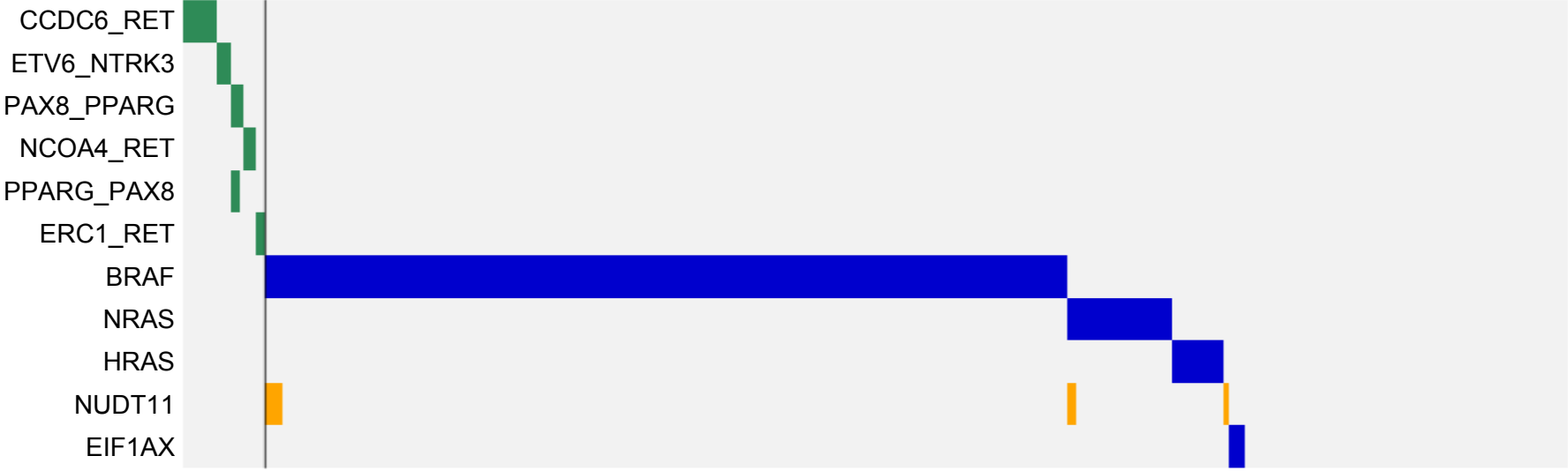
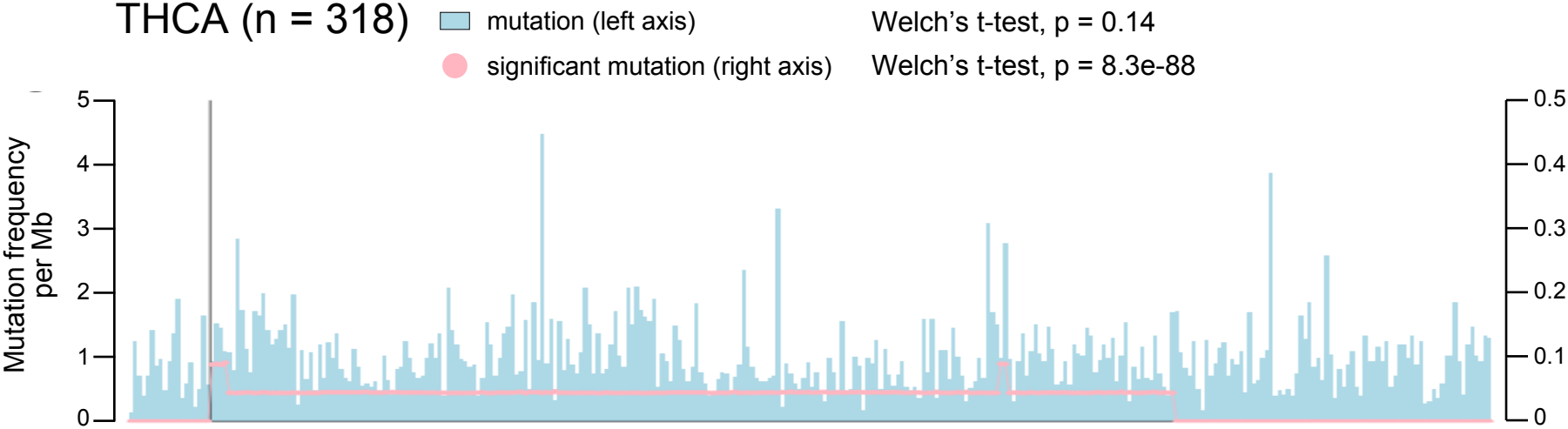
Welch's t-test, p = 0.26

significant mutation (right axis)

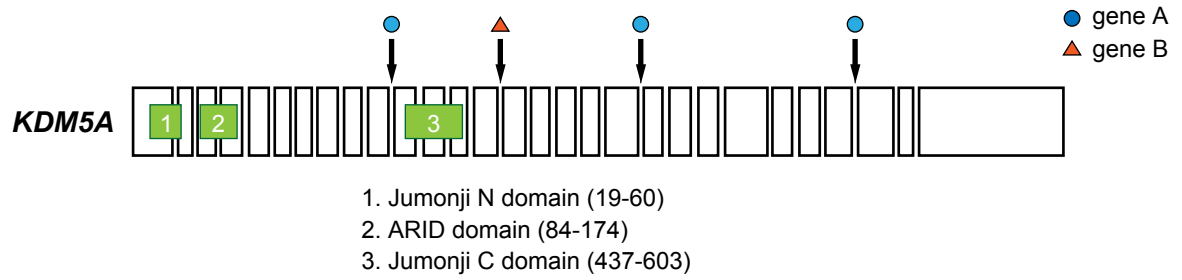
Welch's t-test, p = 0.27



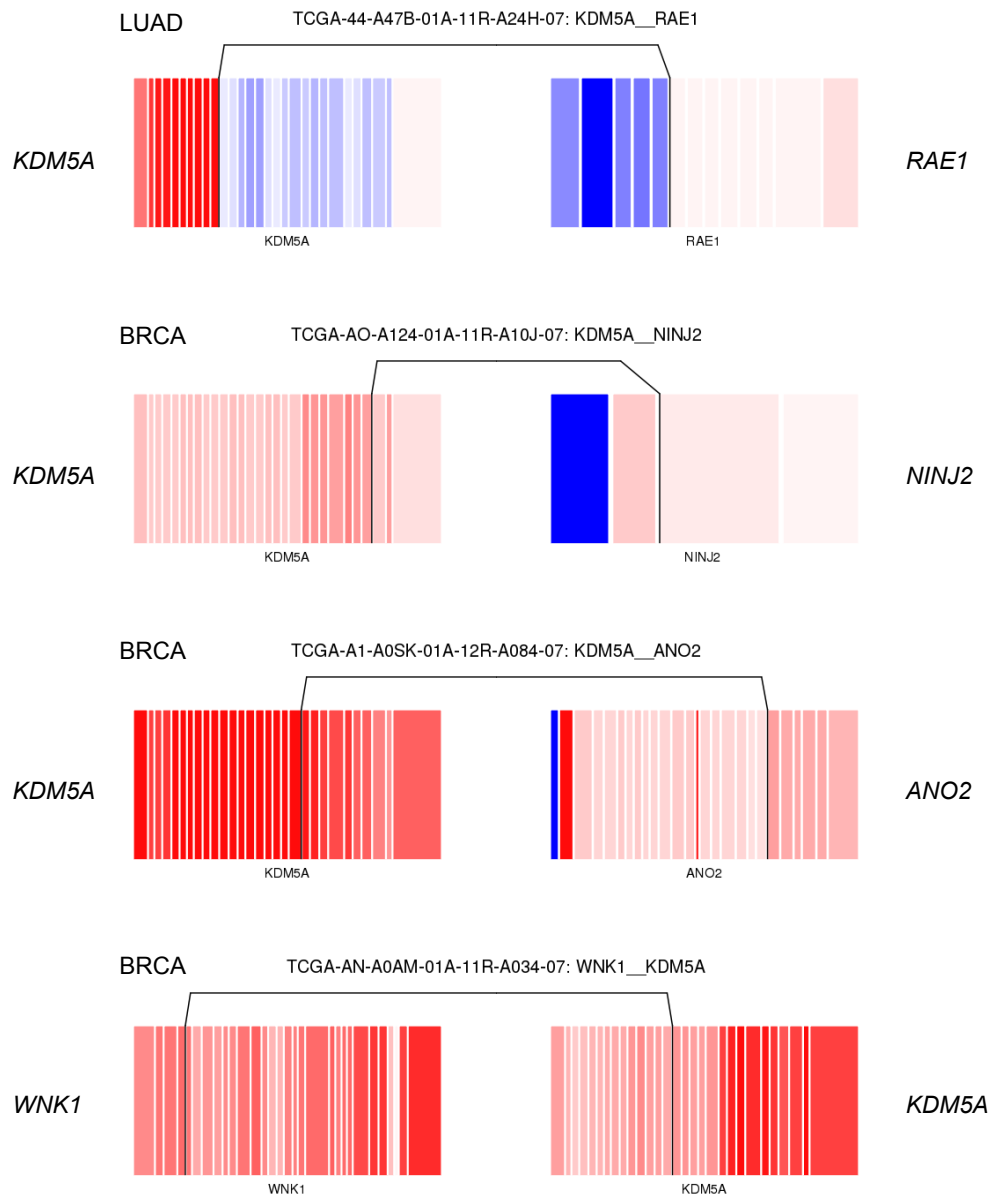
Supplementary Fig. 6



A



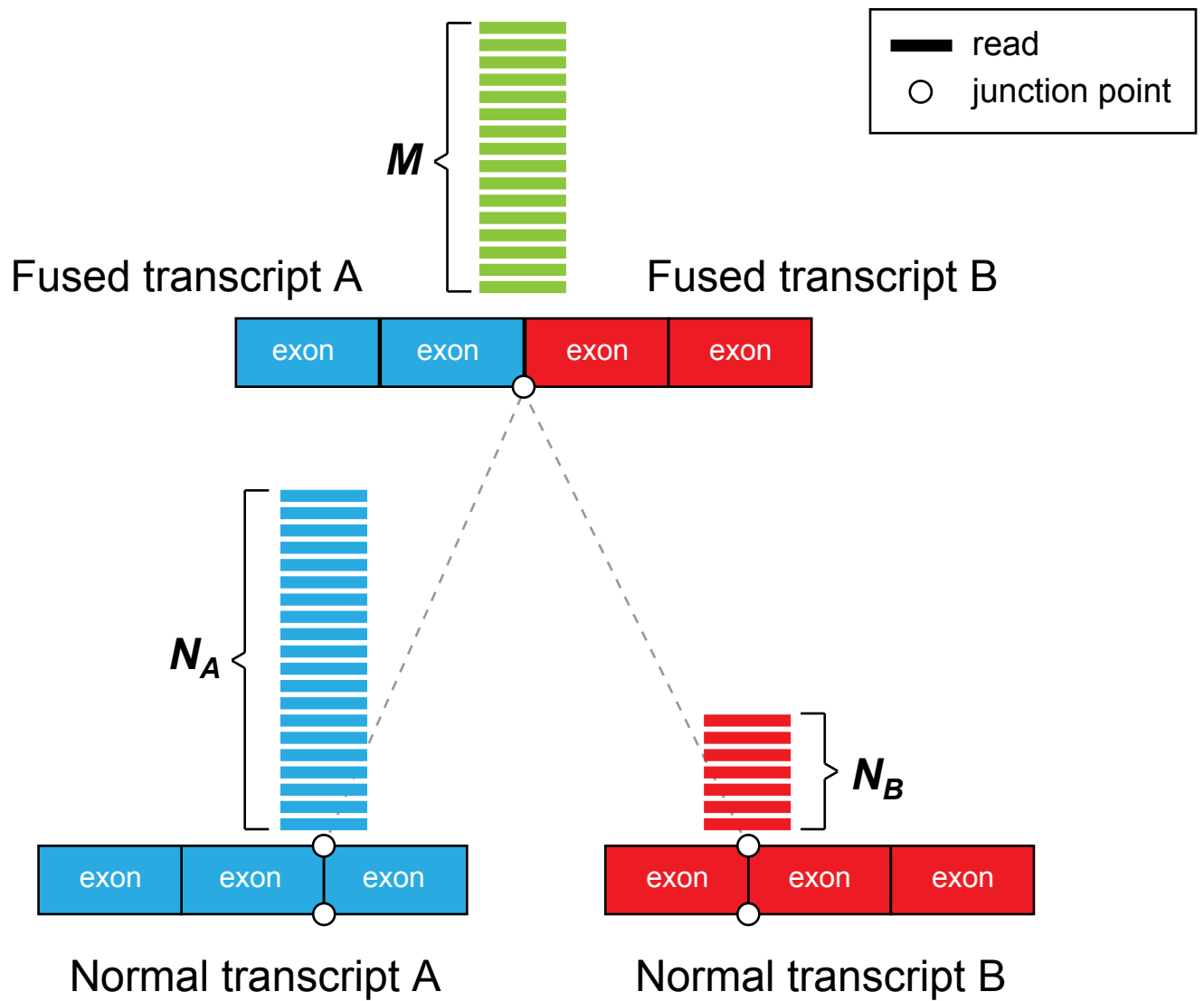
B



Supplementary Figure 7. *KDM5A* fusions as a therapeutic target

(A) Position of each domain in *KDM5A* gene and junction points of *KDM5A* fusions.

(B) Exon expression plots demonstrated Z-normalized exon expression for each exon in thyroid cancers. Red and blue represent relatively high and low exon expression.

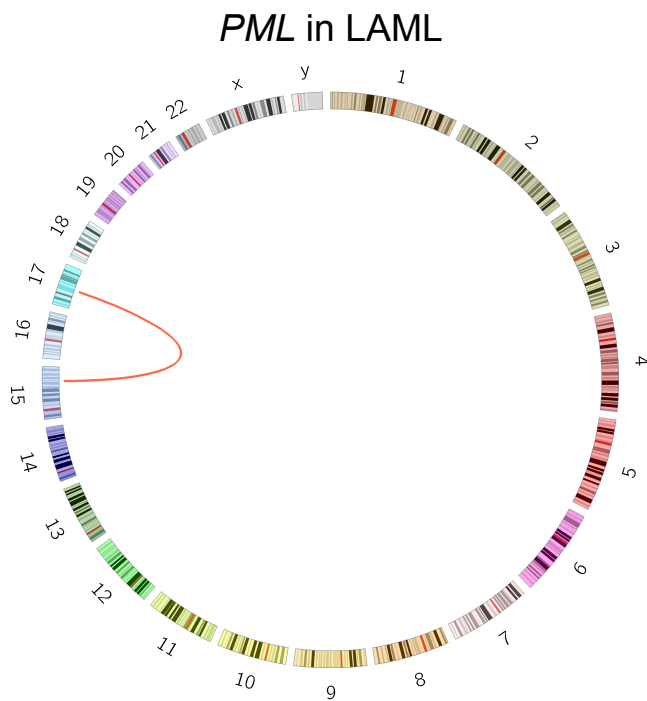


$$\text{Transcript allele fraction (TAF) for transcript A} = \frac{M}{M + N_A}$$

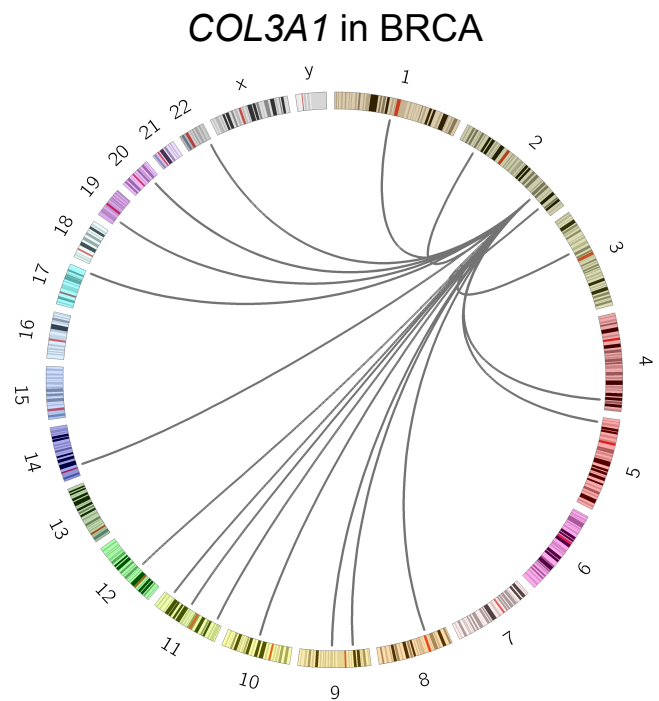
$$\text{Transcript allele fraction (TAF) for transcript B} = \frac{M}{M + N_B}$$

Supplementary Figure 8. Transcript allele fraction (TAF)

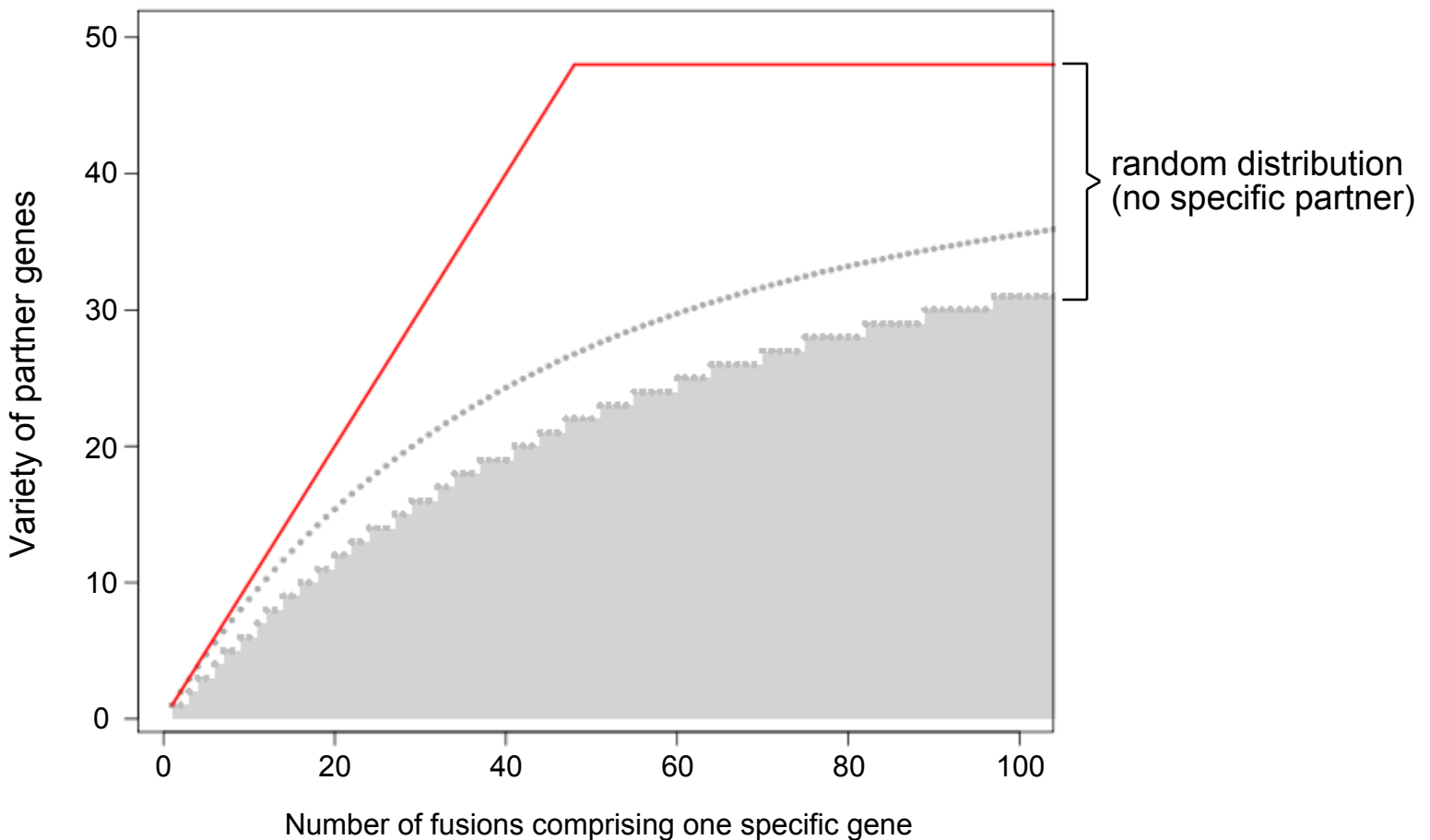
This image shows the definition of transcript allele fraction (TAF). The TAF score is measured as the ratio of junction spanning reads to total reads crossing over junction point mapped to the reference transcript.

A

Variety of partner gene = 1
(Number of fusions = 11)

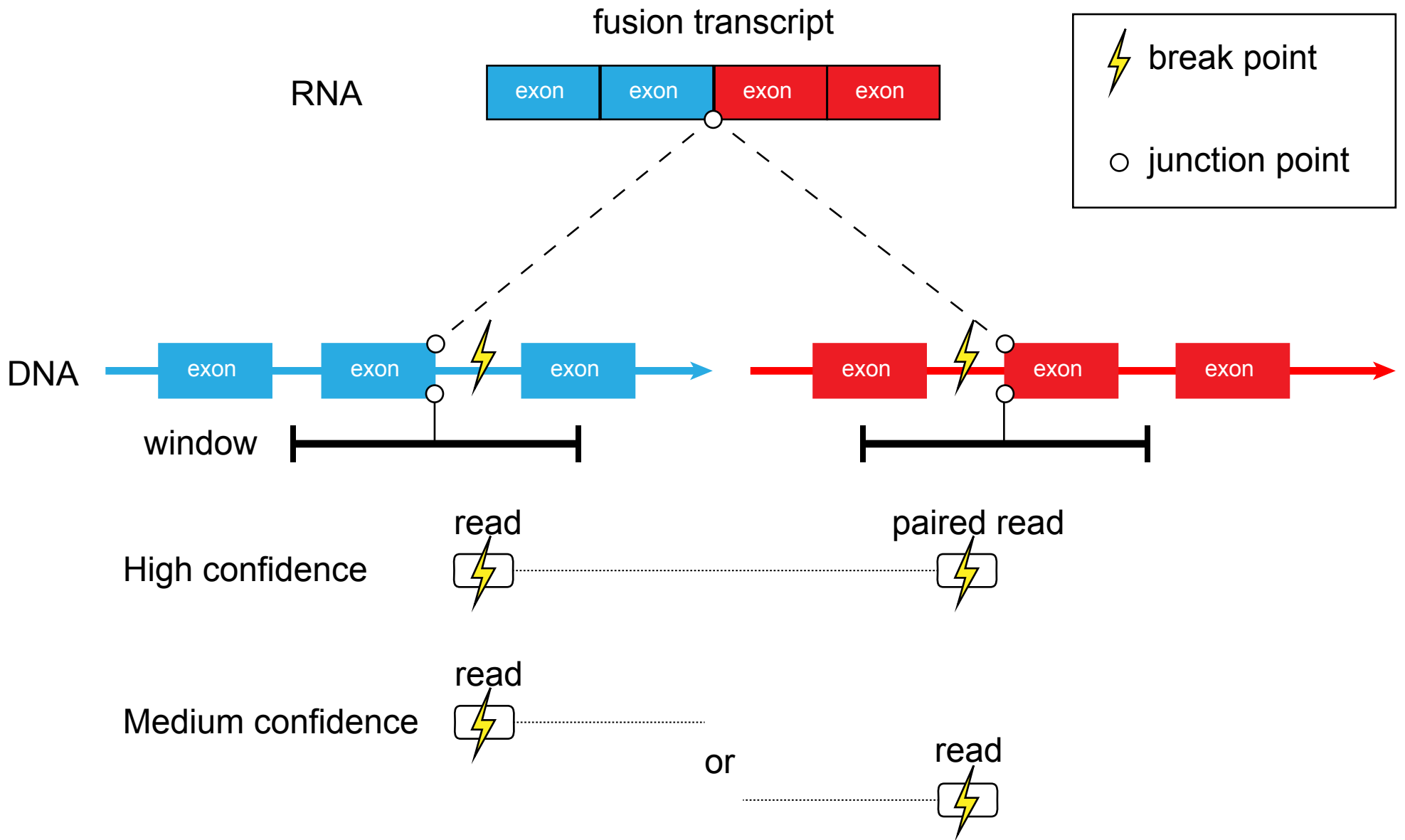


Variety of partner gene = 18
(Number of fusions = 28)

B

Supplementary Figure 9. Diversity of partner gene variety

(A) Circos plots show representative examples indicating low (PML in acute myeloid leukemia) and high (COL3A1 in breast cancer) partner gene variety. (B) The graph demonstrates random distribution of partner gene variety. Red and grey dashed lines depict the maximum and mean of PGV per number of fusion.



Supplementary Figure 10. The concept of validating fusion transcripts

To consider the difference in structure between DNA and RNA, breakpoints (yellow) were searched within the window from the inferred junction point. High confidence validation means the presence of paired end reads showing breakpoints within the window from junction point.