Genomic rearrangements near genes leading to upregulation across a diverse subset of human cancers

Yiqun Zhang[1], Fengju Chen[1], Chad J. Creighton[1,2,3,4]

1. Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA.

2. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

3. Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA.

4. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.

Correspondence to:
Chad J. Creighton (creighto@bcm.edu)

Abbreviations: PCAWG, the Pan-Cancer Analysis of Whole Genomes project; SV, Structural Variant;

Running header: Genomic rearrangements altering transcription

Abstract

Using a dataset of somatic Structural Variants (SVs) in cancers from 2658 patients—1220 with corresponding gene expression data—we identified hundreds of genes for which the nearby presence (within 100kb) of an SV breakpoint was associated with altered expression. For the vast majority of these genes, expression was increased rather than decreased with corresponding SV event. Well-known up-regulated cancer-associated genes impacted by this phenomenon included *TERT*, *MDM2*, *CDK4*, *ERBB2*, and *IGF2*. For specific gene categories, associated SVs were found to potentially impact on the order of 2-4% of cases across various types. SVs upstream of *TERT* involved ~3% of cancer cases and were most frequent in liver-biliary, melanoma, sarcoma, stomach, and kidney cancers. For many genes, SVs were significantly associated with either an increase or decrease in the numbers of enhancer regulatory elements near the promoter. DNA methylation near the gene promoter was often increased with nearby SV breakpoint.

Introduction

Functionally relevant DNA alterations in cancer extend well beyond exomic boundaries. One notable example of this involves *TERT*, for which both non-coding somatic point mutations in the promoter or genomic rearrangements in proximity to the gene have been associated with *TERT* up-regulation[1-3]. Genomic rearrangements in cancer are common and often associated with copy number alterations[4,5]. Conceivably, breakpoints associated with rearrangement could alter the regulation of nearby genes, e.g. by disrupting specific regulatory elements or by translocating regulatory elements from elsewhere in the genome into close proximity to the gene. The Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative has recently assembled over 2600 whole cancer genomes from multiple independent studies representing a wide range of cancer types. These data represent an opportunity for us to survey these cancers for somatic structural variants (SVs) with breakpoints located in proximity to genes. For a sizeable subset of these cancers, data from other platforms in addition to whole genome sequencing, such as RNA expression or DNA methylation, are available for integrative analyses. With a genome-wide analysis, information from multiple genes may be leveraged effectively in order to identify common features involving the observed disrupted regulation of genes impacted by genomic rearrangement.

Results

Widespread impact of somatic SVs on gene expression patterns in cancer

Inspired by recent observations in kidney cancer[3,6], neuroblastoma[7,8], and B-cell malignancies[9], of recurrent genomic rearrangements affecting the chromosomal region proximal to *TERT* and resulting in its up-regulation, we sought to carry out a pan-cancer analysis of all coding genes, for ones appearing similarly affected by rearrangement. We referred to a dataset of somatic SVs called for whole cancer genomes of 2658 patients, representing more than 20 different cancer types and compiled and harmonized by the PCAWG initiative from 47 previous studies (Table S1). Gene expression profiles were available for 1220 of the 2658 patients. We set out to systematically look for genes for which the nearby presence of an SV breakpoint could be significantly associated with changes in expression. In addition to the 0-20 kb region upstream of each gene (previously involved with rearrangements near *TERT*[3]), we also considered SV breakpoints occurring within a gene body, 0-20kb upstream of a gene, 20-50kb upstream of a gene, 50-100kb upstream of a gene, or 0-20kb downstream of a gene (Figure 1a). (SVs located within a given gene were not included in the other upstream or downstream SV sets for that same gene.) For each of the above SV groups, we assessed each gene for correlation between associated SV event and expression. As each cancer type as a whole would have a distinct molecular signature[10], and as genomic rearrangements may be involved in copy alterations[4], both of these were factored into our analysis, using linear models.

For each of the genomic regions relative to genes that were considered (i.e. genes with at least three samples associated with an SV within the given region), we found widespread associations between SV event and expression, after correcting for expression patterns associated with tumor type or copy number (Figure 1b and Table S2). For gene body, 0-20kb upstream, 20-50kb upstream, 50-100kb upstream, and 0-20kb downstream regions, the numbers of significant genes at p<0.001 (corresponding to estimated false discovery rates[11] of less than 5%) were 518, 384, 416, 496, and 302, respectively. For each of these gene sets, many more genes were positively correlated with SV event (i.e. expression was higher when SV breakpoint was present) than were negatively correlated (on the order of 95% versus 5%). Permutation testing of the 0-20kb upstream dataset (randomly shuffling the SV event profiles and computing correlations with expression 1000 times) indicated that the vast majority of the significant genes observed using the actual dataset would not be explainable by random chance or multiple testing (with permutation results yielding an average of 30 "significant" genes with standard deviation of 5.5, compared to 384 significant genes found for the actual dataset). Without correcting for copy number, even larger numbers of genes with SVs associated with increased expression were found (Figure 1b), indicating that many of these SVs would be strongly associated with copy gain. Many of the genes found significant for one SV group were also significant for other SV groups (Figure 1c).

Key driver genes in cancer impacted by nearby SVs

Genes with increased expression associated with nearby SVs included many genes with important roles in cancer (Table 1), such as *TERT* (significant with p<0.001 for regions from 0-20kb downstream to 20-50kb upstream of the gene), *MYC* (significant for gene body SVs), *MDM2* (regions from 0-20kb downstream to 50-100kb upstream), *CDK4* (0-20kb downstream and 20-100kb upstream), *ERBB2* (gene body to 50-100kb upstream), and *IGF2* (0-20kb downstream and 50-100kb upstream). Genes with decreased expression associated with SVs located within the gene included *PTEN* (n=50 out of 1220 cases with an SV and expression data), *STK11* (n=15), *KEAP1* (n=5), *TP53* (n=22), *RB1* (n=55), and *SMAD4* (n=18), where genomic rearrangement would presumably have a role in disrupting important tumor suppressors. Examining the set of genes positively correlated (p<0.001) with occurrence of SV upstream of the gene (for either 0-20kb, 20-50kb, or 50-100kb SV sets), enriched gene categories (Figure 1d) included G-protein coupled receptor activity (70 genes), telomerase holoenzyme complex (*TERT*, *PTGES3*, *SMG6*), eukaryotic translation initiation factor 2B complex (*EIF2S1*, *EIF2B1*, *EIF2B5*), keratin filament (15 genes), and insulin receptor binding (*DOK6*, *DOK7*, *IGF2*, *IRS4*, *FRS2*, *FRS3*, *PTPN11*). When taken together, SVs involving the above categories of genes would potentially impact a substantial fraction of cancer cases, e.g. on the order of 2-5% of cases across various types (Figure 1e). Gene amplification events (defined as five or more copies) could be observed for a number of genes associated with SVs, but amplification alone in many cases would not account for the elevated gene expression patterns observed (Figure 1e).

Translocations involving the region 0-100kb upstream of *TERT* were both inter- and intrachromosomal (Figure 2a and Table S3) and included 170 SV breakpoints and 84 cancer cases, with the most represented cancer types including liver-biliary (n=29 cases), melanoma (n=17 cases), sarcoma (n=15 cases), and kidney (n=9 cases). Most of these SV breakpoints were found within 20kb of the *TERT* start site (Figure 2b), which represented the region where correlation between SV events and *TERT* expression was strongest (Figure 2c, p<1E-14, linear regression model). In neuroblastoma, translocation of enhancer regulatory elements near the promoter was previously associated with *TERT* up-regulation[7,8]. Here, in a global analysis, we examined the number of enhancer elements[12] within a 0.5 Mb region upstream of each rearrangement breakpoint occurring in proximity to *TERT* (for breakpoints where the SV mate was oriented away from *TERT*). While for unaltered *TERT*, 21 enhancer elements are located 0.5 Mb upstream of the gene, on the order of 30 enhancer elements on average were within the 0.5 Mb region adjacent to the *TERT* SV breakpoint (Figure 2d), representing a significant increase (p<1E-6, paired t-test). A trend was also observed, by which SVs closer to the *TERT* start site were associated with a larger number of enhancer elements (Figure 2d, p=0.03, Spearman's correlation).

Consistent with observations elsewhere[4], genomic rearrangements could be associated here with copy alterations for a large number of genes (Figure 1b), including for genes of particular interest including *TERT* and *MDM2* (Figure 3a). However, copy alteration alone would not account for all observed cases of increased expression in conjunction with SV event. For example, with a number of key genes (including *TERT*, *MDM2*, *ERBB2*, *CDK4*), if all amplified cases (i.e. with five or more gene copies) were grouped into a single category, regardless of SV breakpoint occurrence, the remaining SV-involved cases showed significantly increased expression (Figure 3b). Regarding *TERT* in particular, a number of types of genomic alteration may act upon transcription, including upstream SV, *TERT* amplification[13], promoter mutations[1,2], promoter viral integration[14], and *MYC* amplification[15]. Within the PCAWG cohort of 2658 cancer cases, 933 (35%) were altered according to at least one of the above alteration classes, with each class being associated with increased *TERT* mRNA expression (Figure 3c). Upstream SVs in particular were associated with higher *TERT* as compared to promoter mutation or amplification events.

SVs associated with translocated enhancers and altered DNA methylation near genes

Similar to analyses focusing on *TERT* (Figure 2d), we examined SVs involving other genes for potential translocation of enhancer elements. For example, like *TERT*, SVs 0-20kb upstream of *CDK4* were associated with an increased number of upstream enhancer elements as compared to that of the unaltered gene (Figure 4a); however, SVs upstream of *MDM2* were associated with significantly fewer enhancer elements compared to that of the unaltered region (Figure 4a). For the set of 1233 genes with at least 7 SVs 0-20kb upstream and with breakpoint mate on the distal side from the gene, the numbers of enhancer elements 0.5 Mb region upstream of rearrangement breakpoints was compared with the number for the unaltered gene (Figure 4b and Table S4). Of these genes, 24% showed differences at a significance level of p<0.01 (with ~12 nominally significant genes being expected by chance). However, for most of these genes, the numbers of enhancer elements was decreased on average with the SV rather than increased (195 versus 103 genes, respectively), indicating that translocation of enhancers might help explain the observed upregulation for some but not all genes.

We went on to examine genes impacted by nearby SVs for associated patterns of DNA methylation. Taking the entire set of 8256 genes with associated CpG island probes represented on the 27K DNA methylation array platform (available for samples from The Cancer Genome Atlas), the expected overall trend[16] of inverse correlations between DNA methylation and gene expression were observed (Figure 1c). However, for the subset of 263 genes positively correlated in expression with occurrence of upstream SV ($p<0.001$, 0-20kb, 20-50kb, or 50-100kb SV sets), the methylation-expression correlations were less skewed towards negative ($p=0.0001$ by t-test, comparing the two sets of correlation distributions in Figure 1c). Genes positively correlated between expression and methylation included *TERT* and *MDM2*, with many of the same genes also showing a positive correlation between DNA methylation and nearby SV breakpoint (Figure 1d). Regarding *TERT*, a CpG site located in close proximity to its core promotor is known to contain a repressive element[8,17]; non-methylation results in the opening of CTCF binding sites and the transcriptional repression of *TERT*[17]. In the PCAWG cohort, SV breakpoints occurring 0-20kb upstream of the gene were associated with increase CpG island methylation (Figure 1e), while SV breakpoints 20-50kb upstream were not; *TERT* promoter mutation was also associated with increased methylation (Figure 1e).

Discussion

Using a unique dataset of whole genome sequencing and gene expression on tumors from a large number of patients and involving a wide range of cancer types, we have shown here how a phenomenon previously identified for *TERT* is global and would impact a large proportion of genes and of cancer cases. Genomic rearrangements involved with up-regulation of *TERT* have furthermore been shown here to involve a wider range of cancer types, expanded from previous observations made in kidney chromophobe and neuroblastoma. While many of the genes impacted by genomic rearrangement likely represent passengers rather than drivers of the disease, many other genes with canonically established roles in cancer would be impacted. Though any given gene may not be impacted in this way in a large percentage of cancer cases (the more frequently SV-altered gene *TERT* involving less than 3% of cancers surveyed), the multiple genes involved leads to a large cumulative effect in terms of absolute numbers of patients. The impact of somatic genomic rearrangements on global gene expression should therefore be regarded as an important driver mechanism in cancer, alongside that of somatic point mutations, copy number alteration, epigenetic silencing, and germline polymorphisms.

Future efforts can explore the mechanisms involved with specific genes deregulated by nearby genomic rearrangements. Regarding *TERT*-associated SVs, it has been suggested that the observed increase in DNA methylation of the affected region would be the result of massive chromatin remodeling brought about by juxtaposition of the *TERT* locus to strong enhancer elements[8]. However, not all genes found here to be deregulated by SVs would necessarily follow the same patterns as those of *TERT*. For example, not all of the affected genes would have repressor elements being inactivated by DNA methylation, and some genes such as *MDM2* do not show an increase in enhancer numbers with associated SVs but do correlate positively between expression and methylation. There is likely no single mechanism that would account for all of the affected genes, though some mechanisms may be common to multiple genes. Integration of other types of information (e.g. other genome annotation features, data from other platforms, or results of functional studies) may be combined with whole genome sequencing datasets of cancer, in order to gain further insights into the global impact of non-exomic alterations, where the datasets assembled by PCAWG in particular represent a valuable resource.

Acknowledgements

This work was supported in part by National Institutes of Health (NIH) grant P30CA125123 (C. Creighton) and Cancer Prevention and Research Institute of Texas (CPRIT) grant RP120713 C2 (C. Creighton).

Methods

Datasets

Datasets of structural variants (SVs), RNA expression, somatic mutation, and copy number were generated as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. In all, 2671 patients with whole genome data were represented in the PCAWG datasets, spanning a range of cancer types (bladder, sarcoma, breast, liver-biliary, cervix, leukemia, colorectal, lymphoma, prostate, eosophagus, stomach, cns, head/neck, kidney, lung, skin, ovary, pancreas, thyroid, uterus). For SVs, calls were made by three different data centers using different algorithms; calls made by at least two algorithms were used in the downstream analyses. For copy number, the calls made by the Sanger group were used. For somatic mutation of *TERT* promoter, PCAWG variant calls, as well as any additional data available from the previous individual studies[3,14,18,19], were used. *TERT* promoter viral integrations were obtained from ref[14]. Of the 2658 cases, RNA-seq data were available for 1220 cases. For RNA-seq data, alignments by both STAR and TopHat2 were used to generated a combined set of expression calls; FPKM-UQ values (where UQ= upper quartile of fragment count to protein coding genes) were used. Where a small number of patients had multiple tumor sample profiles, one profile was randomly selected to represent the patient. DNA methylation profiles had been generated for 771 cases by The Cancer Genome Atlas using either the Illumina Infinium HumanMethylation450 (HM450) or HumanMethylation27 (HM27) BeadChips (Illumina, San Diego, CA), as previously described[20]. To help correct for batch effects between methylation data platforms (HM450 versus HM27), we used the combat software[21]. For each of 8226 represented genes, an associated methylation array probe mapping to a CpG island was assigned; where multiple probes referred to the same gene, the probe with the highest variation across samples was selected for analysis.

Integrative Analyses

Gene boundaries and locations of enhancer elements were obtained from Ensembl (GRCh37 build). Enhancer elements found in multiple cell types (using Ensembl "Multicell" filter) were used[12]. For each SV 0-20kb upstream of a gene, the number of enhancer elements near the gene that would be represented by the rearrangement was determined (based on the orientation of the SV mate). Gene copies of five or more were called as amplification events. For a given set of SVs associated with a given gene, correlation between expression of the gene and the presence of an SV was assessed using a linear regression model (with log-transformed expression values). In addition to modeling expression as a function of SV event, models incorporating cancer type (one of the 20 major types listed above) as a factor in addition to SV, and models incorporating both cancer type and copy number were also considered. For these linear regression models, genes with at least three samples associated with an SV within the given region were considered. Genes for which SVs were significant (p<0.001) after correcting for cancer type and copy numbers were explored in downstream analyses. The method of Storey and Tibshirini[11] was used to estimate false discovery rates for significant genes.

Statistical Analysis

All P-values were two-sided unless otherwise specified.

Figure Legends

Figure 1. Structural Variants (SVs) associated with altered expression of nearby genes. **(a)** Numbers of SV breakpoints identified as occurring within a gene body, 0-20kb upstream of a gene, 20-50kb upstream of a gene, 50-100kb upstream of a gene, or 0-20kb downstream of a gene. For each SV set, the breakdown by alteration class is indicated. SVs located within a given gene are not included in the other upstream or downstream SV sets for that same gene. **(b)** For each of the SV sets from part a, numbers of significant genes (p<0.001), showing correlation between SV event and expression. Numbers above and below zero point of y-axis denote positively and negatively correlated genes, respectively. Linear regression models also evaluated significant associations when correcting for cancer type (red) and for both cancer type and gene copy number (green). **(c)** Heat map of significance patterns for genes from part b (from the model correcting for both cancer type and gene copy number). Red, significant positive correlation; blue, significant negative correlation; black, not significant (p>0.05); gray, not assessed (less than 3 SV events for given gene in the SV set). **(d)** Significantly enriched Gene Ontology (GO) terms for genes positively correlated (p<0.001) with occurrence of SV upstream of the gene (for either 0-20kb, 20-50kb, or 50-100kb SV sets). P-values by one-sided Fisher's exact test. **(e)** Patterns of SV versus expression for selected gene sets from part d (telomerase holoenzyme complex, top; eukaryotic translation initiation factor 2B complex, middle; insulin receptor binding, bottom). Differential gene expression patterns relative to the median across sample profiles. See also Tables S1 and S2.

Figure 2. SVs associated with *TERT* and its increased expression. (a) Circos plot showing all intra- and interchromosomal rearrangements 0-100kb from the TERT locus. **(b)** By cancer type, SV breakpoint locations within the region ~100kb upstream of *TERT*. Curved line connects two breakpoints common to the same SV. *TERT* promoter, CpG Islands, and CTCF and Myc binding sites along the same region are also indicated. **(c)** Where data available, gene expression levels of *TERT* corresponding to SVs from part a. Where multiple SVs were found in the same tumor, the SV breakpoint that was closest to the *TERT* start site was used for plotting the expression. Expression levels associated with *TERT* promoter (PM) mutation are also represented. Median expression for unaltered cases represents cases without *TERT* alteration (SV, mutation, amplification, viral integration) or *MYC* amplification. **(d)** Numbers of enhancer elements within a 0.5 Mb region upstream of each rearrangement breakpoint are positioned according to breakpoint location. For unaltered *TERT*, 21 enhancer elements are 0.5 Mb upstream of the gene. See also Table S3.

Figure 3. In addition to gene amplification and other genomic alteration events, SVs in proximity to key genes contribute to cases of high expression. (a) For 1220 cancer cases, copy number versus expression for *TERT* (left) and *MDM2* (right). Cases with SV events upstream of the gene are indicated. **(b)** Box plots of expression for *TERT*, *MDM2*, *ERBB2*, and *CDK4* by alteration class ("amp." or gene amplification: 5 or more copies, SV within gene body, SV 0-20kb downstream of gene, SV 0-20kb upstream of gene, SV 20-50kb upstream of gene, SV 50-100kb upstream of gene, or none of the above, i.e. "unaligned"). Cases with both SV and amplification are assigned here within the amplification group. **(c)** Left: Alterations involving *TERT* (SV 0-50kb upstream of gene, somatic mutation in promoter, viral integration within *TERT*

promoter, 5 or more gene copies of *TERT* or *MYC*) found in the set of 1220 cancers cases having both WGS and RNA data available. Right: Box plot of *TERT* expression by alteration class. “TERT amp” group does not include cases with other *TERT*-related alterations (SV, mutation, viral). P-values by Mann-Whitney U-test. n.s., not significant (p>0.05). Box plots represent 5%, 25%, 50%, 75%, and 95%. Points in box plots are colored according to tumor type as indicated in part c.

Figure 4. Translocation of enhancers and altered DNA methylation patterns associated with SVs near genes. (a) For *TERT*, *ERBB2*, *CDK4*, and *MDM2*, average number of enhancer elements within a 0.5 Mb region upstream of each rearrangement breakpoint (considering the respective SV sets occurring 0-20kb upstream of each gene), as compared to the number of enhancers for the unaltered gene. All differences are significant with p<0.01 (paired t-test). Error bars denote standard error. **(b)** For 1233 genes with at least 7 SVs 0-20kb upstream and with breakpoint mate on the distal side from the gene, histogram of t-statistics (paired t-test) comparing numbers of enhancer elements 0.5 Mb region upstream of rearrangement breakpoints with the number for the unaltered gene. Positive versus negative t-statistics denote greater versus fewer enhancers, respectively, associated with the SVs. **(c)** Histogram of t-statistics for correlation between gene expression and DNA methylation (by Pearson’s using log-transformed expression and logit-transformed methylation), for both the entire set of 8256 genes (blue) associated with CpG islands represented on DNA methylation array platform and the subset of 263 genes (red) on methylation platform and positively correlated in expression (p<0.001, “OE” for “over-expressed”) with occurrence of upstream SV (for either 0-20kb, 20-50kb, or 50-100kb SV sets). **(d)** Histogram of t-statistics for correlation between gene expression and SV event (by Pearson’s using logit-transformed methylation), for both the entire set of 2316 genes (blue) with at least 3 SVs 0-20kb upstream and represented on methylation platform and the subset of 97 genes (red) on methylation platform and positively correlated in expression (p<0.001) with occurrence of SV 0-20kb upstream. **(e)** DNA methylation of the CpG site cg02545192 proximal to the TERT core promoter in cases with SV 0-20kb or 20-50kb upstream of TERT, in cases with TERT promoter (PM) activation mutation, in cases with TERT amplification (“amp.”), and in the rest of cases (unaligned). P-values by t-test on logit-transformed methylation beta values. Box plots represent 5%, 25%, 50%, 75%, and 95%. Points in box plots are colored according to tumor type as indicated. See also Tables S4 and S5.

Supplementary Data Files

Table S1, related to Figure 1. Cancer cases examined in this study, with patient-level annotation regarding specific molecular features.

Table S2, related to Figure 1. Complete set of correlations between gene expression and nearby SV event, according to region examined and the regression model applied.

Table S3, related to Figure 2. SVs associated with *TERT*, with associated expression and numbers of enhancer elements within a 0.5 Mb region upstream of each rearrangement breakpoint.

Table S4, related to Figure 4. Numbers of enhancer elements within a 0.5 Mb region upstream of each rearrangement breakpoint, with associated enrichment patterns.

Table S5, related to Figure 4. Correlations between DNA methylation and gene expression, and correlations between DNA methylation and adjacent SV event (for breakpoints occurring 0-20kb upstream).

References

1 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959, doi:10.1126/science.1229259 (2013).

2 Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961 (2013).

3 Davis, C. *et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319-330 (2014).

4 Yang, L. *et al.* Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. *Am J Hum Genet* **98**, 843-856 (2016).

5 Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell.* **153**, 919-929 (2013).

6 Chen, F. *et al.* Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell Rep* **14**, 2476-2489 (2016).

7 Valentijn, L. *et al.* TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nature genetics* **47**, 1411-1414 (2015).

8 Peifer, M. *et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700-704 (2015).

9 Nagel, I. *et al.* Deregulation of the telomerase reverse transcriptase (TERT) gene by chromosomal translocations in B-cell malignancies. *Blood* **116**, 1317-1320 (2010).

10 Hoadley, K. *et al.* Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* **158**, 929-944 (2014).

11 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**, 9440-9445 (2003).

12 Zerbino, D., Wilder, S., Johnson, N., Juettemann, T. & Flicek, P. The Ensembl Regulatory Build. *Genome biology* **16**, 56 (2015).

13 Weir, B. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-898 (2007).

14 Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nature genetics* **48**, 500-509 (2016).

15 Wu, K. *et al.* Direct activation of TERT transcription by c-MYC. *Nature genetics* **21**, 220-224 (1999).

16 Creighton, C. Using large-scale molecular data sets to improve breast cancer treatment. *Breast Cancer Manag* **1**, 57-64 (2012).

17 Lewis, K. & Tollefsbol, T. Regulation of the Telomerase Reverse Transcriptase Subunit through Epigenetic Mechanisms. *Front Genet* **7**, 83 (2016).

18 Fredriksson, N., Ny, L., Nilsson, J. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics* **46**, 1258-1263 (2014).

19 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics* **46**, 1160-1165 (2014).

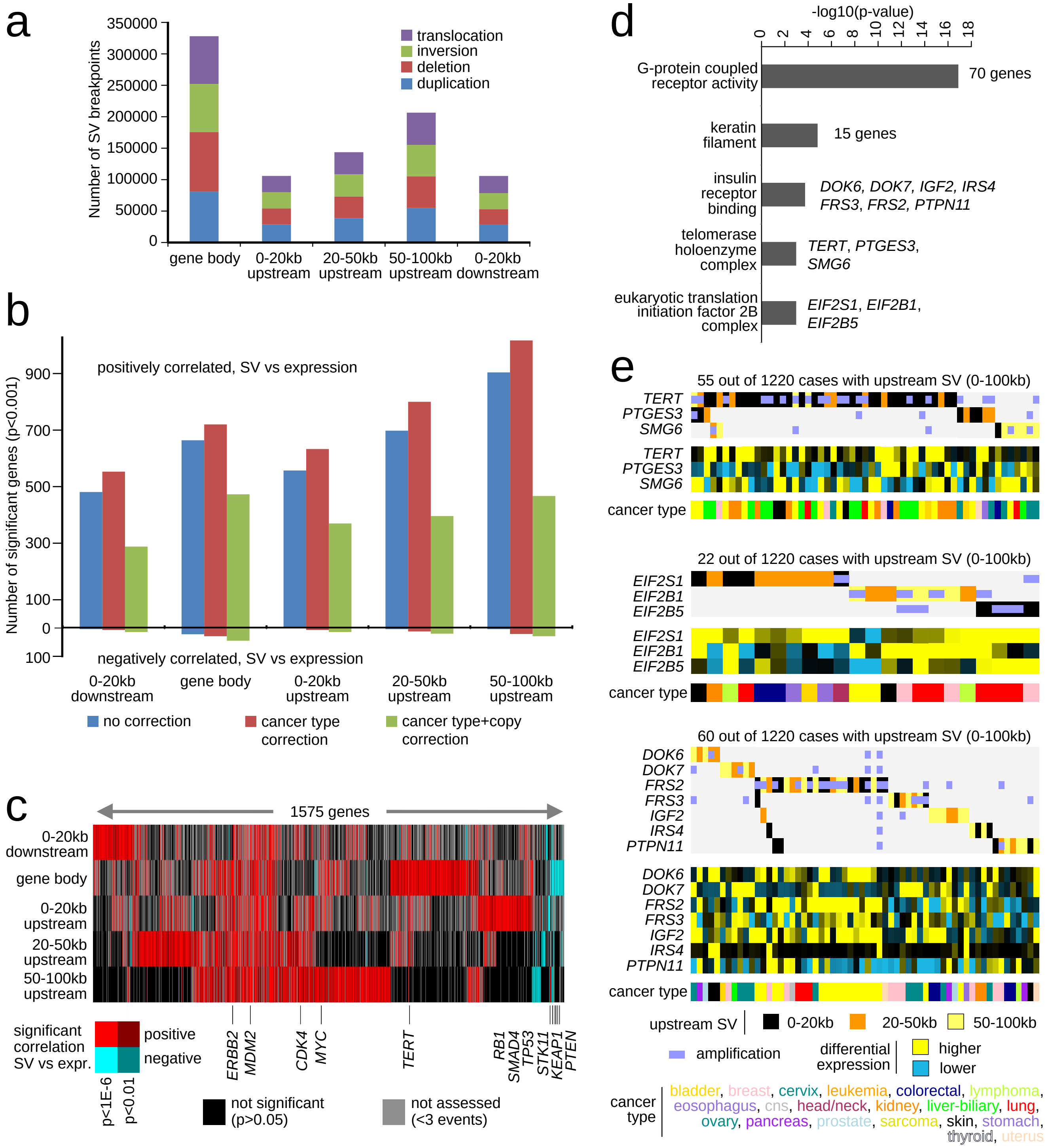
20 The_Cancer_Genome_Atlas_Research_Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49, doi:10.1038/nature12222 (2013).

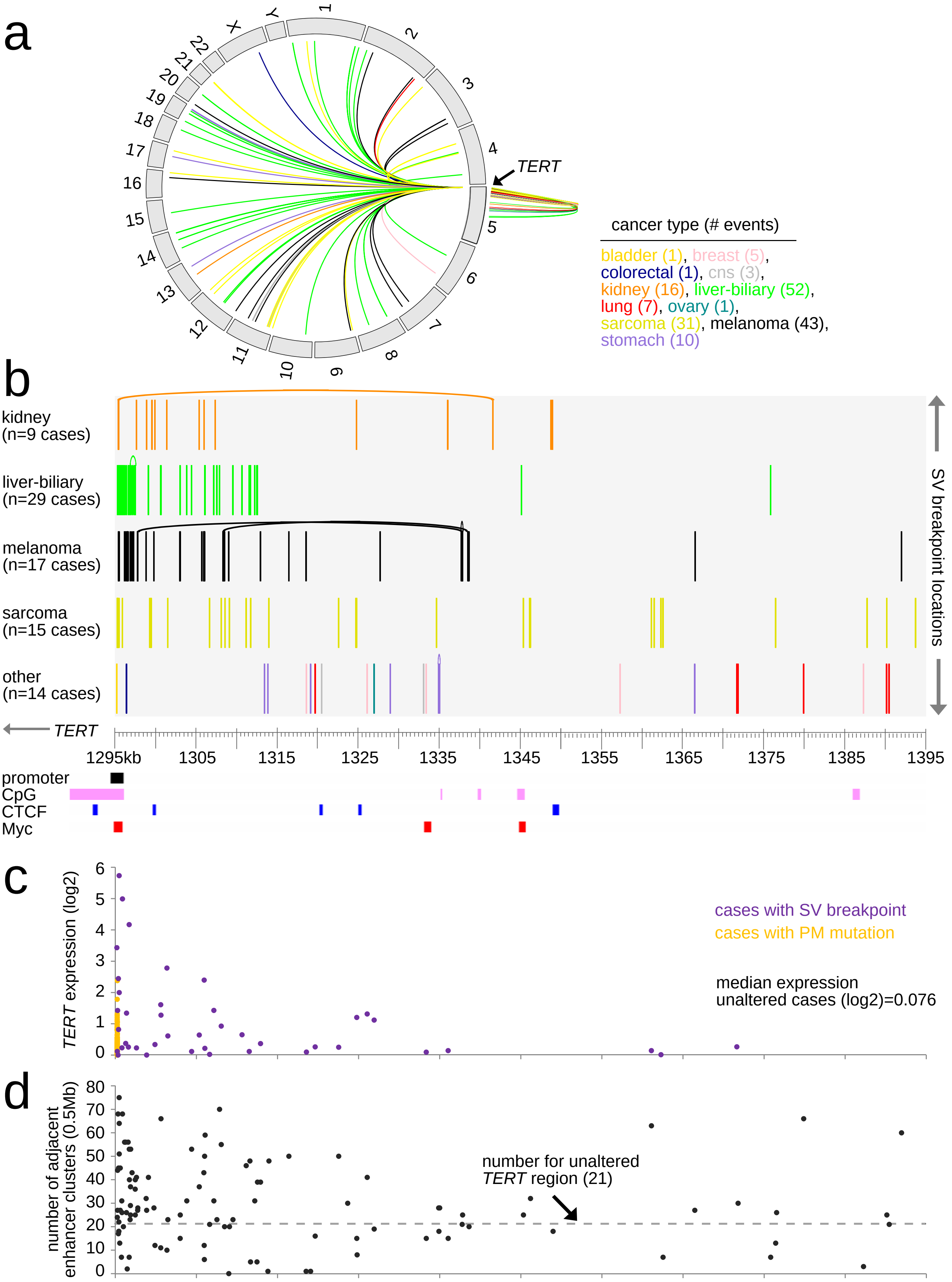
21 Johnson, W., Rabinovic, A. & Li, C. Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).

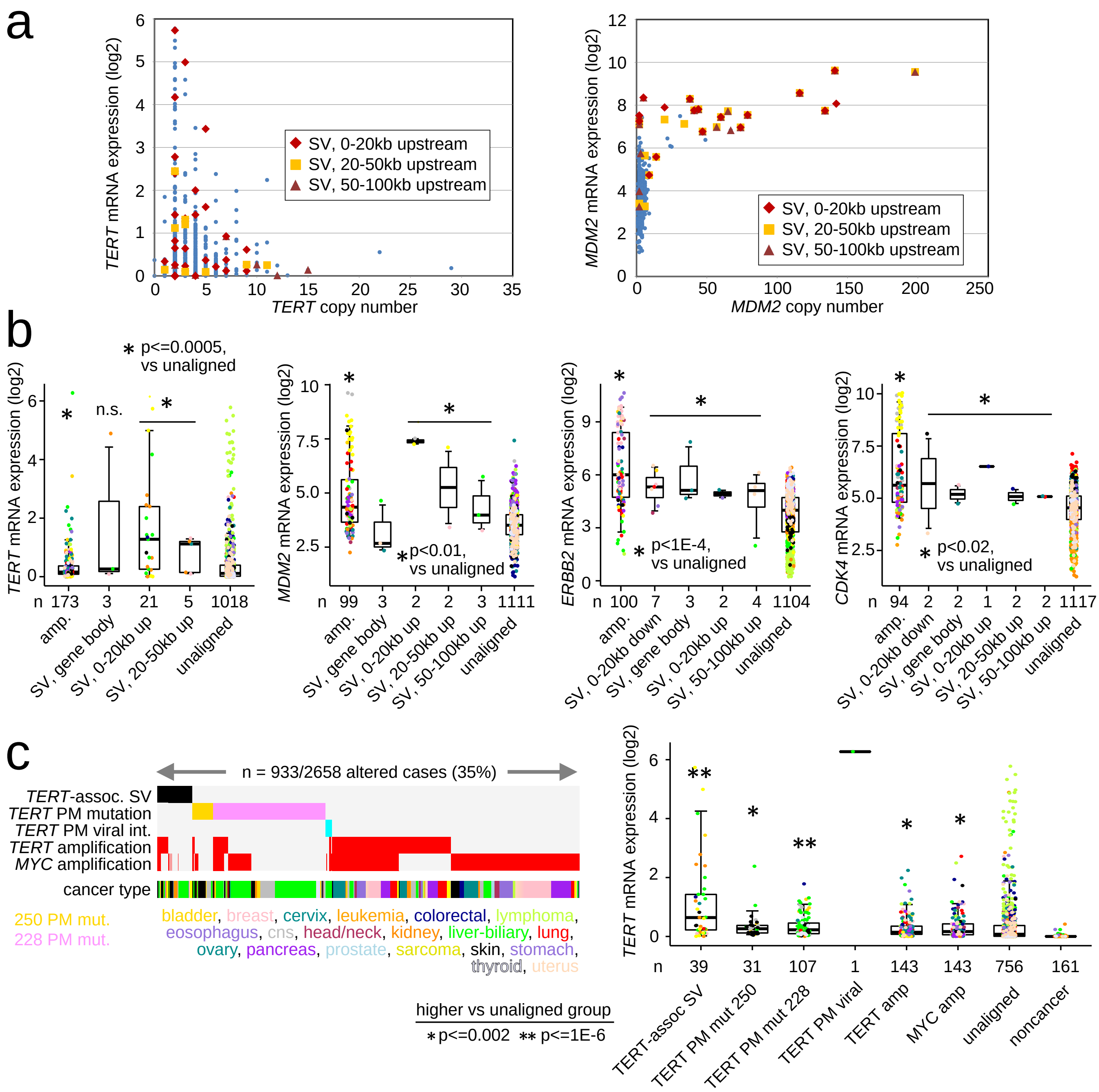
Table 1. Genes positively correlated in expression (p<0.001, corrected for copy number and cancer type) with occurrence of upstream SV, with the gene being previously associated with cancer.

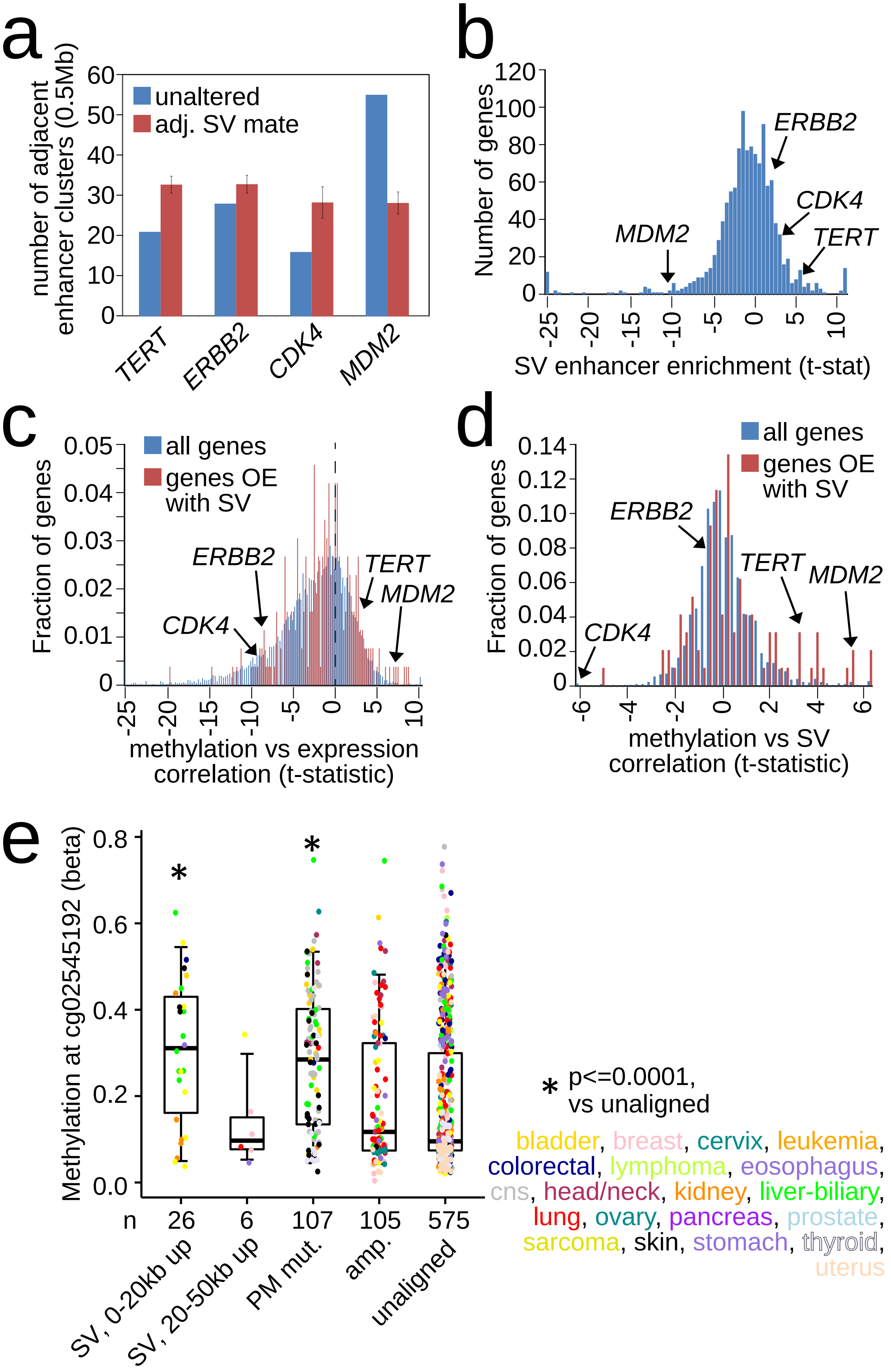
Region:	*0-20kb upstream*		*20-50kb upstream*		*50-100kb upstream*		*gene body*		*0-20kb downstream*	
Gene	*n*	*t*	*n*	*t*	*n*	*t*	*n*	*t*	*n*	*t*
CDK4	16	2.39	27	8.67	23	5.94	13	1.92	21	5.93
ERBB2	13	3.66	17	7.99	34	11.87	23	8.55	15	2
MDM2	17	9.5	22	7.9	21	9.52	20	8.84	19	8.35
TERT	31	8.08	9	2.34	8	0.73	10	7.39	5	6.81
CDK12	7	0.33	14	3.78	14	-0.02	41	2.8	11	3.01
HMGA2	10	3.71	8	4.31	15	1.71	24	2.16	6	-0.84
EGFR	8	1.69	12	4.39	9	2.41	31	5.57	6	4.01
TBL1XR1	3	0.38	9	3.51	9	1.11	32	2.23	4	2.02
MYCL	4	2.23	5	-0.14	10	4.24	0	NA	5	3.05
CCND3	3	2.97	6	4.01	7	4.18	15	4.53	5	1.44
CLTC	7	1.99	4	1.66	5	3.98	14	0.43	6	2.93
PDCD1LG2	3	3.8	8	4.02	4	0.97	9	7.81	6	5.33
PTPN11	4	2.83	3	3.88	7	2.59	7	1.1	3	-0.61
SMARCE1	2	NA	6	4.7	6	3.29	6	0.75	1	NA
PDGFRA	3	3.81	4	0.07	6	0.04	7	1.51	2	NA
NF1	1	NA	3	4.44	8	2.87	65	-2.98	0	NA
CD274	3	3.33	3	1.64	6	1.42	6	5.27	4	5.1
PRKAR1A	2	NA	3	1.3	3	3.39	4	2.29	1	NA
MYB	5	-0.18	3	3.58	0	NA	1	NA	1	NA
FOXL2	2	NA	3	5.27	3	-0.48	0	NA	2	NA
BCL7A	3	2.54	1	NA	3	3.38	7	1.76	3	2.86
SS18	0	NA	3	3.57	4	0.49	8	3.57	1	NA
TFE3	3	3.45	1	NA	2	NA	2	NA	0	NA
NKX2-1	1	NA	3	4.24	2	NA	0	NA	1	NA

Previous cancer association based on membership in the Sanger Cancer Consensus Gene list (http://www.sanger.ac.uk/science/data/cancer-gene-census). Number of cancer cases with SV in given region (n) is from the set of 1220 cases with both expression and SV data. t-statistic (t) based on linear regression model incorporating both cancer type and copy number in addition to SV event; a t-statistic of 3.3 or more approximates to p<0.001. Genes with p<0.001 for 0-20kb upstream, 20-50kb upstream, or 50-100kb upstream regions are included here. See also Table S2.


a
Number of SV breakpoints
350000
300000
250000
200000
150000
100000
50000
0
translocation
inversion
deletion
duplication
gene body
0-20kb upstream
20-50kb upstream
50-100kb upstream
0-20kb downstream
b
Number of significant genes (p<0.001)
900
700
500
300
100
0
100
positively correlated, SV vs expression
negatively correlated, SV vs expression
0-20kb downstream
gene body
0-20kb upstream
20-50kb upstream
50-100kb upstream
no correction
cancer type correction
cancer type+copy correction
c
1575 genes
0-20kb downstream
gene body
0-20kb upstream
20-50kb upstream
50-100kb upstream
significant correlation SV vs expr.
positive
negative
p<1E-6
p<0.01
ERBB2
MDM2
CDK4
MYC
TERT
RB1
SMAD4
TP53
STK11
KEAP1
PTEN
not significant (p>0.05)
not assessed (<3 events)
d
-log10(p-value)
0
2
4
6
8
10
12
14
16
18
G-protein coupled receptor activity
70 genes
keratin filament
15 genes
insulin receptor binding
DOK6, DOK7, IGF2, IRS4 FRS3, FRS2, PTPN11
telomerase holoenzyme complex
TERT, PTGES3, SMG6
eukaryotic translation initiation factor 2B complex
EIF2S1, EIF2B1, EIF2B5
e
55 out of 1220 cases with upstream SV (0-100kb)
TERT
PTGES3
SMG6
TERT
PTGES3
SMG6
cancer type
22 out of 1220 cases with upstream SV (0-100kb)
EIF2S1
EIF2B1
EIF2B5
EIF2S1
EIF2B1
EIF2B5
cancer type
60 out of 1220 cases with upstream SV (0-100kb)
DOK6
DOK7
FRS2
FRS3
IGF2
IRS4
PTPN11
DOK6
DOK7
FRS2
FRS3
IGF2
IRS4
PTPN11
cancer type
upstream SV
0-20kb
20-50kb
50-100kb
amplification
differential expression
higher
lower
cancer type
bladder, breast, cervix, leukemia, colorectal, lymphoma, eosophagus, cns, head/neck, kidney, liver-biliary, lung, ovary, pancreas, prostate, sarcoma, skin, stomach, thyroid, uterus


a
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
TERT
cancer type (# events)
bladder (1), breast (5),
colorectal (1), cns (3),
kidney (16), liver-biliary (52),
lung (7), ovary (1),
sarcoma (31), melanoma (43),
stomach (10)
b
kidney
(n=9 cases)
liver-biliary
(n=29 cases)
melanoma
(n=17 cases)
sarcoma
(n=15 cases)
other
(n=14 cases)
SV breakpoint locations
TERT
1295kb
1305
1315
1325
1335
1345
1355
1365
1375
1385
1395
promoter
CpG
CTCF
Myc
c
TERT expression (log2)
0
1
2
3
4
5
6
cases with SV breakpoint
cases with PM mutation
median expression
unaltered cases (log2)=0.076
d
number of adjacent
enhancer clusters (0.5Mb)
0
10
20
30
40
50
60
70
80
number for unaltered
TERT region (21)


a
TERT mRNA expression (log2)
TERT copy number
SV, 0-20kb upstream
SV, 20-50kb upstream
SV, 50-100kb upstream
MDM2 mRNA expression (log2)
MDM2 copy number
b
* p<=0.0005, vs unaligned
TERT mRNA expression (log2)
n.s.
n 173 3 21 5 1018
amp.
SV, gene body
SV, 0-20kb up
SV, 20-50kb up
unaligned
MDM2 mRNA expression (log2)
* p<0.01, vs unaligned
n 99 3 2 2 3 1111
SV, 50-100kb up
ERBB2 mRNA expression (log2)
* p<1E-4, vs unaligned
n 100 7 3 2 4 1104
SV, 0-20kb down
CDK4 mRNA expression (log2)
* p<0.02, vs unaligned
n 94 2 2 1 2 2 1117
c
n = 933/2658 altered cases (35%)
TERT-assoc. SV
TERT PM mutation
TERT PM viral int.
TERT amplification
MYC amplification
cancer type
250 PM mut.
228 PM mut.
bladder, breast, cervix, leukemia, colorectal, lymphoma, eosophagus, cns, head/neck, kidney, liver-biliary, lung, ovary, pancreas, prostate, sarcoma, skin, stomach, thyroid, uterus
TERT mRNA expression (log2)
n 39 31 107 1 143 143 756 161
TERT-assoc SV
TERT PM mut 250
TERT PM mut 228
TERT PM viral
TERT amp
MYC amp
unaligned
noncancer
higher vs unaligned group
* p<=0.002 ** p<=1E-6


a
number of adjacent enhancer clusters (0.5Mb)
60
50
40
30
20
10
0
unaltered
adj. SV mate
TERT
ERBB2
CDK4
MDM2
b
Number of genes
120
100
80
60
40
20
0
-25
-20
-15
-10
-5
0
5
10
SV enhancer enrichment (t-stat)
ERBB2
CDK4
TERT
MDM2
c
Fraction of genes
0.05
0.04
0.03
0.02
0.01
0
all genes
genes OE with SV
ERBB2
TERT
MDM2
CDK4
-25
-20
-15
-10
-5
0
5
10
methylation vs expression correlation (t-statistic)
d
Fraction of genes
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
all genes
genes OE with SV
ERBB2
TERT
MDM2
CDK4
-6
-4
-2
0
2
4
6
methylation vs SV correlation (t-statistic)
e
Methylation at cg02545192 (beta)
0.8
0.6
0.4
0.2
0.0
*
*
n
26
6
107
105
575
SV, 0-20kb up
SV, 20-50kb up
PM mut.
amp.
unaligned
* p<=0.0001, vs unaligned
bladder, breast, cervix, leukemia, colorectal, lymphoma, eosophagus, cns, head/neck, kidney, liver-biliary, lung, ovary, pancreas, prostate, sarcoma, skin, stomach, thyroid, uterus