Comprehensive genome and transcriptome analysis reveals genetic basis for gene fusions in cancer

Nuno A. Fonseca[1*], Yao He[2*], Liliana Greger[1], PCAWG3, Alvis Brazma[1], Zemin Zhang[2]

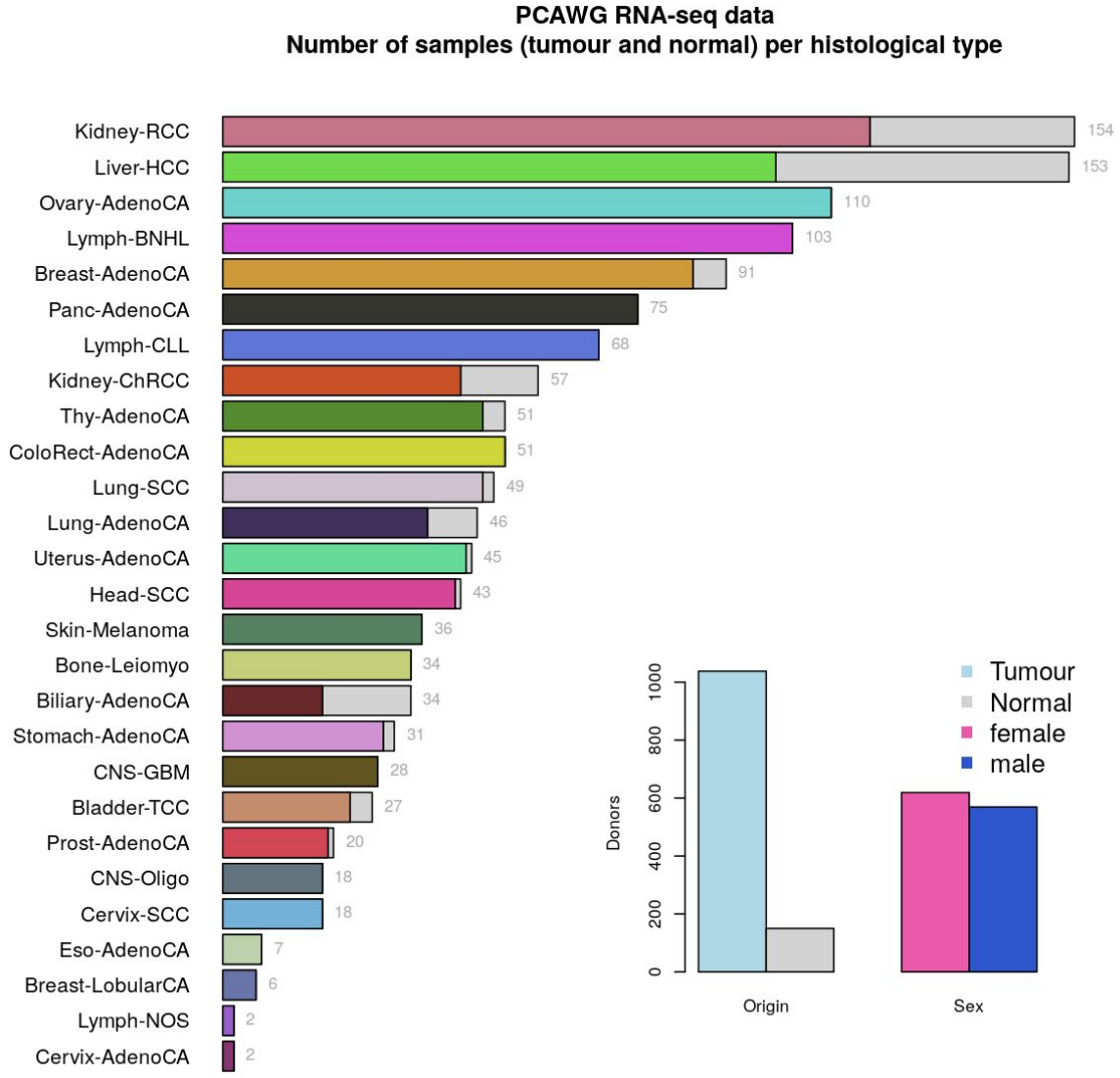
[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK;

[2]Peking-Tsinghua Centre for Life Sciences, BIOPIC, and Beijing Advanced Innovation Centre for Genomics, Peking University, Beijing, 100871, China

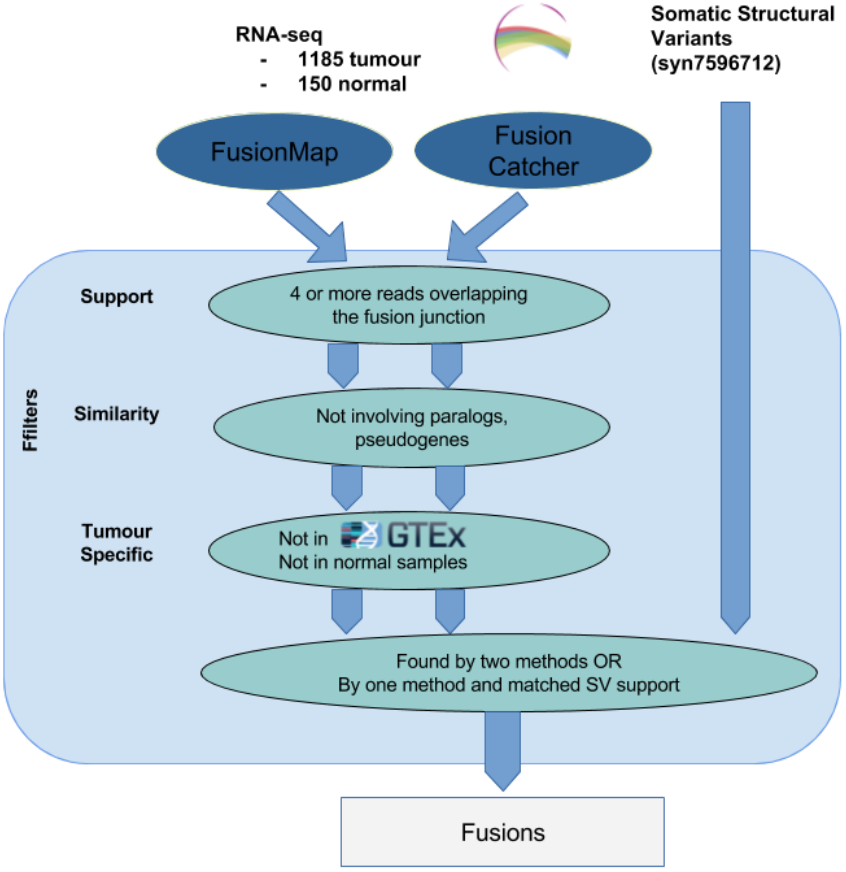
*Joint first authors

Supplementary Figures

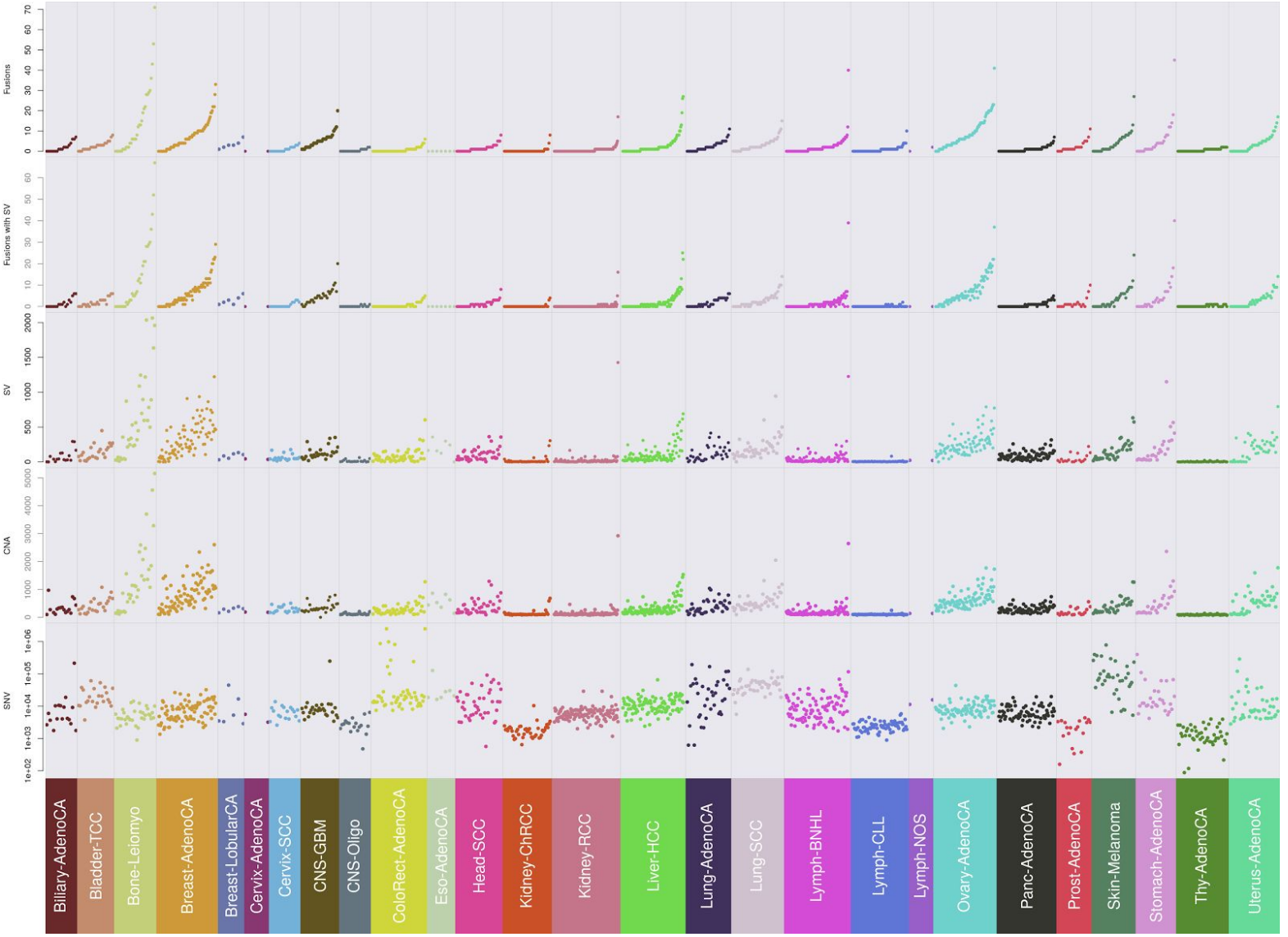
Supplementary Figure 1: ICGC RNA-seq data for diverse cancer types.



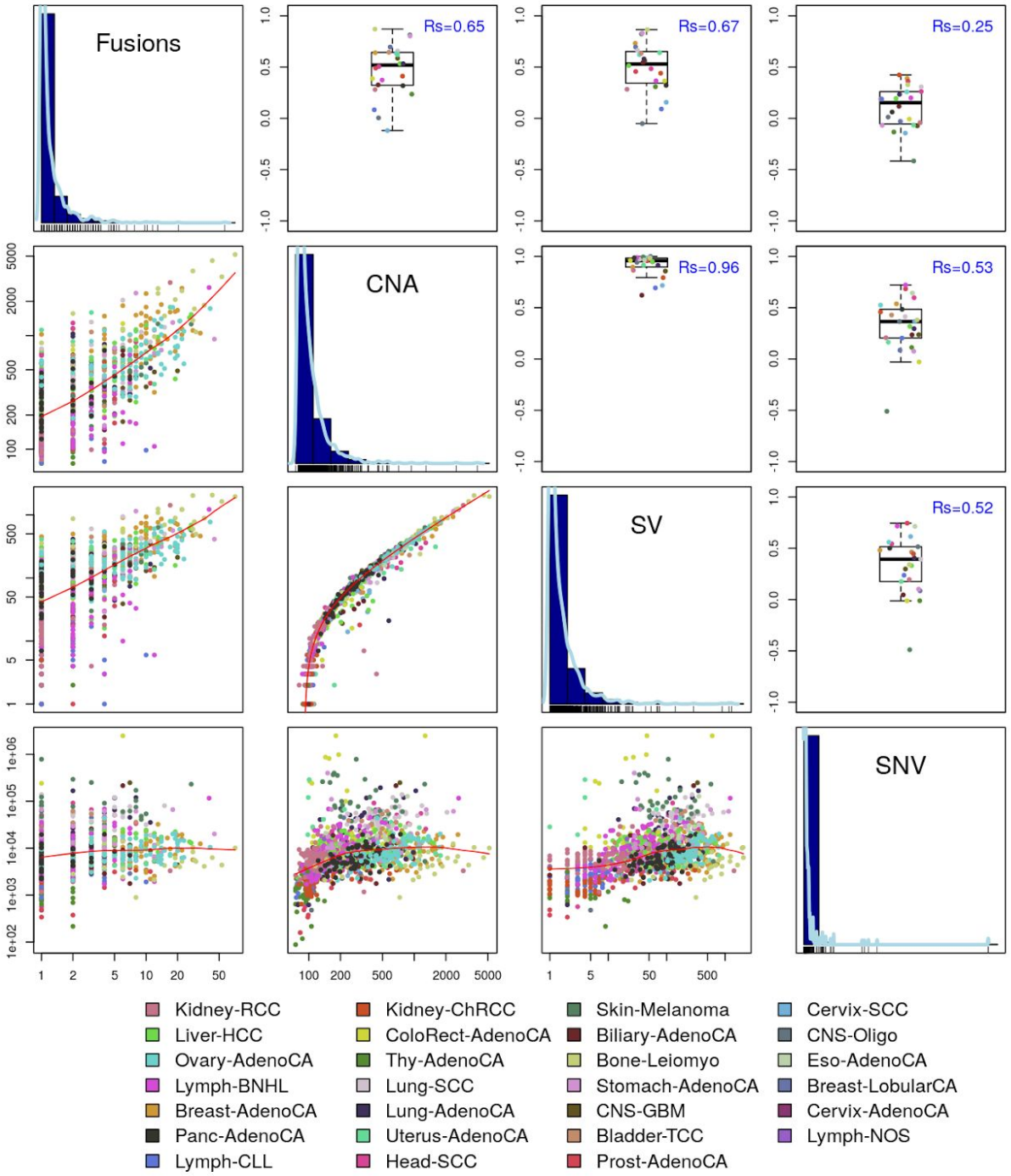
Supplementary Figure 2: Gene fusion detection pipeline. Fusions between any two genes were identified based on two different gene fusion detection tools: FusionMap and FusionCatcher. To reduce the number of false positive fusions, the two sets of fusions were filtered to exclude those with strong homology to each other or those with occurrence in normal samples from the PCAWG cohort and GTEx (phs000424.v4.p1). Finally, only fusions found by both fusion detection pipelines and/or with matched structural variant support were included in the final set of fusions.



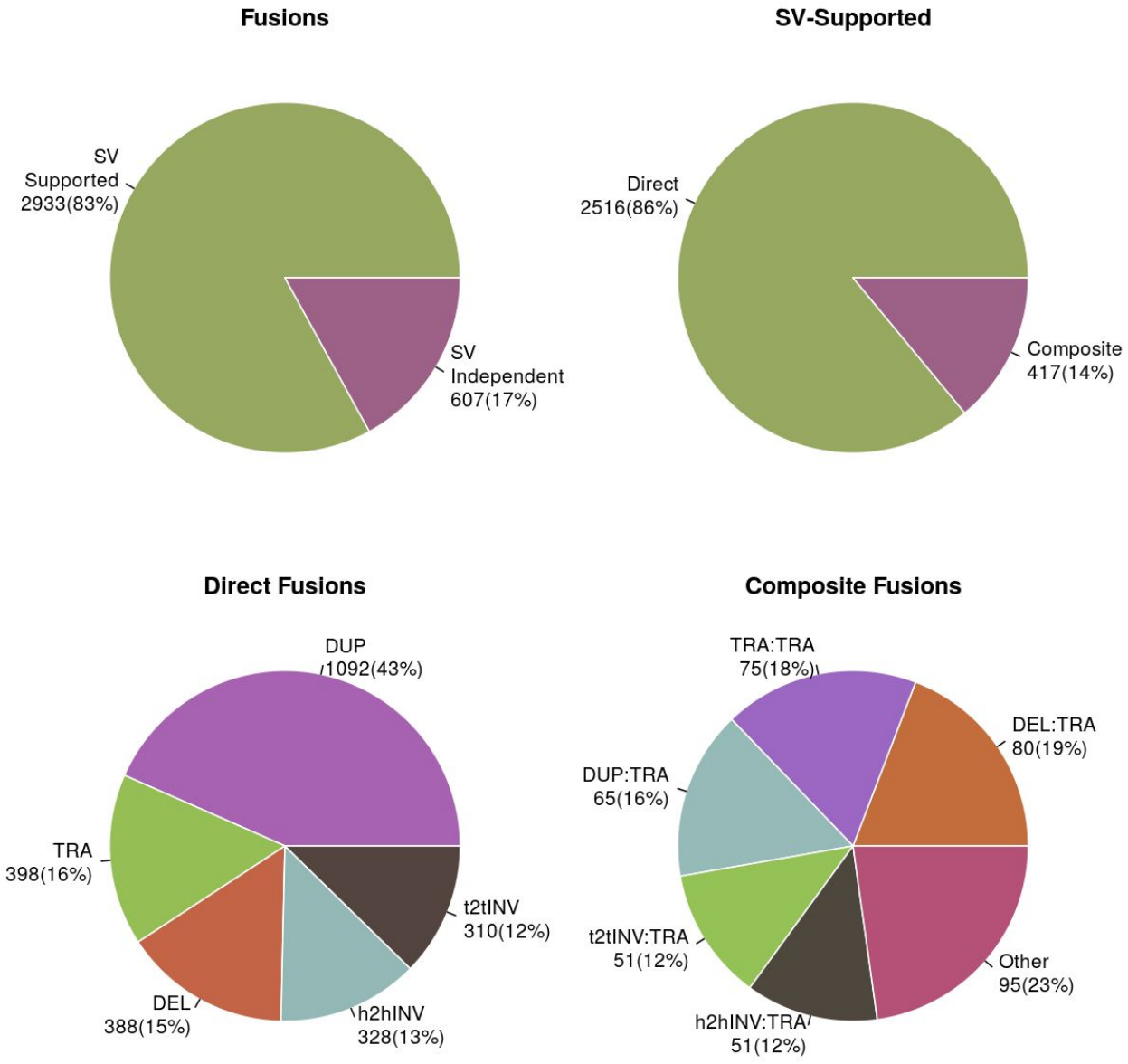
Supplementary Figure 3: Comparison of gene fusion distribution in the context of structural variations (SV), copy number alteration (CNA), and single nucleotide variations (SNV). The numbers of gene fusions per sample and respective numbers of fusions with SV support, SVs, CNAs, and SNVs are plotted according to histotypes. Each dot corresponds to a sample, and the order of the (matched) samples across horizontal panels is preserved and is based on the number of fusions.



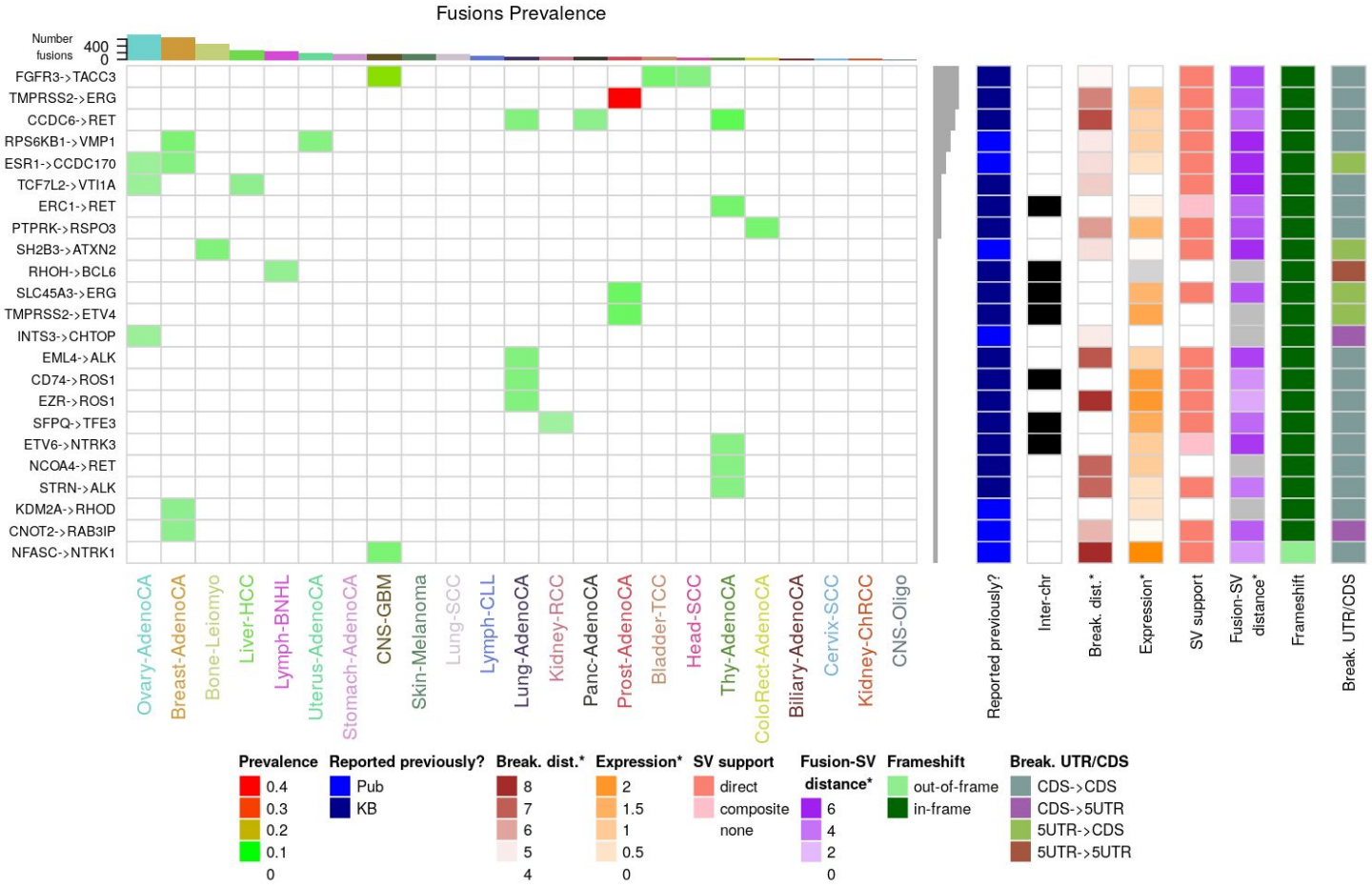
Supplementary Figure 4: Number of gene fusions per sample and respective number of fusions, structural variants (SV), copy number alterations (CNA), and single nucleotide variants (SNV). The diagonal histograms shows the distribution of the number of alterations per sample. The upper triangle presents the Spearman correlation between two types of alterations per histological type (dot) and together with the overall spearman correlation (in blue). The bottom triangle contains scatter plots contrasting the number of alterations for each sample (dot).



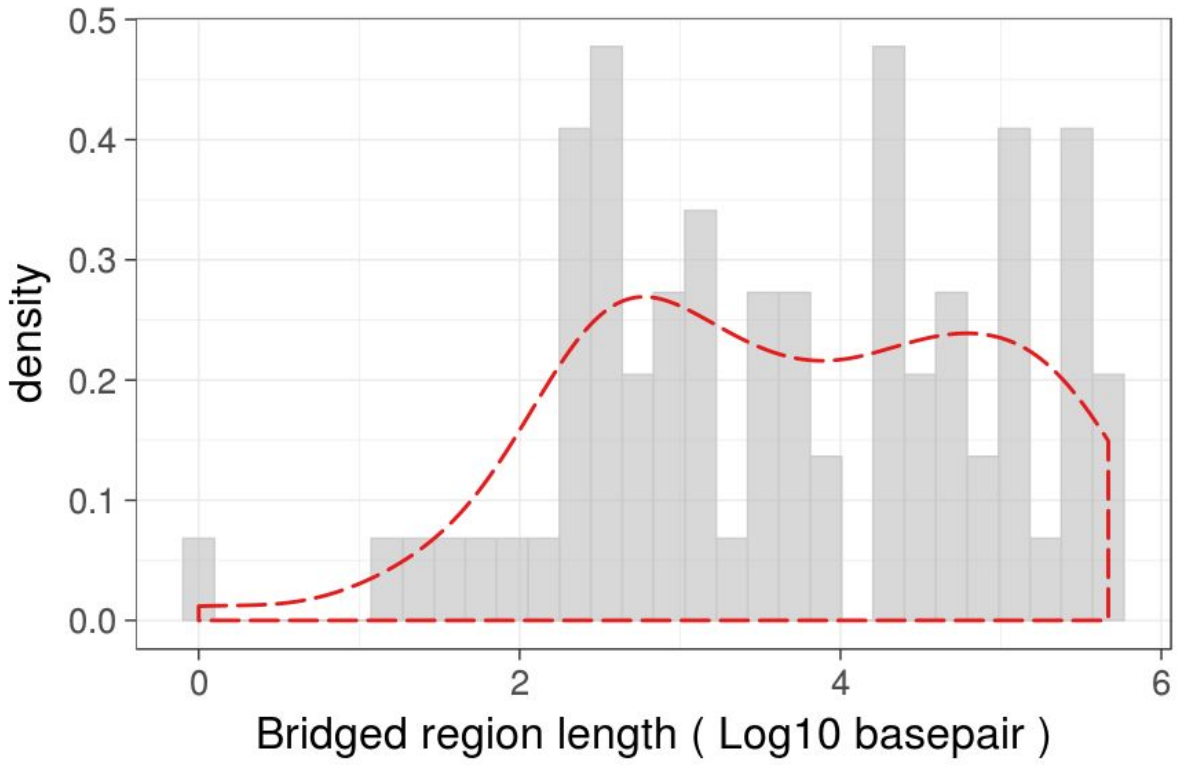
Supplementary Figure 5: Distribution of the different types of fusions. A) SV-support fusions and SV independent fusions. B) direct SV support fusions and composite fusions. C) Structural variants types associated with the direct SV support fusions: deletions (DEL), duplications (DUP), translocations (TRA), head to head inversion (h2hINV) and tail to tail inversion (t2tINV). D) Structural variants types associated with the composite SV fusions.



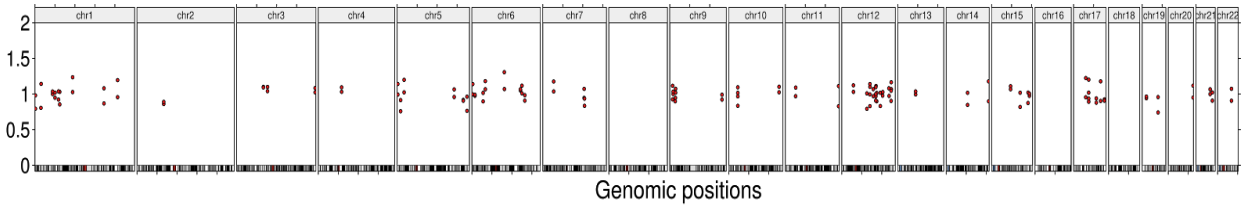
Supplementary Figure 6: Gene fusions previously reported with high confidence (online methods). Top histogram shows the total number of fusions per histotype while the left histogram corresponds to the number of times each fusion was detected across all histotypes. Each cell in the central matrix shows the prevalence of a fusion in a histological type. ChimerDB 3.0 [PMID:27899563] was used as a reference of previously reported gene fusions. The expression column presents the median expression of the putative transcript. The values with a star ('*') are log10 and distance unit is in bp.



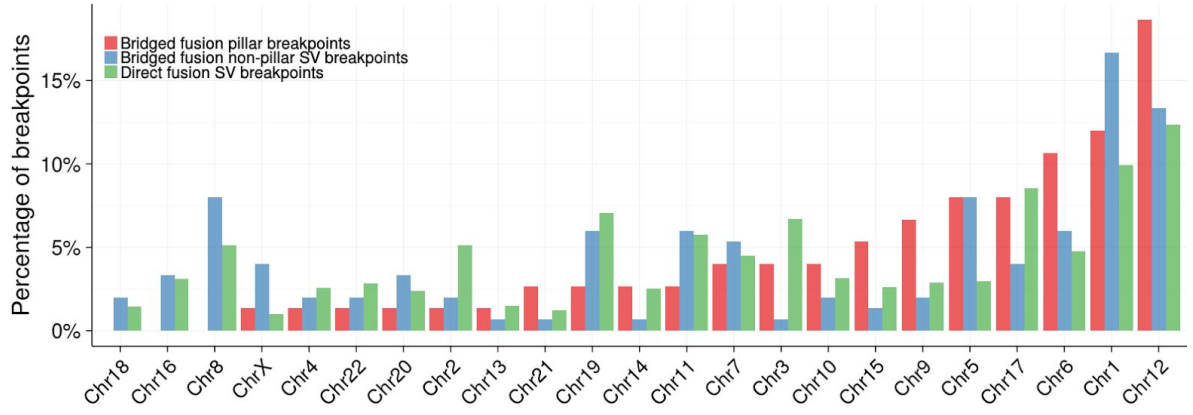
Supplementary Figure 7: Distribution of the lengths of bridged regions for bridged fusions. The density histogram shows the length distribution for those 75 bridged regions in Log10 scale.



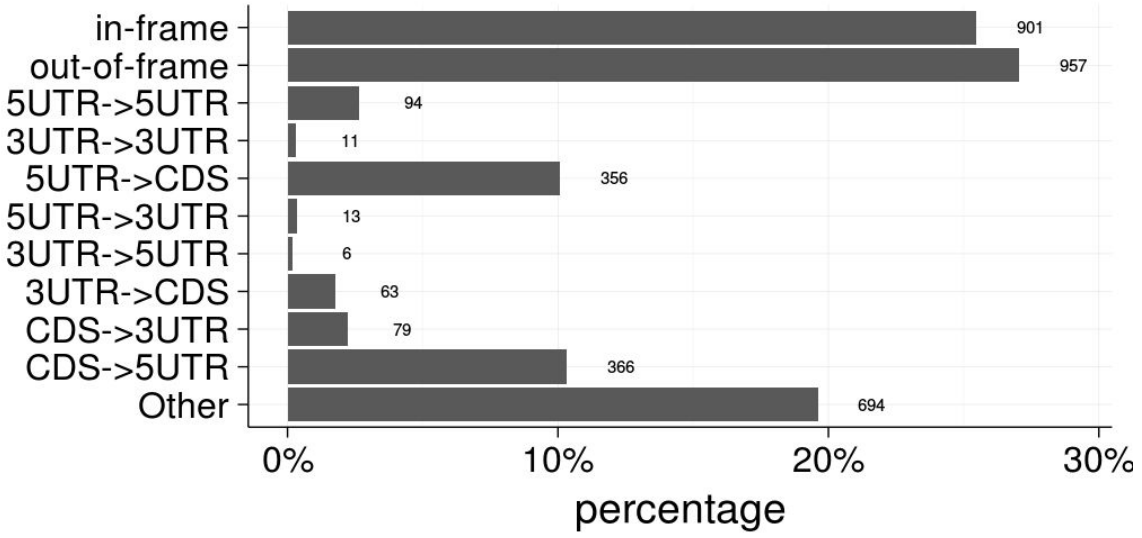
Supplementary Figure 8: Chromosomal distribution of bridged fusion pillar breakpoints. Each dot represents one of two pillar breakpoints, with x axis indicating the chromosome position. The Y axis simply represents random jitter variations to avoid overplotting for the dots.



Supplementary Figure 9: Percentages of breakpoint types distributed on different chromosomes. Red bars represent bridged fusion pillar breakpoints, blue bars the non-pillar SV breakpoints of bridged fusions, and green bars SV breakpoints of direct SV-supported fusions. Compared with direct fusion breakpoints, bridged fusion pillar breakpoints are enriched on the chromosome 12 (Odds Ratio: 1.62, Fisher exact test, p-value = 0.032). The chromosomes are ordered by the percentages of pillar breakpoints followed by the percentages of non-pillar breakpoints.

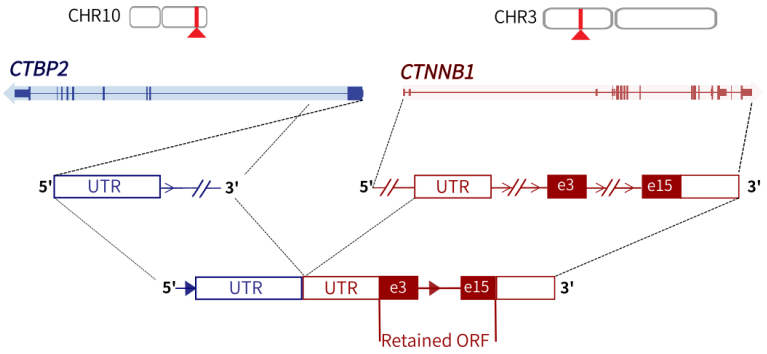


Supplementary Figure 10: Relative abundance of different types of fusions. Fusions involving non-coding genes were assigned to the “Other” category. The open reading frame for each fusion transcript was based on the dominant isoform or the longest CDS transcript of each fusion gene partner.

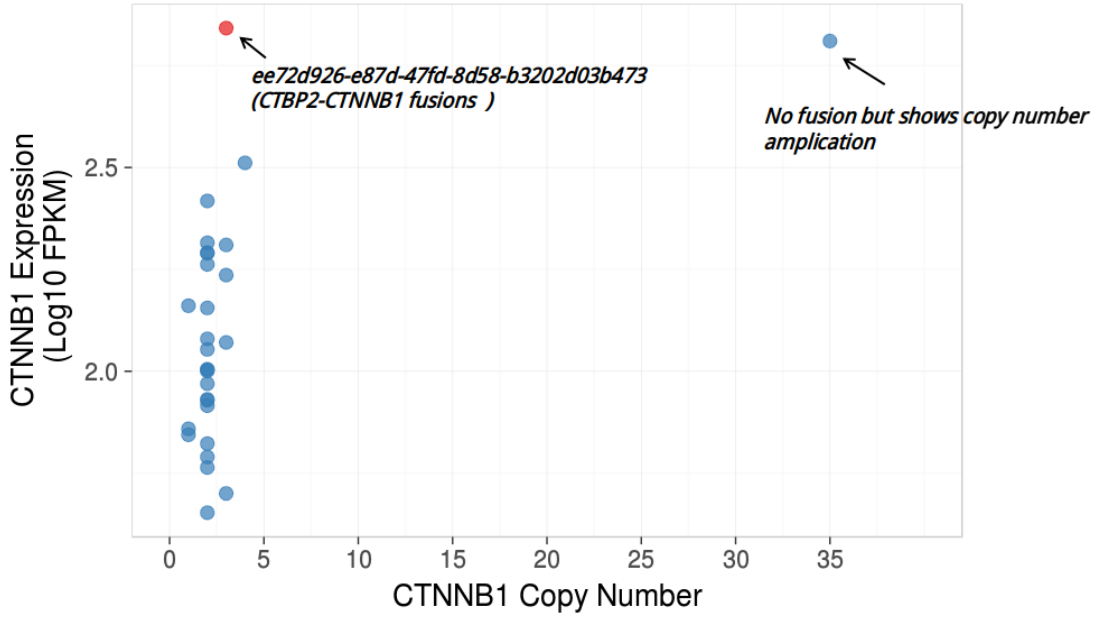


Supplementary Figure 11: *CTBP2-CTNNB1* as an example of "Retained ORF" fusion. A) a cartoon depicting the location, orientation and exon-intron architecture of the two genes involved. B) a scatter plot of *CTNNB1* DNA copy number versus mRNA expression across all ICGC gastric cancer samples.

A

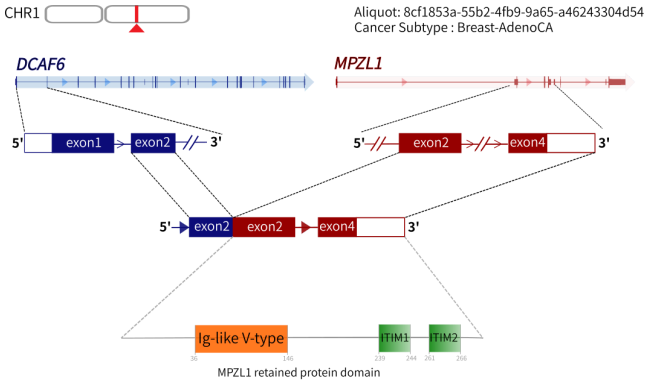


B

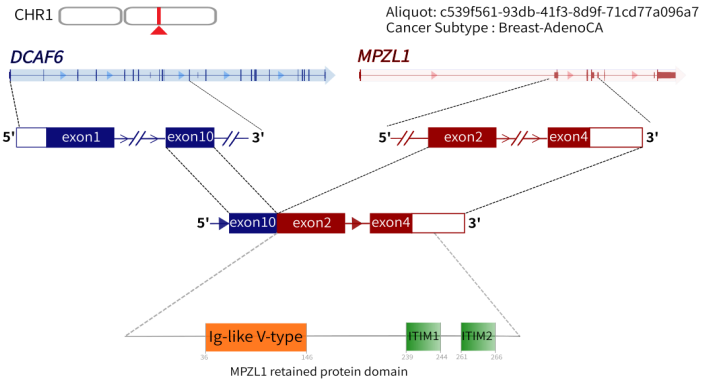


Supplementary Figure 12: The recurrent *DCAF6-MPZL1* fusion. A) Two independent *DCAF6-MPZL1* fusions observed in breast tumours, as supported by distinct SV events. Cartoon depicting the location, orientation and exon-intron architecture of the *DCAF6-MPZL1* fusion. Protein domains, including Ig-like V-type, ITIM1 and ITIM2, are retained in the fusion products. B) Combined *MPZL1* expression and DNA copy number analysis for *MPZL1*-involving fusions, showing the high expression of *MPZL1* in the fusion-containing sample.

A

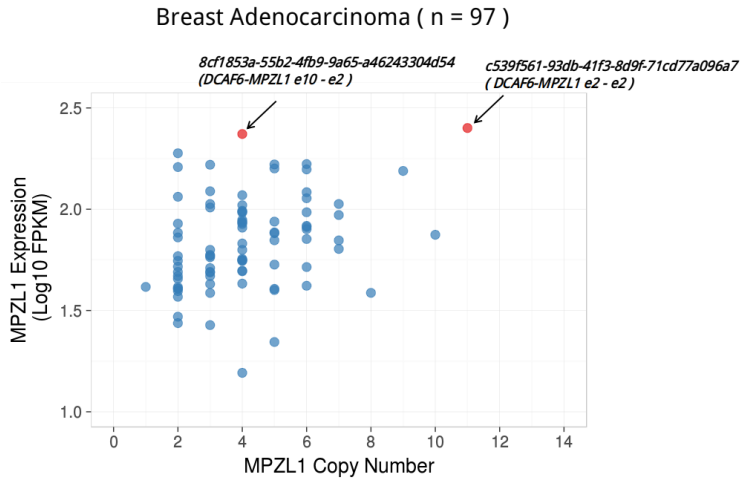


Cartoon depicting the location, orietation and exon-intron architeture of the DCAF6-MPZL1 fusion on the genome
Ig-like V-type: Ig-like V-Type domain;
ITIM1: immunoreceptor tyrosine-based inhibitor motif 1
ITIM2: immnuoreceptro tyrosine-based inhibitor motif 2

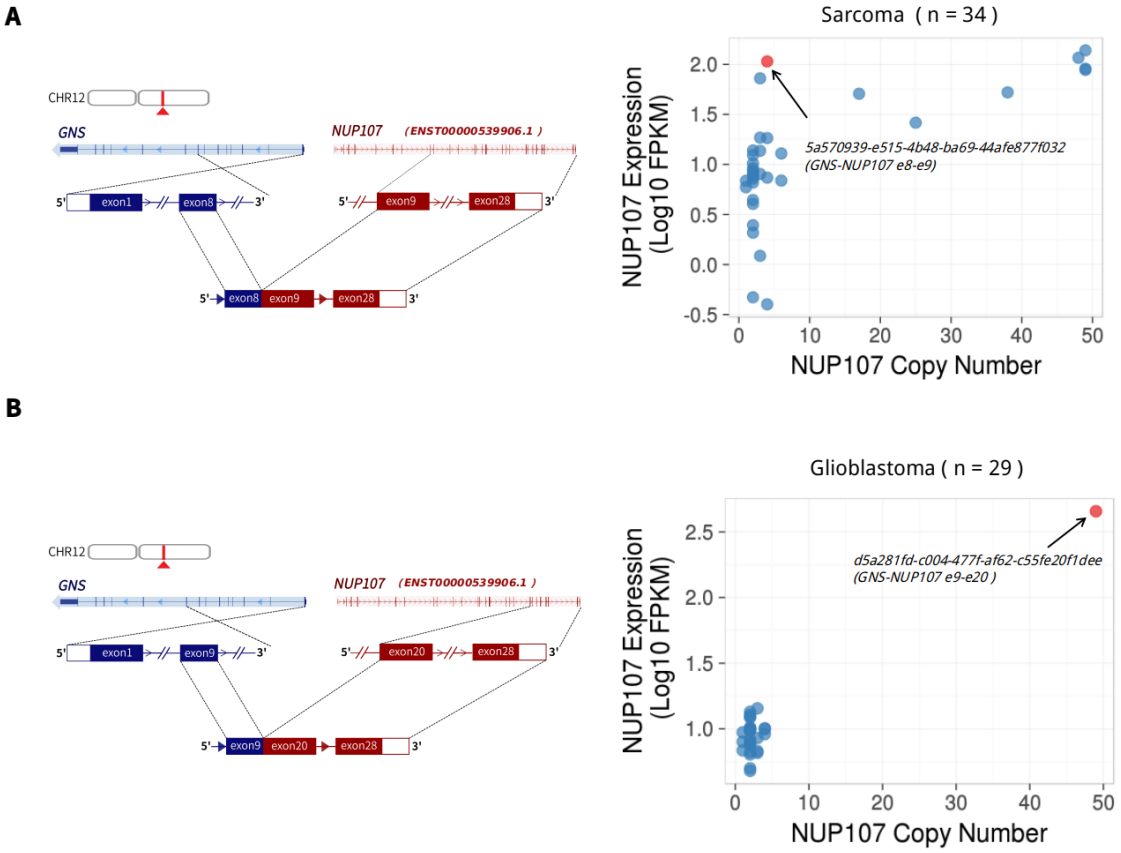


Cartoon depicting the location, orietation and exon-intron architeture of the DCAF6-MPZL1 fusion on the genome
Ig-like V-type: Ig-like V-Type domain;
ITIM1: immunoreceptor tyrosine-based inhibitor motif 1
ITIM2: immnuoreceptro tyrosine-based inhibitor motif 2

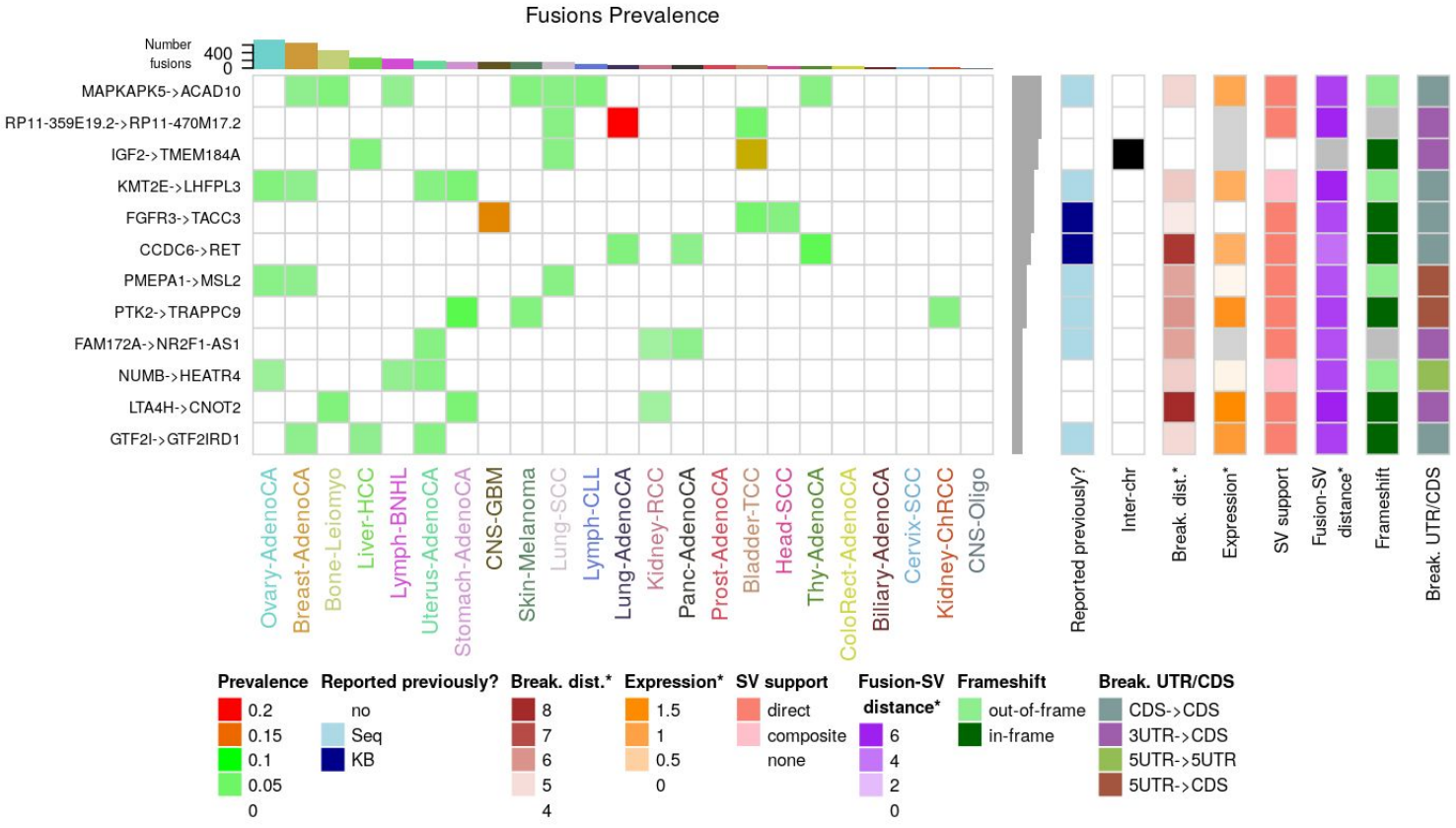
B



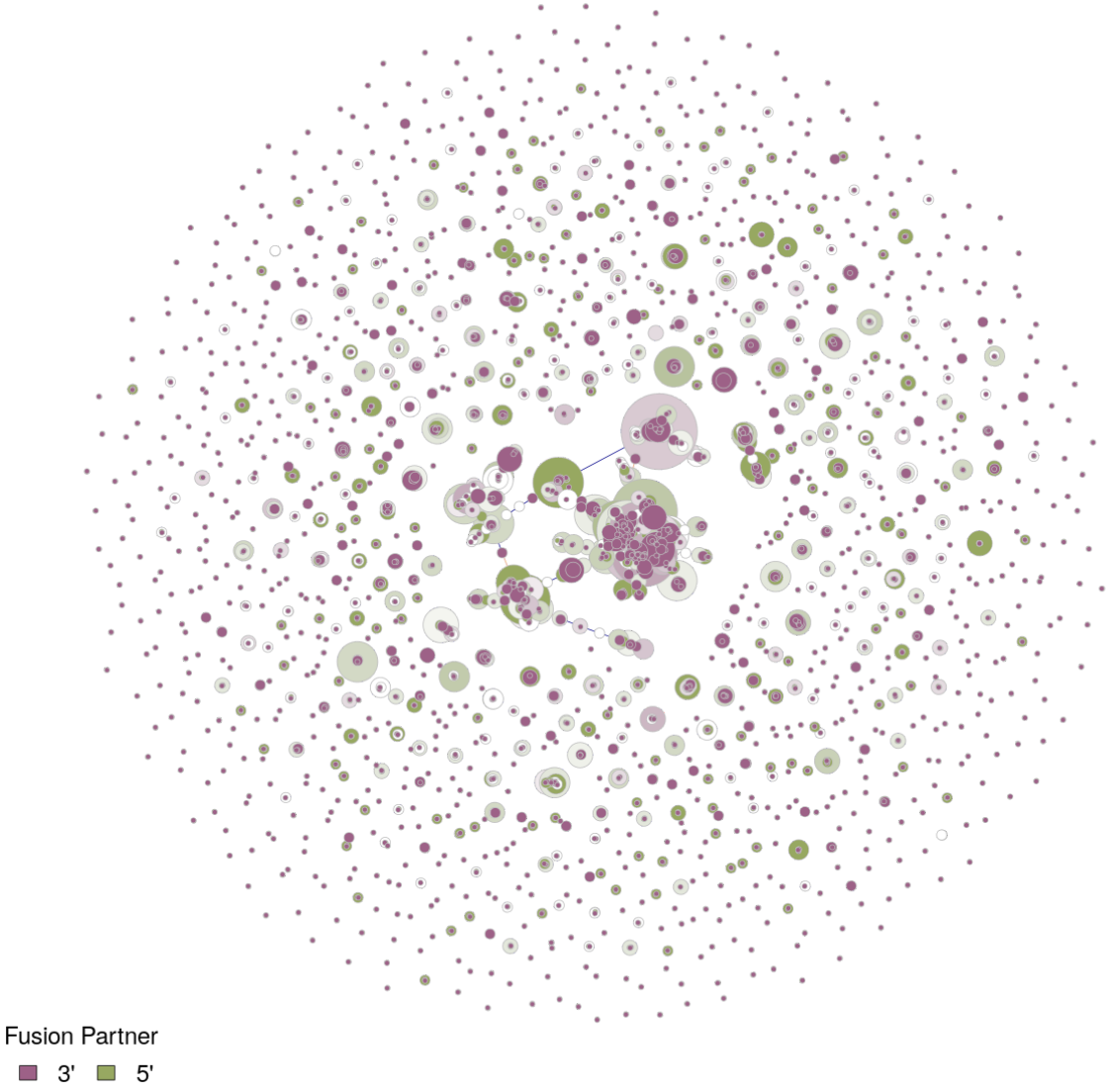
Supplementary Figure 13: The recurrent *GNS-NUP107* fusion. A) A *GNS-NUP107* fusion observed in sarcoma, with the underlying SV support. Cartoon depicting the location, orientation and exon-intron architecture of the *GNS-NUP107* fusion. The scatterplot shows *NUP107* DNA copy number versus mRNA expression across all ICGC sarcoma samples. B) Another *GNS-NUP107* fusion observed in glioblastoma, as supported by different SV events. Cartoon depicting the location, orientation and exon-intron architecture of the *GNS-NUP107* fusion. The scatterplot shows *NUP107* DNA copy number versus mRNA expression across all ICGC glioblastoma samples.



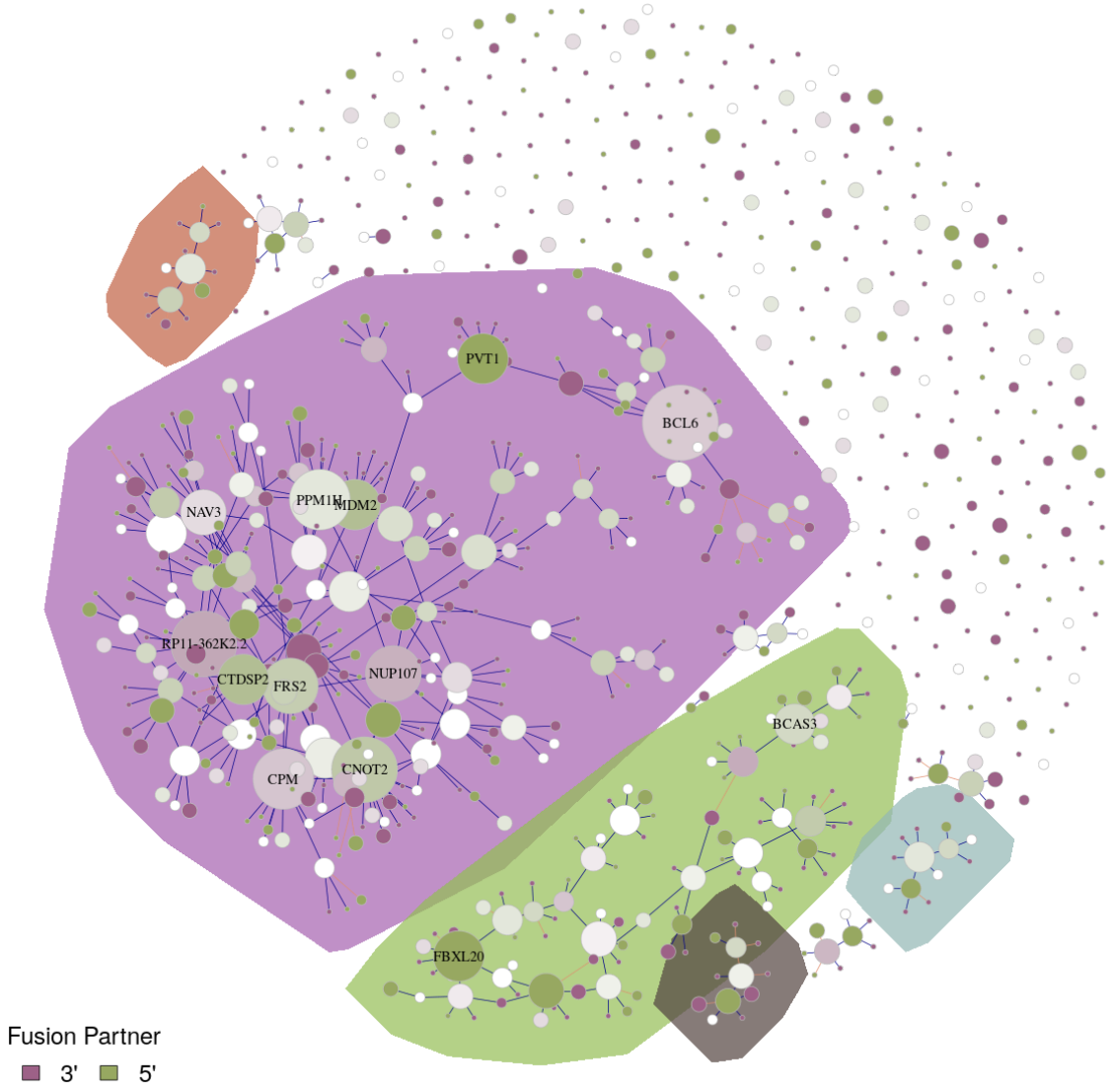
Supplementary Figure 14: The top 12 most recurrent fusions. Top histogram shows the total number of fusions per histotype while the left histogram corresponds to the number of times each fusion was detected across all histotypes. Each cell in the central matrix shows the prevalence of a fusion in a histological type. ChimerDB 3.0 [PMID:27899563] was used as a reference of previously reported gene fusions. The expression column presents the median expression of the putative transcript. The values with a star ('*') are in the log10 scale and distance unit is bp.



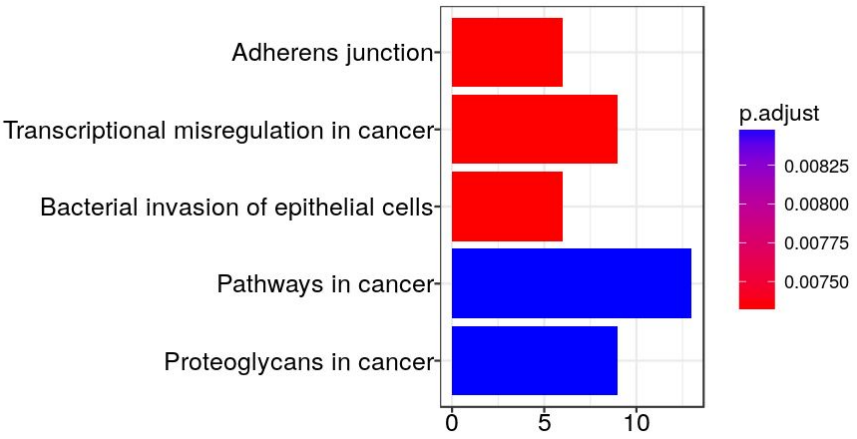
Supplementary Figure 15: Gene fusion pairing landscape. Genes are represented as nodes and the size of a node is proportional to the number of gene fusion partners. Two nodes are connected if one fusion was detected involving the two genes. Nodes and connections are only shown between genes with more than one gene partner (promiscuous genes). Non-promiscuous genes are not displayed. The color gradient indicates if a gene is involved more often in a fusion as 5' (red) or 3' (green) gene or both (white).



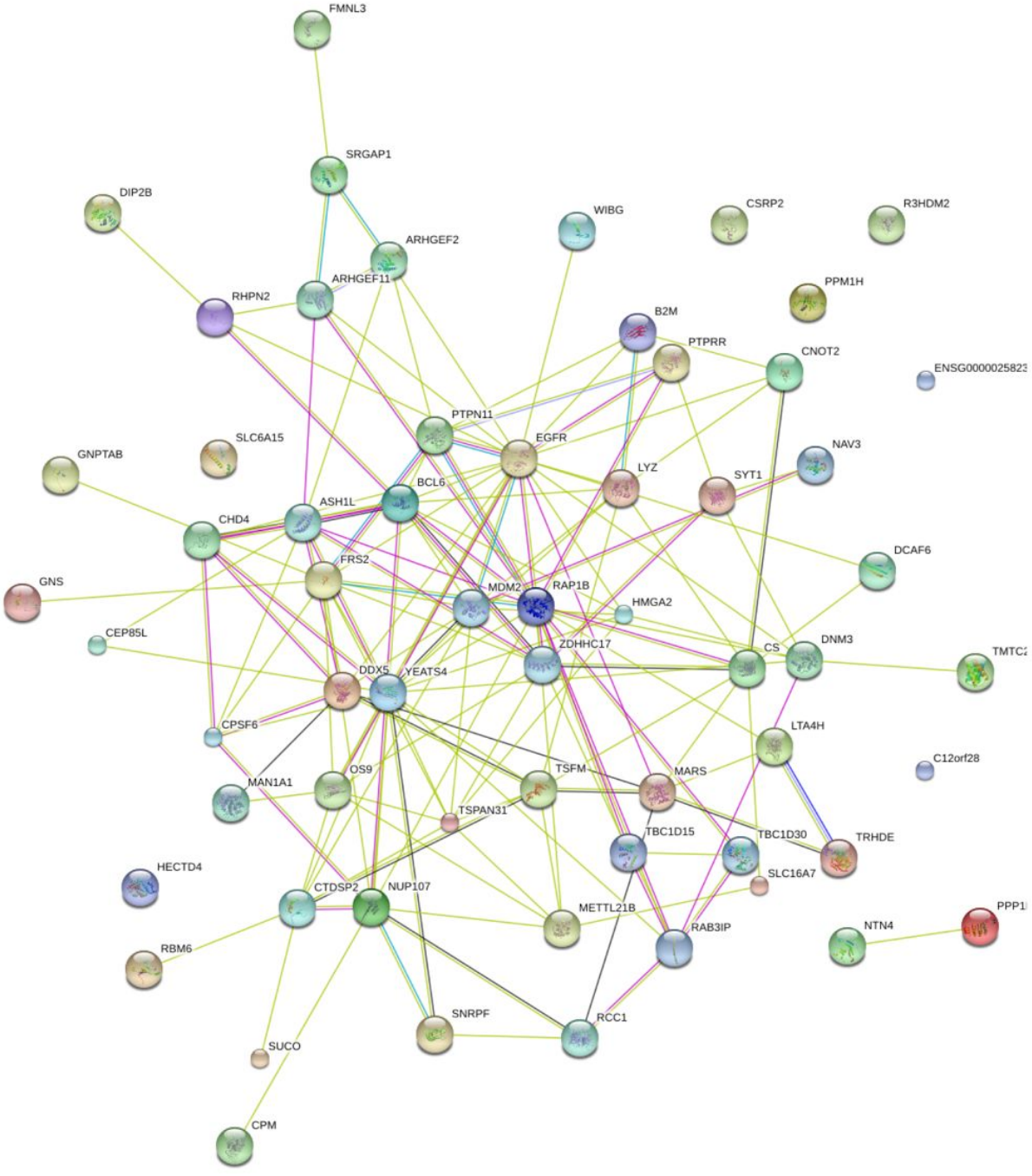
Supplementary Figure 16: Zoomed-in view of the larger clusters of promiscuous gene fusion partners. Genes are represented as nodes and the size of a node is proportional to the number of gene fusion partners. Two nodes are connected if one fusion was detected involving the two genes. Nodes and connections are only shown between genes with more than 3 gene partners (promiscuous genes). Non-promiscuous genes are not displayed. The color gradient indicates if a gene is involved more often in a fusion as 5' (red) or 3' (green) gene or both (white). The five connected clusters with at least 10 promiscuous genes are highlighted in different colors.



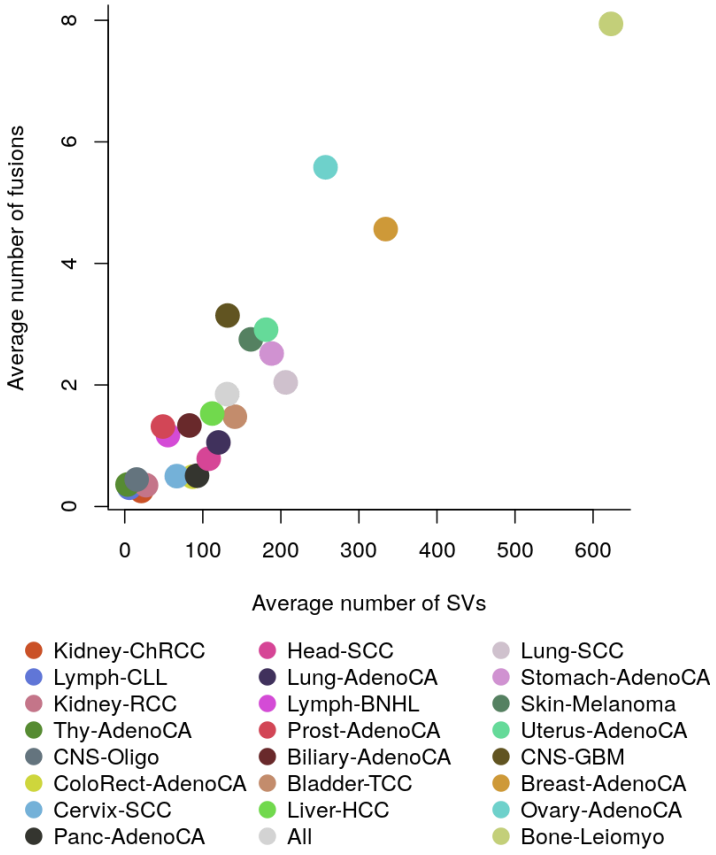
Supplementary Figure 17: Significantly enriched KEGG biological pathways for the genes in the clusters with at least 10 genes (Benjamini-Hochberg corrected p-values, corrected p-value cut-off of 0.01). The color of the bars reflects the corrected p-value and the size of the bars the number of genes.



Supplementary Figure 18: Protein-protein association network for the promiscuous genes in the greatest cluster based on the STRING database [PMID:25352553]. The nodes represent proteins; the edges represent known functional associations (based on experimental evidence and curated databases) and predicted functional associations (based, e.g., on text-mining, co-expression, gene fusions, ...). The network has significantly more interactions (150) than expected (91) (p-value < 0.0000001).



Supplementary Figure 19: Number of fusions per sample across histotypes. To avoid possible bias due to keeping fusions with SV support, only the 2268 fusions detected independently of SVs were considered (see online methods). Compared to considering all fusions the Pearson correlation between the average number of fusions and the average number of SVs per sample is marginally decreased from 0.96 (Fig. 1B) to 0.93.



Supplementary Figure 20: The breakpoints of promiscuous genes do not show enrichment in common fragile sites (two sided Wilcoxon rank sum test, P= 0.1239). Fusion genes with promiscuous gene partners are overlapped with human common fragile sites (online methods).