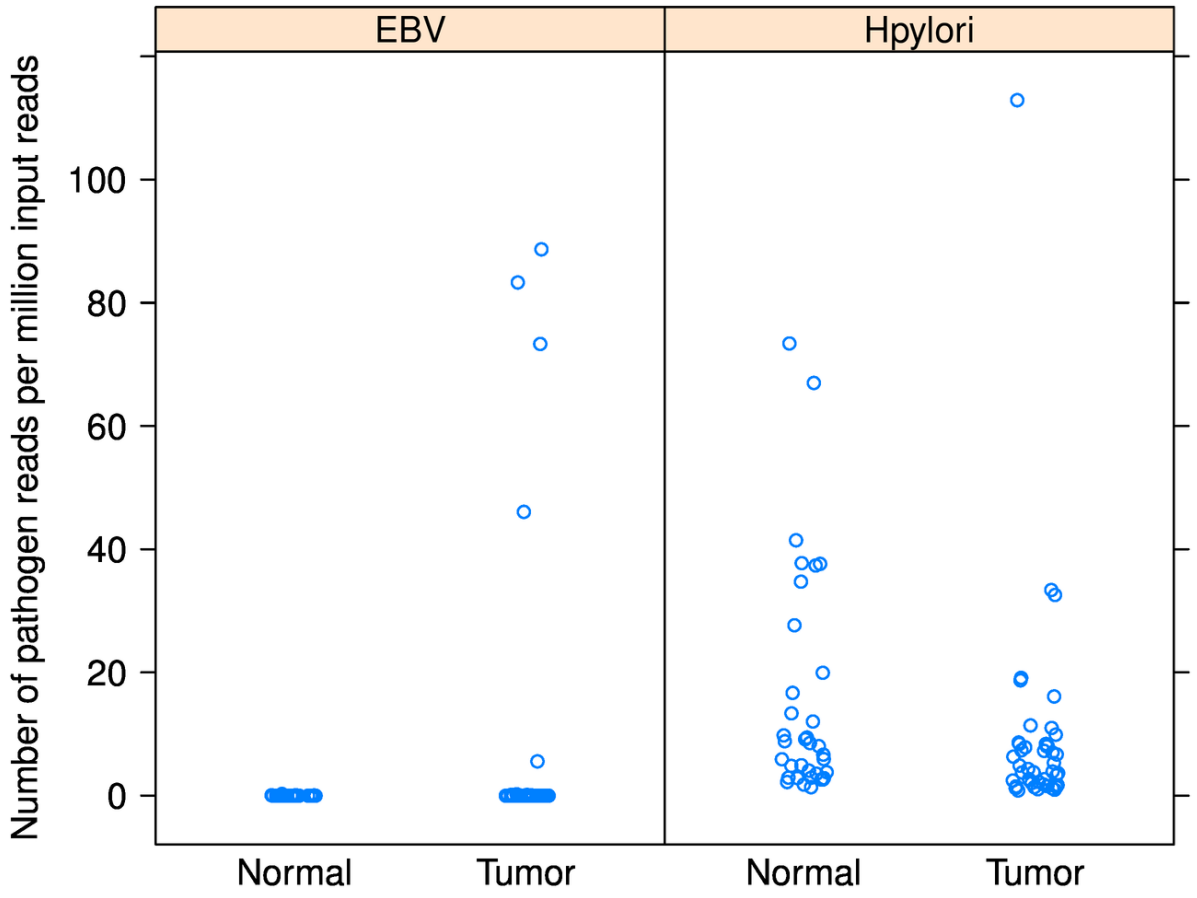
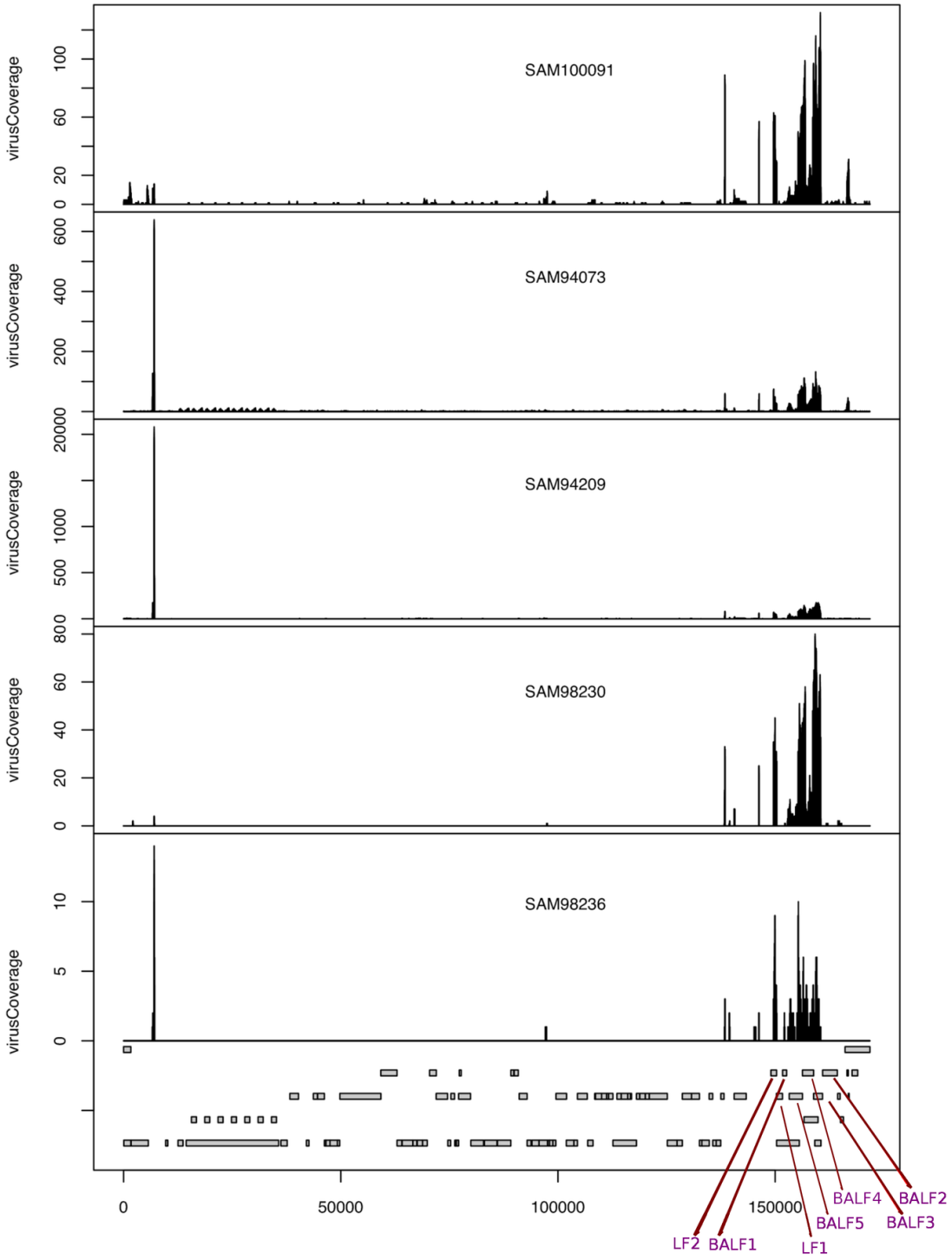
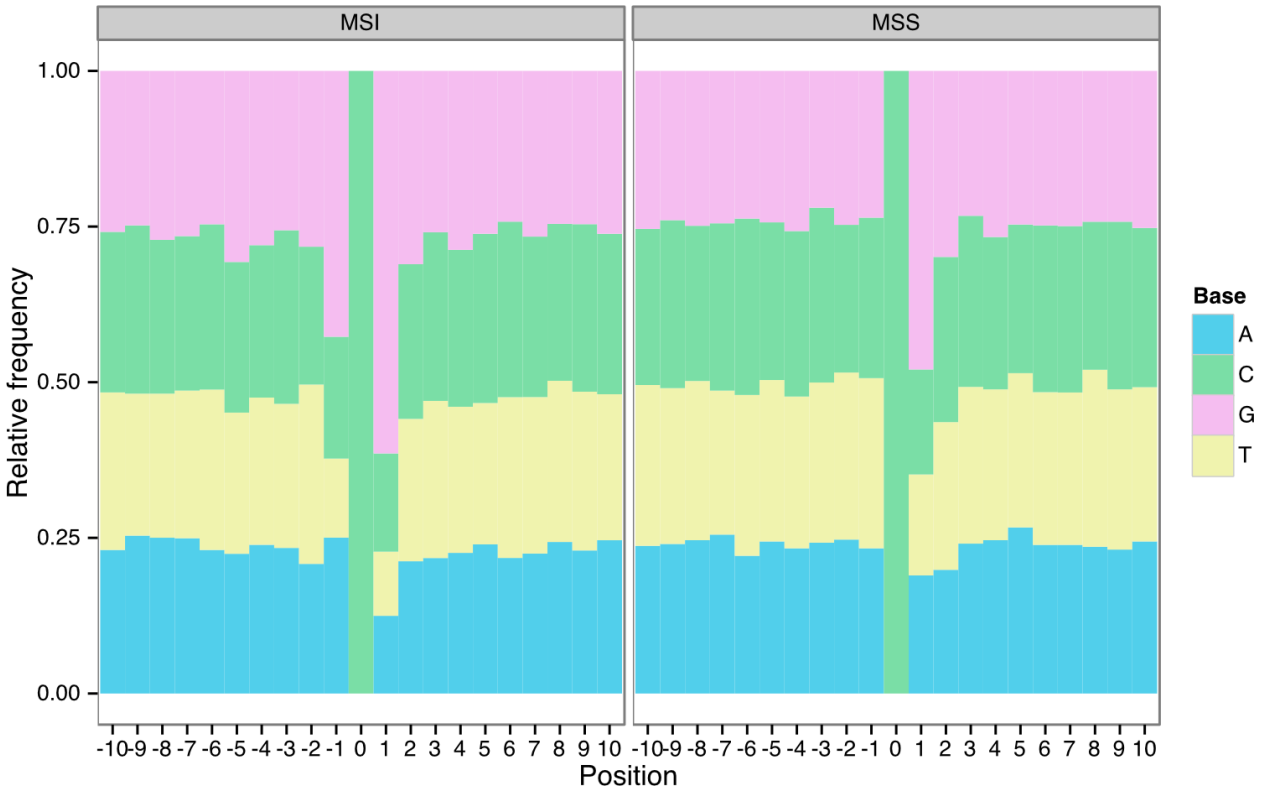


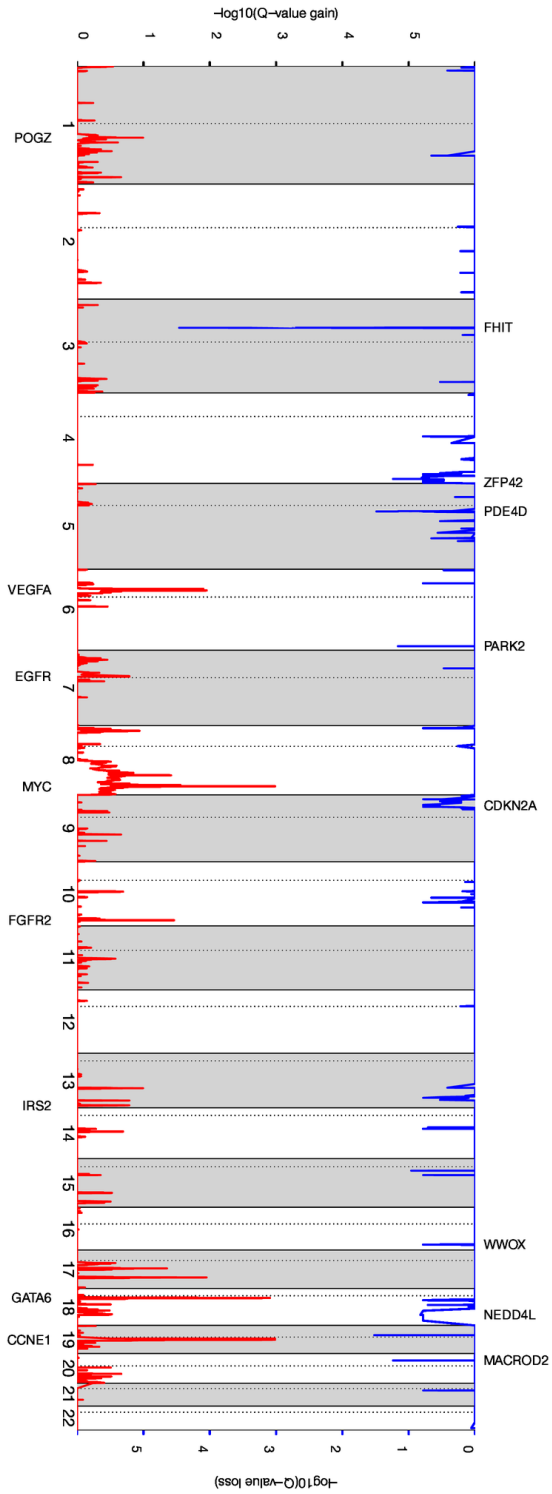
Supplementary Figure 1. Pathogen status from RNAseq data.
Five of the tumor but none of the normal tissue samples have detectable Epstein-Barr virus (EBV, *Human herpesvirus 7*) sequences. Sequences of *Helicobacter pylori,* the infection of which has been associated with gastritis and gastric cancer, were found in the majority of our gastric tissue samples, regardless of the normal/tumor status.



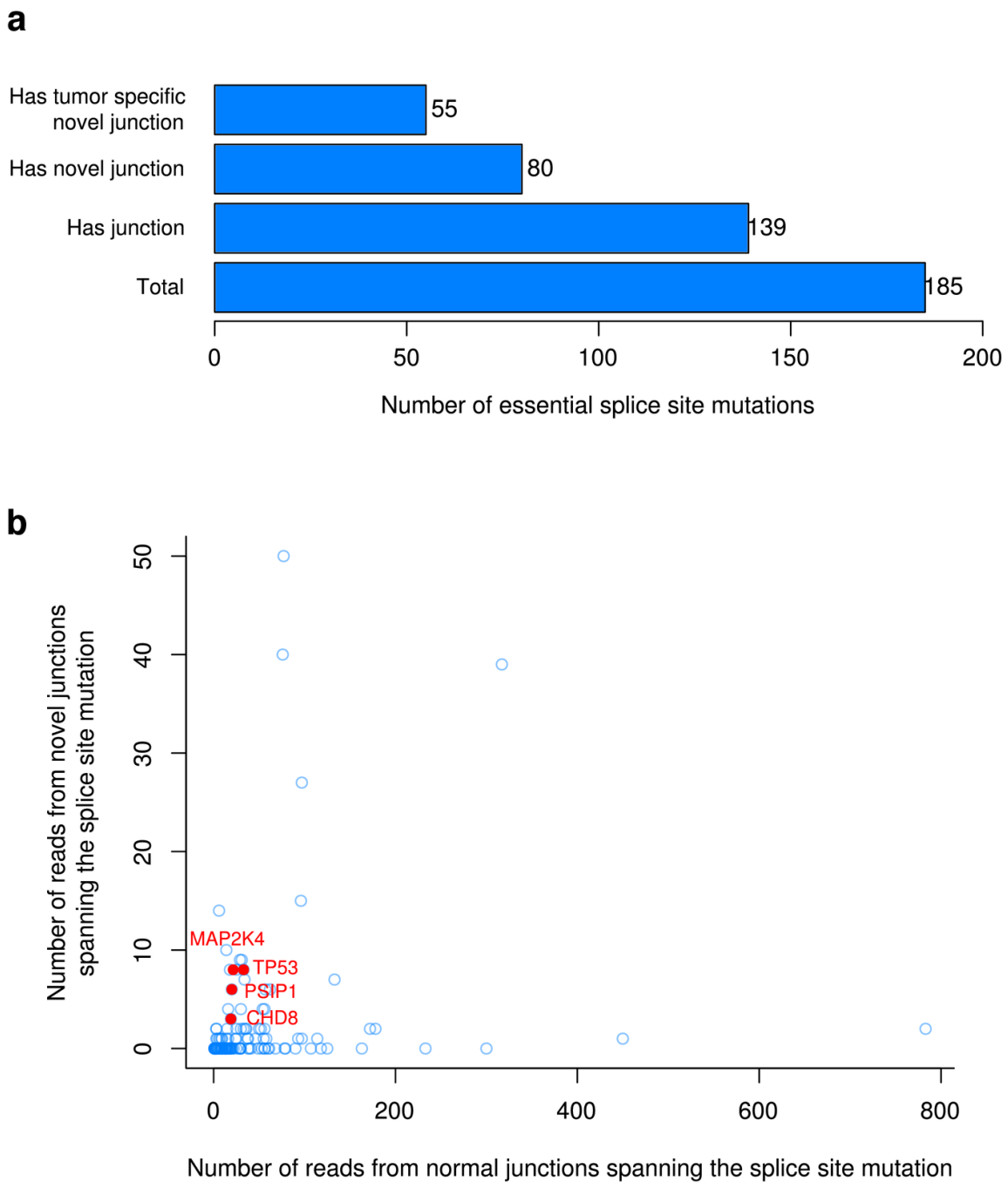
Supplementary Figure 2. Mapping of pathogen reads in the EBV genome.
Coverage of EBV genomes by RNAseq reads from the five gastric tumors. X-axis represents the coordinate in the EBV genome. Annotated EBV genes are shown on the bottom. Most EBV reads are mapped to the *BALF* gene cluster.



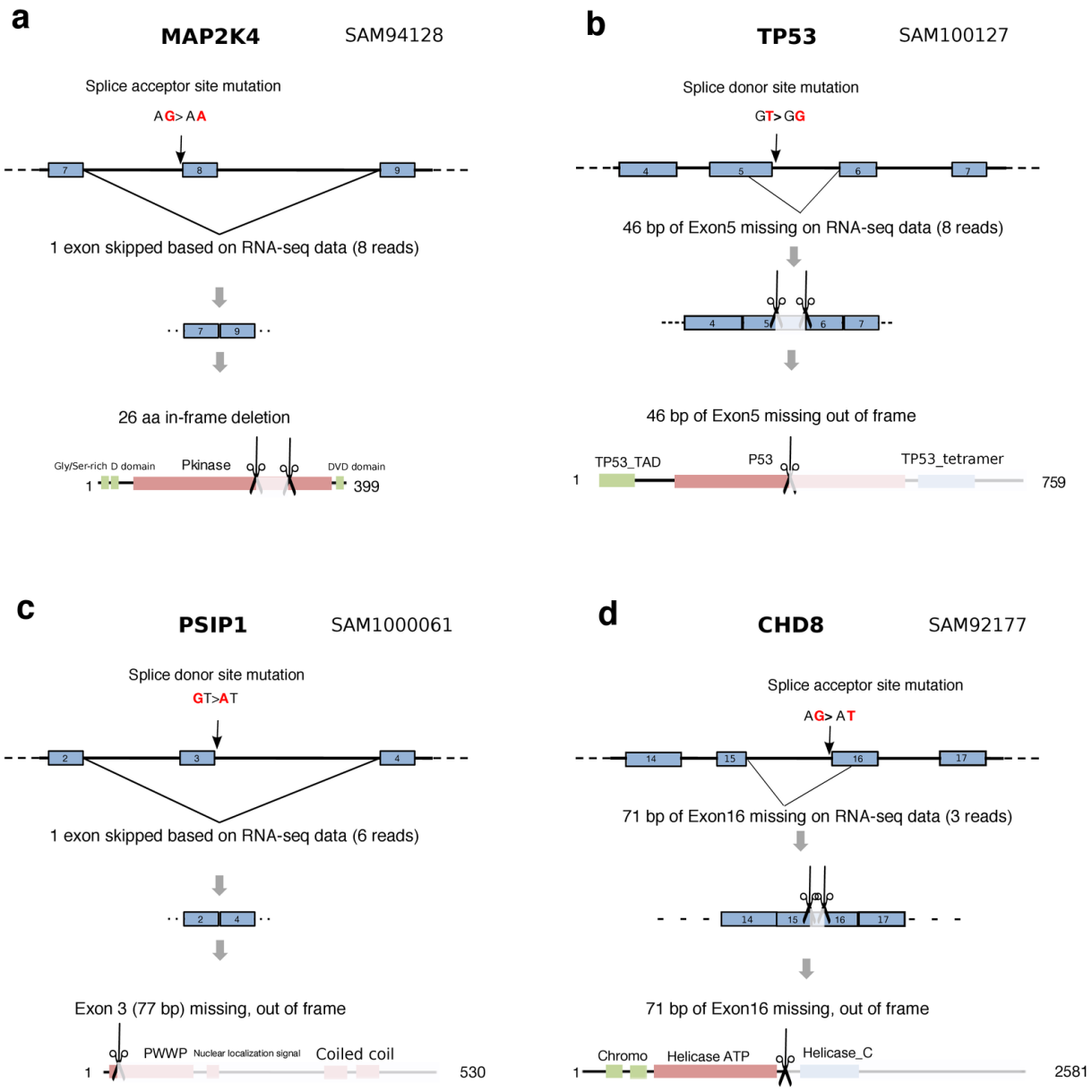
Supplementary Figure 3. Nucleotide context of somatic C-T transitions.
61% of C-T variations in MSI samples and 48% in MSS samples are observed to have a following G nucleotide, indicating a preference for CpG dinucleotide context.



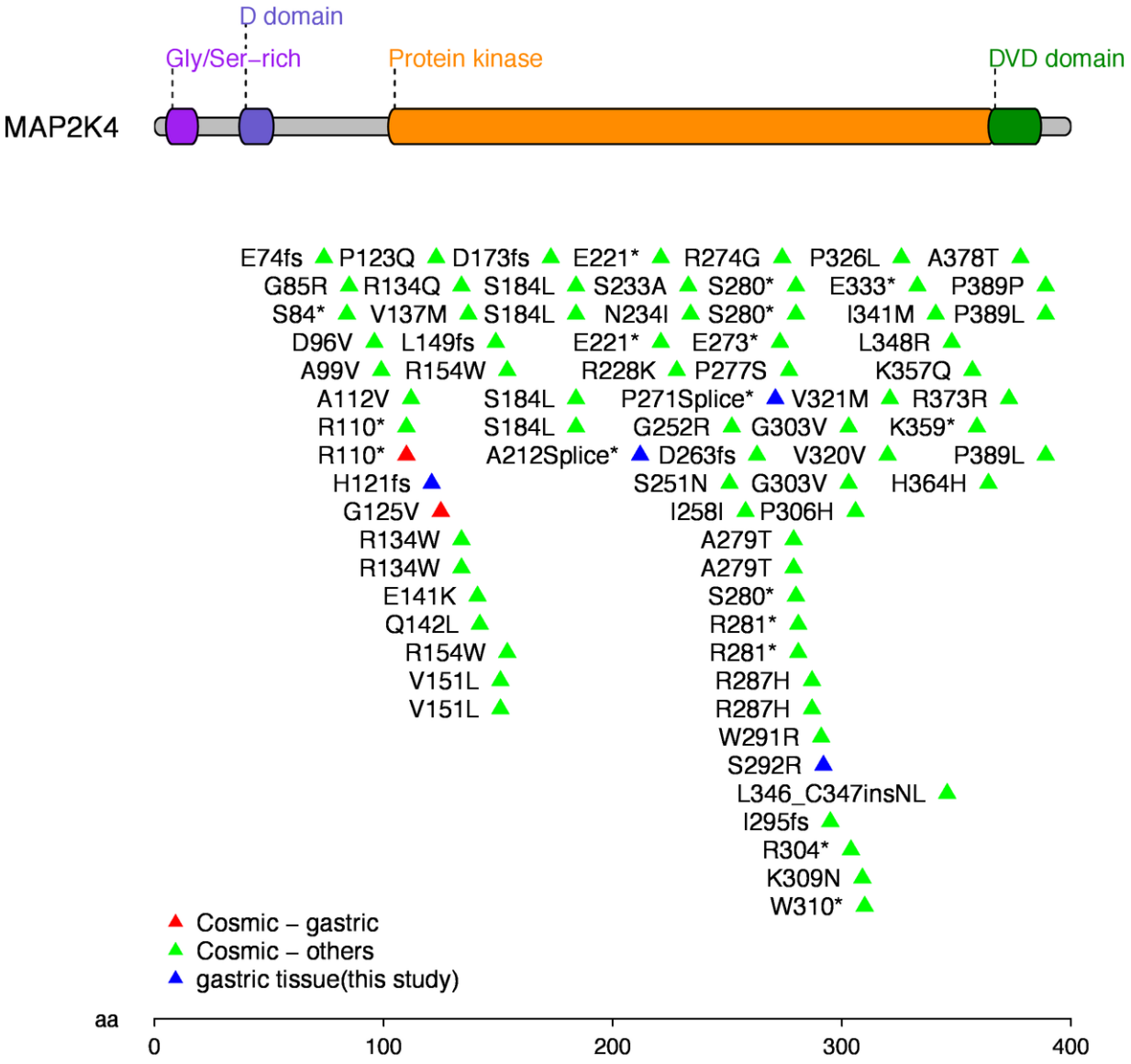
Supplementary Figure 4. GISTIC analysis of SNP Array-based DNA copy number showed recurrent gain and loss of multiple genes characteristic of gastric cancer.
Negative log10 False Discovery Rates are plotted in red for Gain and in blue for Loss. Genes representative of the most significant peaks are labeled.



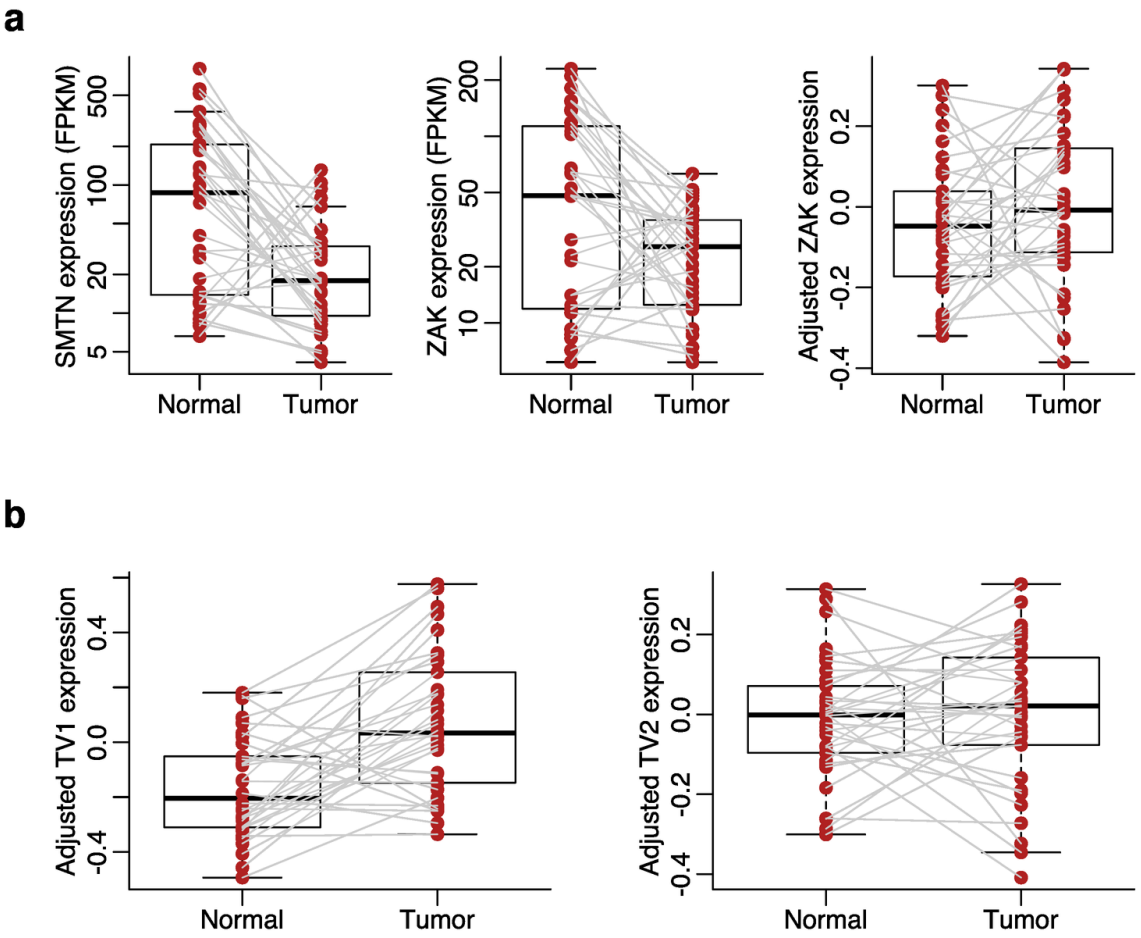
Supplementary Figure 5. Number of reads supporting the normal junctions vs. aberrant junctions spanning splice site mutations.
(a) Number of essential splice site mutations identified in gastric tumors, and the subset of the mutations associated with tumor-specific novel junctions. **(b)** Consistent with the heterozygous nature of these mutations, we typically observed both known splice junctions and aberrant ones spanning the same mutation locus, but the numbers of reads supporting the normal junctions are much higher than those supporting aberrant junctions (Supplementary Data 13).



Supplementary Figure 6. Four examples of aberrant splicing events associated with splice site mutations in cancer-related genes.
Top: the mutation in the essential splice site; middle: aberrant splicing events supported by RNAseq reads; bottom: predicted consequence at the protein-level.

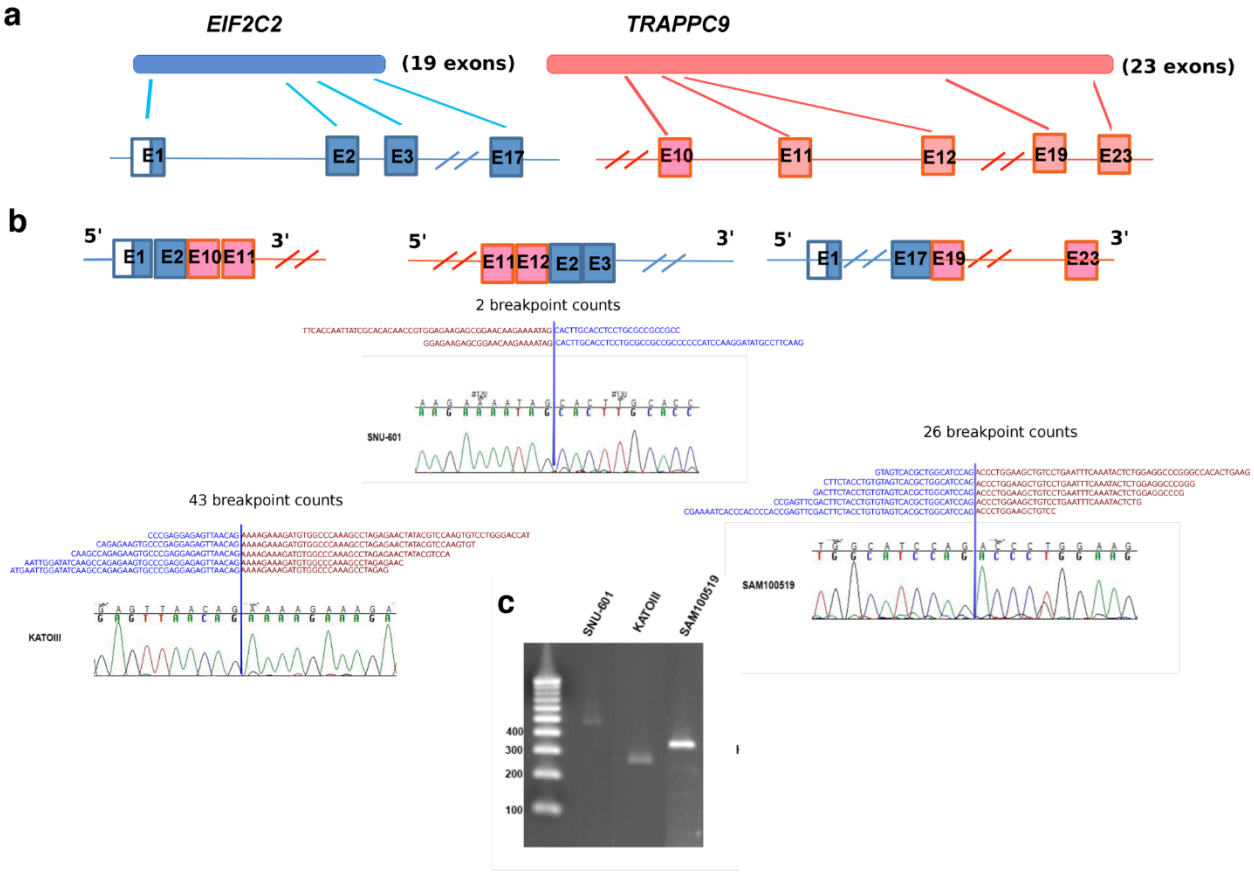


Supplementary Figure 7. Protein-altering mutations in *MAP2K4*.
Protein-altering mutations in MAP2K4 are compiled from the COSMIC[1] database and the current study, and plotted along the protein sequence.

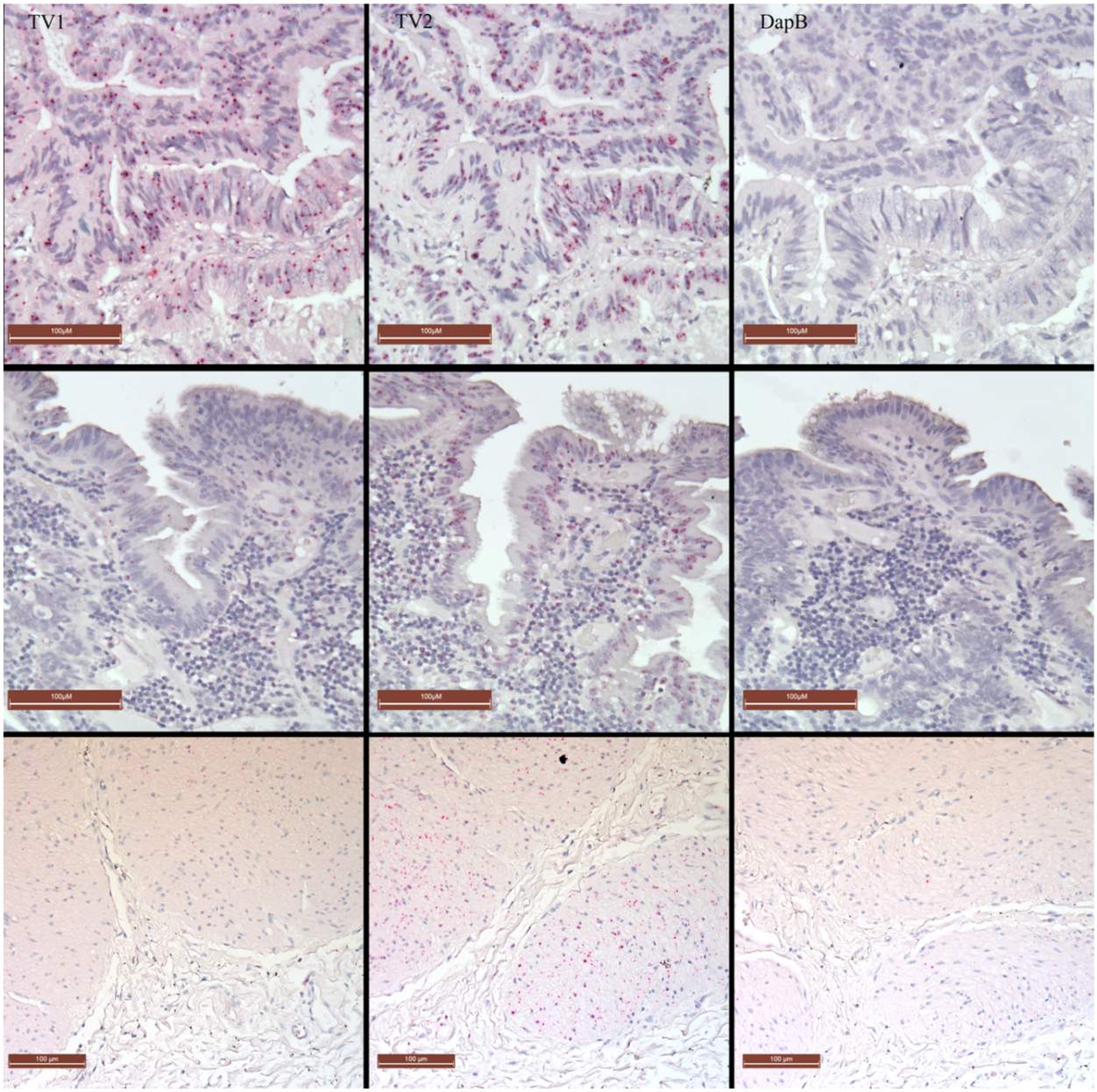


Supplementary Figure 8. ZAK isoform expression and adjustment for smooth muscle contamination.

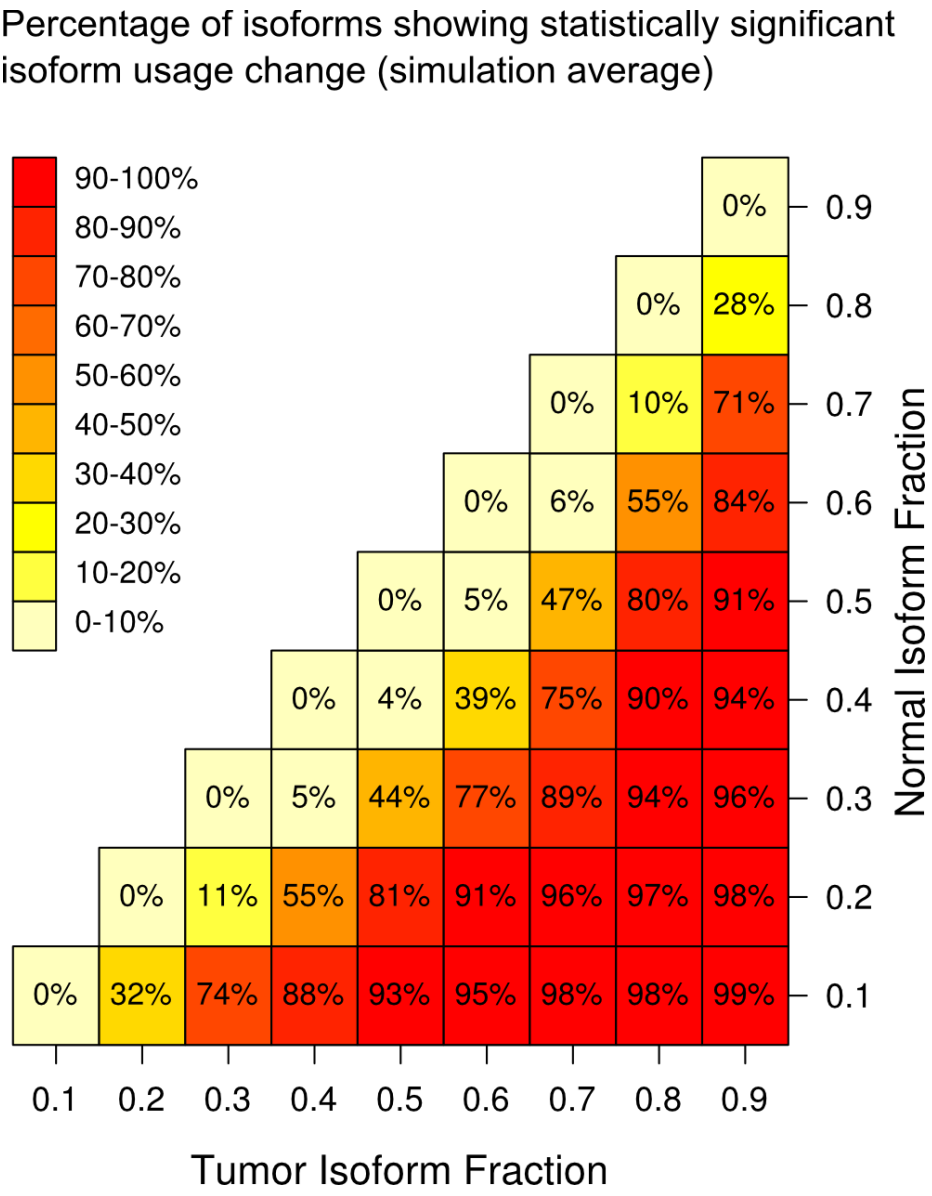
(a) Normal gastric tissue samples in our collection (34 normal samples and 34 tumors) have varying degree of smooth muscle contamination, as indicated by higher level and higher variability of smoothelin (SMTN) expression (left panel). Without correcting for the extent of contamination, ZAK gene expression appeared lower in the gastric tumors (middle panel); however, after the adjustment, ZAK expression was slightly higher in the tumors (right panel). The units of expression in the left and middle panels are FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values derived from cufflinks[2], and that in the right panel is the residuals of the linear model lm(log(ZAK) ~ log(SMTN)). Dots represent samples. Grey lines connect matched tumor and normal samples. The boxes in the box-and-whisker plots represent the interquartile range between the first and third quartiles; the dashed lines (whiskers) extend to the most extreme data points which is no more than 1.5 times the interquartile range from the box. **(b)** ZAK TV1 expression is higher in gastric tumors than the adjacent normal tissues (left panel), while TV2 expression remains unchanged (right panel). ZAK TV1 and TV2 expression level (FPKM) were derived using cufflinks[2]. To adjust for smooth muscle contamination, we fit a linear model lm(log(ZAK.TV) ~ log(SMTN)), and used the residuals of the model as the 'Adjusted expression' for ZAK TV1 and TV2. Dots represent samples. Grey lines connect matched tumor and normal samples.



Supplementary Figure 9. EIF2C2- TRAPPC9 fusions in gastric tumor and cell lines.
(a) Exon organization of the *EIF2C2* and *TRAPPC9* genes. **(b)** From left to right, three different fusion transcripts of *EIF2C2-TRAPPC9* were identified in gastric cell lines (KATOIII and SNU-601) and in a gastric tumor (SAM100519). PCR products (in panel **c**) as shown in the gel picture were submitted for Sanger sequencing. In each case, the Sanger sequences (bottom) confirmed the results from RNAseq (top).



Supplementary Figure 10. Isoform specific ZAK expression by in situ hybridization. Representative images show ZAK TV1 and ZAK TV2 expression in gastric carcinoma (top row), normal adjacent tumor mucosa (middle row) and gastric smooth muscle (bottom row). Chromogenic red staining denotes level of expression. Scale is depicted for each image. DapB is used as a non specific negative control. Scale bars represent 100 μm.



Supplementary Figure 11. Statistical power of negative binomial regression model on assessing differential isoform usage.
Isoform reads were simulated as described in Supplementary Material section 3. For a given underlying fraction of isoform reads (among all gene reads) in tumor and normal samples, negative binomial regression was employed to detect isoforms showing differential isoform usage. Fraction of isoforms with adjusted p-value < 0.1 is shown. This represents true positive rate for off-diagonal entries and false positive rate for diagonal entries. The power of our approach increases with fold change (distance from the diagonal) and with larger observed counts (e.g., higher power for 0.4 vs. 0.8 2-fold change than for 0.1 vs. 0.2). The results for simulations with the roles of tumor and normal samples exchanged (above diagonal) are similar and are omitted.

Supplementary Note 1: SNP array data generation and copy number analysis

Illumina HumanOmni2.5_4v1 arrays were used to assay 140 samples (49 primary tumor–normal pairs, 40 gastric cell line and 2 unpaired primary tumors) for genotype, DNA copy number and loss of heterozygosity (LOH) at ~2.5 million SNP positions. These samples all passed our quality control metrics for sample identity and data quality. A subset of 2,295,239 high-quality SNPs was selected for all analyses. After making modifications to permit use with Illumina array data, we applied the PICNIC algorithm to estimate total copy number, allele-specific copy number and LOH, as described[3]. Recurrent genomic regions with DNA copy gain and loss were identified using GISTIC, version 2.0[4].

GISTIC analysis of the 2.5M copy number data identified several recurrently amplified and deleted genes (Supplementary Figure 3). Fragile site genes *FHIT* and *WWOX* are frequently deleted as well as the tumor suppressor gene *CDKN2A*. Frequently amplified genes include *MYC, EGFR, VEGFA, FGFR2, CCNE1* and *GATA6. ERBB2* was not identified by GISTIC as recurrently amplified but we detected three samples with focal *ERBB2* amplifications and three others with broader gains.

For the DNA copy number analysis shown in Figure 2c, we further consider the relative DNA copy number by correcting for the overall cell ploidy. To do this, we median centered the copy number ratios for each sample using the median copy number ratio of all autosomes. We then used thresholds of 0.15 and -0.4 for calling gains and losses respectively. Such threshold consideration was calibrated based on estimated normal cell contamination in our tumor samples.

Supplementary Note 2: Gene expression analysis

Gene expression profiles of the gastric tumors and noncancerous tissues were investigated using RNAseq. Unsupervised hierarchical clustering effectively separated tumors from normal samples. Differential expression analysis on the count data was performed using the R package 'DESeq2'[5], which is based on a negative binomial distribution and uses shrinkage estimation for the variance of the distribution. The expression level of smoothelin (SMTN, a smooth muscle expression marker) was included in our model to correct for the extent of the smooth muscle inclusion. The DESeq2 design formula is: GeneCount ~ SMTN + patientID + CancerStatus, where SMTN is the normalized count (i.e. size factor adjusted count) for SMTN. 165 genes were

upregulated and 256 genes were downregulated in tumors using a stringent cutoff (fold change > 3 and FDR < 10^{-10}). Observed changes such as upregulation of TOP2A and CKS2[6] and downregulation of GKN1, GKN2 and TFF1[7,8] were largely consistent with previous reports. Notably, most of the canonical markers for gastric cancer were present in the set of differentially expressed genes including serum markers MUC5AC[9], Reprimo[10] and Pepsinogen C[11]. The prognostic marker CDH17[12] ($p=2.8*10^{-19}$) and the diagnostic marker INHBA[13] ($p=5.6*10^{-25}$) were also highly differentially expressed. Gene ontology enrichment analysis on the 421 differentially expressed genes revealed that a significant proportion of genes with higher expression in tumors are involved in cell cycle ($p=10^{-6}$). This agrees with an exon-array based study on 80 pairs of gastric cancer and adjacent noncancerous tissues[14]. Among the cell cycle associated genes with high expression in gastric cancer were canonical cancer genes such as the BUB1 mitotic checkpoint kinase and homeobox gene HOXA13. In contrast, digestion, transmembrane transport and ion transport were enriched in downregulated genes including pepsinogen A and cholecystokinin A receptor, which is also consistent with previous findings.

Supplementary Note 3: Validation of fusion transcripts

Identification of fusion transcripts was performed as described[3].

Gene fusions were validated by RT PCR with both tumor and matched normal samples. 500 ng of total RNA was reverse transcribed to complementary DNA (cDNA) with a High Capacity cDNA Reverse Transcription kit following manufacturer's instructions (Life Technologies). 50 ng of cDNA was amplified in a 25 µl reaction containing 0.5 µM of each primer, 300 µM of each deoxynucleoside triphosphate and 2.5 U of LongAmp Taq DNA polymerase (New England Biolabs). PCR was performed with an initial denaturation at 95°C for 3 min followed by 35 cycles of 95°C for 10 s, 56°C for 1 min and 68°C for 30 s, and a final extension step at 68°C for 10 min. Three µl of PCR product was run on 1.2% agarose gel to identify samples containing the gene fusion. Specific PCR products were purified with either a MinElute PCR Purification kit or MinElute Gel Extraction kit (Qiagen). The purified reactions were sequenced using Sanger sequencing on an ABI 3730 XL as per manufacturer's instructions (Life Technologies) with PCR primers specific to each fusion. The Sanger sequencing trace files were analyzed using Sequencher (Gene Codes Corp.).

In addition to the gene fusions identified in gastric tumor samples, we also found similar gene fusions containing the same partners in cancer cell lines. One of the examples (*EIF2C2* and *TRAPPC9*-derived fusion transcripts) was further analyzed (Supplementary Figure 7).

Stomach carcinoma cell lines, KATOIII and SNU-601 expressed RNA fusions of the *EIF2C2* and *TRAPPC9* genes. Both cell lines were cultured in RPMI 1640 supplemented with 10% FBS (Sigma). Total RNAs from these two cell lines was extracted from a confluent 10-cm plate using RNeasy kit following manufacturer's instructions (Qiagen). 1 µg of total RNA was reverse transcribed as described above. 50 ng of cDNA was then amplified in a 20 µl reaction containing 0.5 µM of each primer and Phusion High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs). PCR was performed with an initial denaturation at 98°C for 1 min followed by 30 cycles of 98°C for 15 seconds, 58°C for 30 seconds and 72°C for 30 s, and a final extension step at 72°C for 10 min. One microliter of PCR product was run on a 2% E-gel (Invitrogen) to identify samples containing the gene fusion.

EIF2C2 and *TRAPPC9* genes are tandem adjacent genes ~75 kb apart in chromosome 8. The fusion transcripts that involve these two genes have different structure in the clinical sample and in the cancer cell lines. In the tumor SAM100519, exon 17 of the *EIF2C2* was fused to exon 19 of *TRAPPC9*, leading to a C-terminal truncated EIF2C2 protein (757 amino acids) fused to an N-terminal truncated *TRAPPC9* (297 amino acids). In the KATO III stomach carcinoma cell line, exon 2 of *EIF2C2* was fused to exon 10 of *TRAPPC9*, resulting in out of frame truncation of *EIF2C2* after exon 2. In an additional stomach carcinoma cell line SNU-601, exon 12 of *TRAPPC9* was fused to exon 2 of *EIF2C2*, generating a fusion protein composed of the first 687 amino acids of TRAPPC9 protein and almost the entire EIF2C2 protein missing just the first 7 amino acids. Interestingly, both *EIF2C2* and *TRAPPC9* transcripts are highly expressed in SNU-601 cells compared to other gastric cell lines, suggesting that the fusion event might result in overexpression of the two genes.

Supplementary Methods

Simulation of negative binomial regression models for assessing differential isoform usage

To assess the performance of our approach to detecting differential isoform usage, we conducted a set of simulations. We selected a subset of genes with only 2 transcript isoforms (to avoid added complexity) and with sufficient coverage across the samples in our panel. For

each sample and each gene we set the underlying proportion of expressed transcripts corresponding to shorter isoform to be 0.1, 0.2, …, 0.9 (with the corresponding proportion of longer isoform expressed molecules being 0.9, 0.8, …, 0.1). Given the underlying shorter isoform proportion F_{short}, we simulated the observed number N_{short,sample} of fragments from this isoform in a given sample as following negative binomial distribution with mean E(N_{short,sample})=N_{gene,sample}*F_{short}, where N_{gene,sample} is the number of reads aligned to the gene in the sample of interest in the original (non-simulated) data set. We also supplied the dispersion parameter (as described below), which together with the mean completely parametrized the distribution. Using this negative binomial distribution, in each simulation we produced the desired number of fragments n_{short,sample} and then generated the actual fragments by sampling 75bp paired-end fragments with constant insert size of 150bp uniformly at random from the (spliced) isoform sequence. Fragments from longer of the two gene isoforms were generated analogously, with F_{long}=1-F_{short} within a given simulation, and a dispersion estimate specific to the longer isoform was used to parametrize the distribution. Fragments thus simulated were aligned to the genome using same settings as the original data. 5 simulations were performed for each combination of simulation settings.

For each isoform, we estimated isoform-specific value of dispersion to be used within negative binomial model as follows. We assumed that when randomly generating a set of fragments from an isoform, each fragment will have the same probability of being isoform-specific, independent of other fragments, and that this process gave rise to our tally of isoform-specific read counts in the original data. Under this assumption, the dispersion observed among isoform-specific counts can be used as an estimate of the dispersion among all isoform fragment counts (including non-isoform-specific). We therefore used our counts of isoform-specific fragments across all samples (tumor and normal) to estimate the dispersion, after adjusting for observed gene counts within each sample. This dispersion parameter was then used to generate number of isoform fragments during simulations.

For each isoform, and for any two levels F_{tumor} and F_{normal} in 0.1, 0.2, …, 0.9, we had 5 sets of simulated samples where F_{isoform,sample}=F_{tumor} in all tumor samples and F_{isoform, sample}=F_{normal in all normal samples}. We applied our mixed effects negative binomial regression model to each of those 5 sets and obtained 5 estimates of the fraction of isoforms that are found to show statistically significant differential isoform usage. The estimate of our detection power is shown in **Supplementary Figure 11**. We find that in cases of F_{tumor}=F_{normal}, only 2/5875 isoforms tested across simulations showed Bonferroni-

adjusted p-value < 0.1. Conversely, the detection power was very high for 2- or larger fold changes.

In situ hybridization of ZAK isoforms

Non-isotopic *in situ* hybridization (ISH) was performed on 4 µm FFPE sections using QuantiGene® ViewRNA ISH Tissue Assay (Affymetrix/Panomics) following the manufacturer's protocol on a Tecan platform equipped to carry out non-isotopic *in situ* hybridization.

Gene-specific probe sets for detection of human ZAK isoform 1 mRNA (VA1-15607) and ZAK isoform 2 mRNA (VA1-15608), target region 1190-2158 and 1626-2774 respectively in Genbank accessions NM_016653 & NM_133646 were used on tissue samples. A probe set to *Bacillus subtilis* dihydropicolinate reductase (dapB) (VF1-11712), target region 1363-2044 in Genbank accession L38424 was used as a negative control. Probe sets specific to human (VA1-10203) Ubiquitin C, target region 342-1275 in Genbank accession NM_021009 was used as a positive control for comparing overall mRNA levels.

Horseradish peroxidase (HRP) conjugated label probe was used, followed by TSA™ (tyramide signal amplification) to increase sensitivity (Perkin Elmer NEL748001KT). Briefly, TSA Plus DIG stock solution (digoxigenin) was diluted 1:50 in 1x Plus Amplification Diluent and applied to sections and incubated for 10 minutes at room temperature. This was followed by incubation with anti-DIG-AP (Roche 11093274910) diluted 1:500 in TNB blocking buffer with 4% lamb serum (Gibco, 16070-096) for 30 minutes at room temperature. Vulcan Fast Red substrate (Biocare, FR805S) was used for chromogenic detection.

Hybridized target mRNAs were visualized with bright field microscopy using an Olympus BX51 microscope equipped with a Qimaging Retiga SRV camera. Metamorph software was used to capture images.

Supplementary References

1. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–950 (2011).
2. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).
3. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488,** 660–664 (2012).

4. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
5. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
6. El-Rifai, W., Frierson, H. F., Jr, Harper, J. C., Powell, S. M. & Knuutila, S. Expression profiling of gastric adenocarcinoma using cDNA array. *Int. J. Cancer J. Int. Cancer* **92,** 832–838 (2001).
7. Hippo, Y. *et al.* Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res.* **62,** 233–240 (2002).
8. Moss, S. F. *et al.* Decreased expression of gastrokine 1 and the trefoil factor interacting protein TFIZ1/GKN2 in gastric cancer: influence of tumor histology and relationship to prognosis. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **14,** 4161–4167 (2008).
9. Xu, Y., Zhang, L. & Hu, G. Potential application of alternatively glycosylated serum MUC1 and MUC5AC in gastric cancer diagnosis. *Biol. J. Int. Assoc. Biol. Stand.* **37,** 18–25 (2009).
10. Bernal, C. *et al.* Reprimo as a potential biomarker for early detection in gastric cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **14,** 6264–6269 (2008).
11. Kon, O. L. *et al.* The distinctive gastric fluid proteome in gastric cancer reveals a multi-biomarker diagnostic profile. *BMC Med. Genomics* **1,** 54 (2008).
12. Lee, H.-J. *et al.* Gene expression profiling of metaplastic lineages identifies CDH17 as a prognostic marker in early stage gastric cancer. *Gastroenterology* **139,** 213–225.e3 (2010).
13. Takeno, A. *et al.* Integrative approach for differentially overexpressed genes in gastric cancer by combining large-scale gene expression profiling and network analysis. *Br. J. Cancer* **99,** 1307–1315 (2008).
14. Cui, J. *et al.* An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. *Nucleic Acids Res.* **39,** 1197–1207 (2011).