nature COMMUNICATIONS

ARTICLE

　　OPEN

High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated *cis*-regulatory alterations

Yiqun Zhang[1], Fengju Chen[1], Nuno A. Fonseca [2], Yao He[3,4], Masashi Fujita [5], Hidewaki Nakagawa[5], Zemin Zhang[3,4], Alvis Brazma [2], PCAWG Transcriptome Working Group, PCAWG Structural Variation Working Group, Chad J. Creighton [1,6,7,8]* & PCAWG Consortium

The impact of somatic structural variants (SVs) on gene expression in cancer is largely unknown. Here, as part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, which aggregated whole-genome sequencing data and RNA sequencing from a common set of 1220 cancer cases, we report hundreds of genes for which the presence within 100 kb of an SV breakpoint associates with altered expression. For the majority of these genes, expression increases rather than decreases with corresponding breakpoint events. Up-regulated cancer-associated genes impacted by this phenomenon include *TERT*, *MDM2*, *CDK4*, *ERBB2*, *CD274*, *PDCD1LG2*, and *IGF2*. *TERT*-associated breakpoints involve ~3% of cases, most frequently in liver biliary, melanoma, sarcoma, stomach, and kidney cancers. SVs associated with up-regulation of PD1 and PDL1 genes involve ~1% of non-amplified cases. For many genes, SVs are significantly associated with increased numbers or greater proximity of enhancer regulatory elements near the gene. DNA methylation near the promoter is often increased with nearby SV breakpoint, which may involve inactivation of repressor elements.

[1] Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA. [2] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK. [3] BIOPIC and College of Life Sciences, Peking University, Beijing 100871, China. [4] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China. [5] Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo 108-8639, Japan. [6] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [7] Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA. [8] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. PCAWG Transcriptome Working Group and PCAWG Structural Variation Working Group authors and their affiliations appear at the end of the paper. PCAWG Consortium members and their affiliations appear in the Supplementary Information. *email: creighto@bcm.edu

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* upregulation[1–3]. Genomic rearrangements in cancer are common and often associated with copy number alterations[4,5]. Breakpoints associated with rearrangement can potentially alter the regulation of nearby genes, e.g., by disrupting specific regulatory elements or by translocating *cis*-regulatory elements from elsewhere in the genome into close proximity to the gene. Recent examples of rearrangements leading to "enhancer hijacking"—whereby enhancers from elsewhere in the genome are juxtaposed near genes, leading to overexpression—include a distal *GATA2* enhancer being rearranged to ectopically activate *EVI1* in leukemia[6], activation of GFI1 family oncogenes in medulloblastoma[7], and 5p15.33 rearrangements in neuroblastoma juxtaposing strong enhancer elements to *TERT*[8]. By integrating somatic copy alterations, gene expression data, and information on topologically associating domains (TADs), a recent pan-cancer study uncovered 18 genes with overexpression resulting from rearrangements of *cis*-regulatory elements (including enhancer hijacking)[9]. Genomic rearrangement may also disrupt the boundary sites of insulated chromosome neighborhoods, resulting in gene upregulation[10].

The PCAWG Consortium aggregated whole-genome sequencing data from 2658 cancers across 38 tumor types generated by the ICGC and TCGA projects. These sequencing data were reanalyzed with standardized, high-accuracy pipelines to align to the human genome (reference build hs37d5) and identify germline variants and somatically acquired mutations[11]. These data involve a comprehensive and unified identification of somatic substitutions, indels, and structural variants (SVs, representing genomic rearrangement events, each event involving two breakpoints from different genomic coordinates becoming fused together), based on "consensus" calling across three independent algorithmic pipelines, together with initial basic filtering, quality checks, and merging[11]. Whole-genome sequencing offers much better resolution in SV inference over that of whole exome data or SNP arrays[4,9]. These data represent an opportunity for us to survey this large cohort of cancers for somatic SVs with breakpoints located in proximity to genes. For a sizeable subset of cases in the PCAWG cohort, data from other platforms in addition to whole-genome sequencing, such as RNA expression or DNA methylation, are available for integrative analyses, with 1220 cases having both whole-genome and RNA sequencing.

While SVs can result in two distant genes being brought together to form fusion gene rearrangements (e.g., *BCR-ABL1* or *TMPRSS2-ERG*)[12], this present study focuses on SVs impacting gene regulation in the absence of fusion events or copy number alterations, e.g., SVs with breakpoints occurring upstream or downstream of the gene and involving rearrangement of *cis*-regulatory elements. In a recent study involving integration of gene expression with low-pass whole-genome sequencing for more than 1000 cancer cases[13], evidence for a widespread impact of somatic SVs on gene expression patterns was observed, though a noted limitation with that study involved the level of coverage (~6–8×) of low-pass sequencing. With a genome-wide analysis involving a large sample size and much deeper sequencing coverage (~30–60×), information from multiple genes may be leveraged more effectively, in order to identify common features involving the observed disrupted regulation of genes impacted by somatic genomic rearrangement.

In this present study, we utilize the PCAWG datasets in order to analyze high coverage whole-genome sequencing data from 1220 individuals. Integrating SV calls with gene expression data,

we observe a widespread impact of somatic structural variants on gene expression patterns, independent of copy number alterations, involving key oncogenes and tumor suppressor genes. Mechanisms involved with SV-mediated gene deregulation, as observed here, include enhancer hijacking and altered DNA methylation.

Results

Widespread impact of somatic SVs on gene expression. Inspired by recent observations in kidney cancer[3,14], neuroblastoma[8,15], and B-cell malignancies[16], of recurrent genomic rearrangements affecting the chromosomal region proximal to *TERT* and resulting in its upregulation, we sought to carry out a pan-cancer analysis of all coding genes, for ones appearing similarly affected by somatic rearrangement. We referred to a dataset of somatic SVs called for high coverage 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 (Supplementary Data 1). 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 20–50 kb upstream of a gene, 50–100 kb upstream of a gene, within a gene body, or 0–20 kb downstream of a gene (Fig. 1a). (SV breakpoints 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 group would have a distinct molecular signature[17], and as genomic rearrangements may be involved in copy alterations[4,13], 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 breakpoint 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 (Fig. 1b and Supplementary Fig. 1a and Supplementary Data 2). For gene body, 0–20 kb upstream, 20–50 kb upstream, 50–100 kb upstream, and 0–20 kb downstream regions, the numbers of significant genes at $p < 0.001$ (corresponding to estimated false discovery rates[18] of <4%, Supplementary Data 2) 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 0–20 kb 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 with 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 (Fig. 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 (Fig. 1c). Tumor purity, tumor ploidy, and total number of SV breakpoints were not found to represent significant confounders (Supplementary Fig. 1b). High numbers of statistically significant genes were also found when
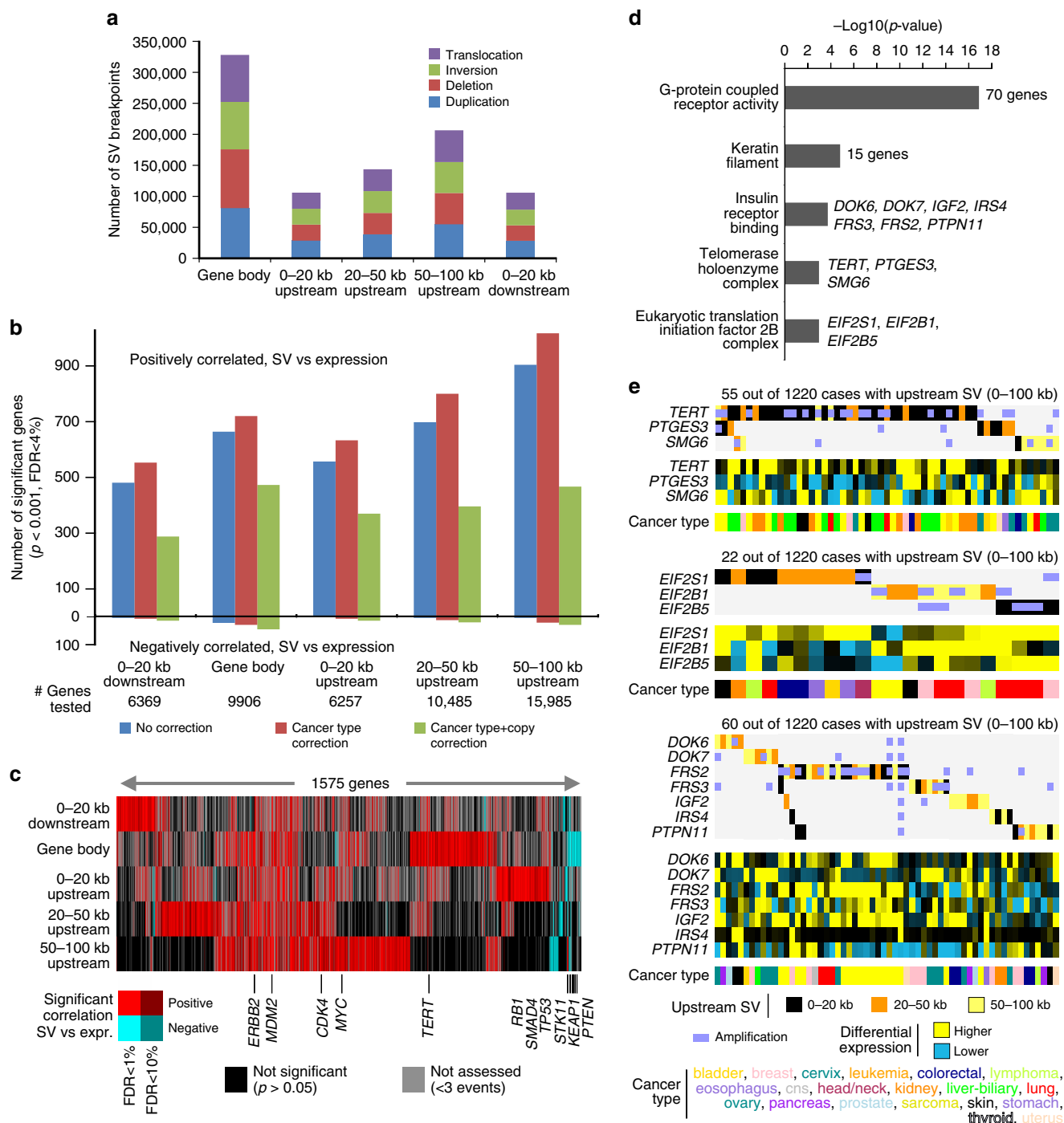
Fig. 1 Structural Variant (SV) breakpoints associated with altered expression of nearby genes. a Numbers of SV breakpoints identified as occurring within a gene body, 0–20 kb upstream of a gene, 20–50 kb upstream of a gene, 50–100 kb upstream of a gene, or 0–20 kb downstream of a gene. For each SV set, the breakdown by alteration class is indicated. SVs with breakpoints 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 ($p < 0.001$, FDR < 4%), showing correlation between expression and associated SV event. 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 (<3 SV events for given gene in the given genomic region). **d** Significantly enriched Gene Ontology (GO) terms for genes positively correlated ($p < 0.001$ and FDR < 4%) with occurrence of SV upstream of the gene (for either 0–20 kb, 20–50 kb, or 50–100 kb 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 Supplementary Data 1, 2 and Supplementary Fig. 1.

Table 1 Selected genes positively correlated in expression with occurrence of upstream SV breakpoint.

Region:	0–20 kb upstream		20–50 kb upstream		50–100 kb upstream		Gene body		0–20 kb 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

Table lists the genes positively correlated in expression ($p < 0.001$ and FDR < 4%, corrected for copy number and cancer type) with occurrence of upstream SV breakpoint, with the gene being previously associated with cancer. 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 $p < 0.001$ or FDR < 4%. Genes with $p < 0.001$ for 0–20 kb upstream, 20–50 kb upstream, or 50–100 kb upstream regions are included here. "NA", not assessed (less than three cases involved). See also Supplementary Data 2

examining regions further upstream or downstream of genes, up to 1 Mb (Supplementary Fig. 1c).

Key driver genes in cancer impacted by nearby SV breakpoints. Genes with increased expression associated with nearby SV breakpoints included many genes with important roles in cancer (Table 1), such as *TERT* (significant with $p < 0.001$ for regions from 0–20 kb downstream to 20–50 kb upstream of the gene), *MYC* (significant for gene body SV breakpoints), *MDM2* (regions from 0–20 kb downstream to 50–100 kb upstream), *CDK4* (0–20 kb downstream and 20–100 kb upstream), *ERBB2* (gene body to 50–100 kb upstream), *CD274* (0–20 kb downstream to 50–100 kb upstream), *PDCD1LG2* (0–20 kb downstream to 20–50 kb upstream), and *IGF2* (0–20 kb downstream and 50–100 kb upstream). Genes with decreased expression associated with SV breakpoints located within the gene included *PTEN*[19] ($n = 50$ cases with an SV breakpoint out of 1220 cases with expression data available), *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; for other genes, SV breakpoints within the gene could potentially impact intronic regulatory elements, or could represent potential fusion events (though in a small fraction of cases)[12,13]. Examining the set of genes positively correlated ($p < 0.001$, FDR < 4%) with occurrence of SV breakpoint upstream of the gene (for either 0–20 kb, 20–50 kb, or 50–100 kb SV sets), enriched gene categories (Fig. 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 (Fig. 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 (Fig. 1e).

Translocations involving the region 0–100 kb upstream of *TERT* were both inter- and intrachromosomal (Fig. 2a and Supplementary Data 3) 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 20 kb of the *TERT* start site (Fig. 2b), which represented the region where correlation between SV events and *TERT* expression was strongest (Fig. 2c, d, $p < 1E-14$, linear regression model). In neuroblastoma, translocation of enhancer regulatory elements near the promoter was previously associated with *TERT* upregulation[8,15]. Here, in a global analysis, we examined the number of enhancer elements[20] within a 0.5 Mb region upstream of each rearrangement breakpoint occurring in proximity to *TERT* (for breakpoints where the breakpoint 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 (Fig. 2e), representing a significant increase ($p < 1E-6$, paired t-test). A trend was also observed, by which SV breakpoints closer to the *TERT* start site were associated with a larger number of enhancer elements (Fig. 2d, $p < 0.03$, Spearman's correlation).

Consistent with observations elsewhere[4,13], genomic rearrangements could be associated here with copy alterations for a large number of genes (Fig. 1b), including genes of particular interest
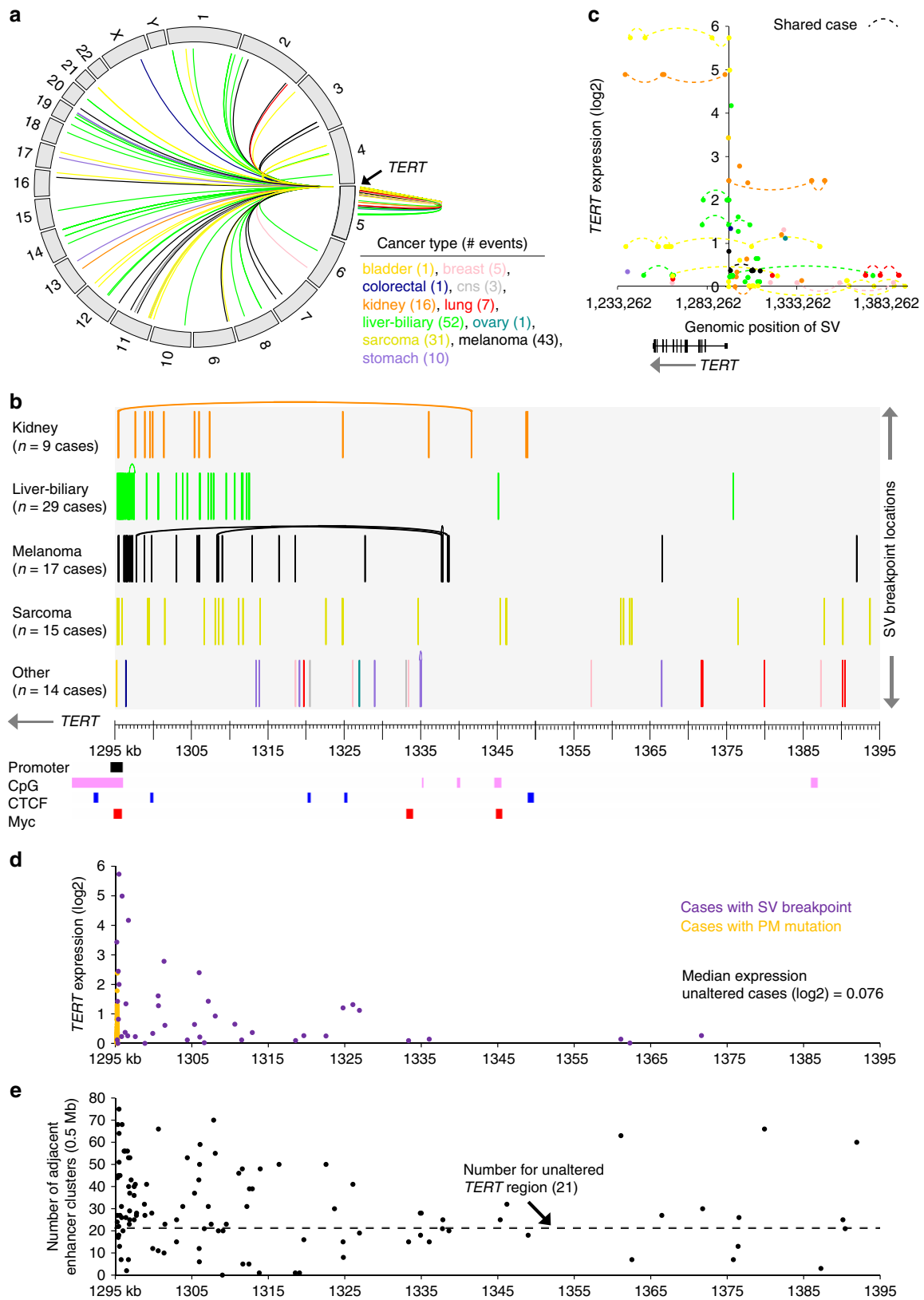
such as *TERT* and *MDM2* (Fig. 3a). However, copy alteration alone would not account for all observed cases of increased expression in conjunction with SV event. For example, with a number of key genes (including *TERT*, *MDM2*, *ERBB2*, *CDK4*), when all amplified cases (i.e., with five or more gene copies) were grouped into a single category, regardless of SV breakpoint occurrence, the remaining SV-involved cases showed significantly increased expression (Fig. 3b). Regarding *TERT* in particular, a number of types of genomic alteration may act upon transcription, including upstream SV breakpoint, *TERT* amplification[21], promoter mutations[1,2], promoter viral integration[22], and *MYC* amplification[23]. Within the PCAWG cohort of 2658 cancer cases,

Fig. 2 SVs associated with *TERT* and its increased expression. a Circos plot showing all intra- and interchromosomal rearrangements 0–100 kb from the TERT locus. **b** By cancer type, SV breakpoint locations within the region ~100 kb 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** Gene expression levels of *TERT* corresponding to SVs with breakpoints located in the genomic region 0–20 kb downstream to 100 kb upstream of the gene (116 SV breakpoints involving 47 cases). **d** Where data available, gene expression levels of *TERT* corresponding to SVs from part (**b**). Expression levels associated with *TERT* promoter (PM) mutation are also represented. Median expression for unaltered cases represents cases without *TERT* alteration (SV, promoter mutation, amplification, viral integration) or *MYC* amplification. For part (**d**), 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. **e** 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 were 0.5 Mb upstream of the gene. See also Supplementary Data 3.

933 (35%) were altered according to at least one of the above alteration classes, with each class being associated with increased *TERT* mRNA expression (Fig. 3c). Upstream SV breakpoints in particular were associated with higher *TERT* as compared with promoter mutation or amplification events.

SVs associated with *CD274* (PD1) and *PDCD1LG2* (PDL1)—genes with important roles in the immune checkpoint pathway—were associated with increased expression of these genes (Fig. 4a and Supplementary Data 4). Out of the 1220 cases with gene expression data, 19 harbored an SV breakpoint in the region involving the two genes, both of which reside on chromosome 9 in proximity to each other (Fig. 4b, considering the region 50 kb upstream of *CD274* to 20 kb downstream of *PDCD1LG2*). These 19 cases included lymphoma (*n* = 5), lung (4), breast (2), head and neck (2), stomach (2), colorectal (1), and sarcoma (1). Six of the 19 cases had amplification of one or both genes, though on average cases with associated SV had higher expression than cases with amplification but no SV breakpoint (Fig. 4a, *p* < 0.0001 *t*-test on log-transformed data). For most of the 19 cases, the SV breakpoint was located within the boundaries of one of the genes (Fig. 4a), while both genes tended to be elevated together regardless of the SV breakpoint position (Fig. 4b). We examined the 19 cases with associated SVs for fusions involving either *CD274* or *PDCD1LG2*, and we identified a putative fusion transcript for *RNF38->PDCD1LG2* involving three cases, all of which were lymphoma. No fusions were identified involving *CD274*.

Translocated enhancers and altered DNA methylation. Similar to analyses focusing on *TERT* (Fig. 2d), we examined SVs involving other genes for potential translocation of enhancer elements. For example, like *TERT*, SVs with breakpoints 0–20 kb upstream of *CDK4* were associated with an increased number of upstream enhancer elements as compared with that of the unaltered gene (Fig. 5a); however, SV breakpoints upstream of *MDM2* were associated with significantly fewer enhancer elements compared with that of the unaltered region (Fig. 5a). For the set of 1233 genes with at least 7 SV breakpoints 0–20 kb 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 (Fig. 5b and Supplementary Data 5). Of these genes, 24% showed differences at a significance level of *p* < 0.01 (paired *t*-test, with ~12 nominally significant genes being expected by chance, FDR = 4%). However, for most of these genes, the numbers of enhancer elements was decreased on average with the SV breakpoint rather than increased (195 versus 103 genes, respectively), indicating that translocation of greater numbers of enhancers might help explain the observed upregulation for some but not all genes. For other genes (e.g., *HOXA13* and *CCNE1*), enhancer elements on average were positioned in closer proximity to the gene as a result of the genomic rearrangement (Fig. 5c). Of 829 genes examined (with at least 5 SV

breakpoints 0–20 kb upstream and with breakpoint mate on the distal side from the gene, where the breakpoint occurs between the gene start site and its nearest enhancer in the unaltered scenario), 8.3% showed a significant decrease (*p* < 0.01, paired *t*-test, FDR = 10.8%) in distance to the closest enhancer on average as a result of the SV breakpoint, as compared with 1% showing a significance increase in distance.

We went on to examine genes impacted by nearby SV breakpoints 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[24] of inverse correlations between DNA methylation and gene expression were observed (Fig. 6a and Supplementary Fig. 2 and Supplementary Data 6). However, for the subset of 263 genes positively correlated in expression with occurrence of upstream SV breakpoint (*p* < 0.001 and FDR < 4%, 0–20 kb, 20–50 kb, or 50–100 kb SV sets), the methylation-expression correlations were less skewed toward negative (*p* = 0.0001 by *t*-test, comparing the two sets of correlation distributions in Fig. 5a). 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 (Fig. 6a). Regarding *TERT*, a CpG site located in close proximity to its core promotor is known to contain a repressive element[8,25]; non-methylation results in the opening of CTCF binding sites and the transcriptional repression of *TERT*[25]. In the PCAWG cohort, SV breakpoints occurring 0–20 kb upstream of the gene were associated with increased CpG island methylation (Fig. 6b), while SV breakpoints 20–50 kb upstream were not; *TERT* promoter mutation was also associated with increased methylation (Fig. 6c).

Discussion
Using a unique dataset of high coverage 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 genomic rearrangement of regions nearby genes, leading to gene upregulation—a phenomenon previously observed for individual genes such as *TERT*—globally impacts a large proportion of genes and of cancer cases. Genomic rearrangements involved with upregulation of *TERT* in particular have furthermore been shown here to involve a wide range of cancer types, expanded from previous observations made in individual cancer types such as kidney chromophobe and neuroblastoma. While many of the genes impacted by genomic rearrangement in this present study likely represent passengers rather than drivers of the disease, many other genes with canonically established roles in cancer would be impacted. Outside information can be brought to bear in distinguishing driver from passenger genes, including significant mutation or copy number alteration patterns[26,27], experimental data, and domain-specific expertise. Though any given gene may not be impacted in a large percentage of cancer
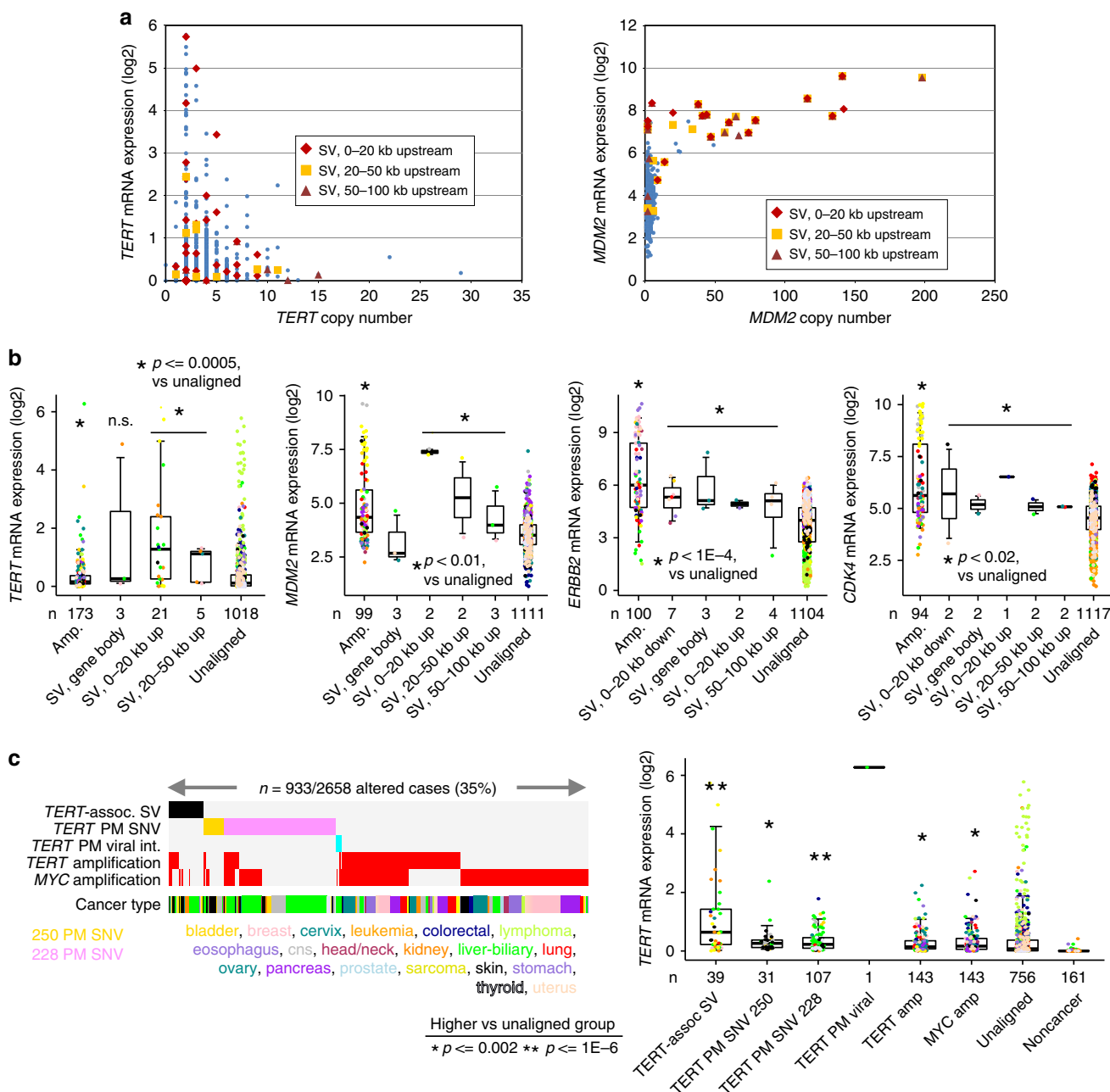
Fig. 3 SV breakpoints in proximity to key genes uniquely contribute to cases of high expression. a For 1220 cancer cases, copy number versus expression for *TERT* (left) and *MDM2* (right). Cases with SV events upstream of the gene are indicated. **b** Box plots of expression for *TERT*, *MDM2*, *ERBB2*, and *CDK4* by alteration class ("amp." or gene amplification: 5 or more copies, SV breakpoint within gene body, SV breakpoint 0–20 kb downstream of gene, SV breakpoint 0–20 kb upstream of gene, SV breakpoint 20–50 kb upstream of gene, SV breakpoint 50–100 kb upstream of gene, or none of the above, i.e., "unaligned"). Cases with both SV breakpoint and amplification are assigned here within the amplification group. Asterisks ("*") denote statistically significant differences versus unaligned group as indicated. **c** Left: Alterations involving *TERT* (SV breakpoint 0–50 kb 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 whole-genome sequencing 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, Single Nucleotide Variant or "SNV", viral). P-values by Mann–Whitney U-test; "*" denotes significant differences versus unaligned group with $p <= 0.002$, and "**" denotes significant differences with $p < 1E-6$. 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**).

cases (the more frequently SV-altered gene *TERT* involving <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 altered *cis*-regulation should therefore be regarded as an important driver mechanism in cancer, alongside that of somatic point mutations, copy number alteration, epigenetic silencing, gene fusions, and

germline polymorphisms. Our results have implications for personalized or precision medicine, which tends to be primarily focused on mutations within coding regions.

While the role of genomic rearrangements in altering the *cis*-regulation of specific genes within specific cancer types has been previously observed, our present pan-cancer study demonstrates that this phenomenon is more extensive and impacts a far greater
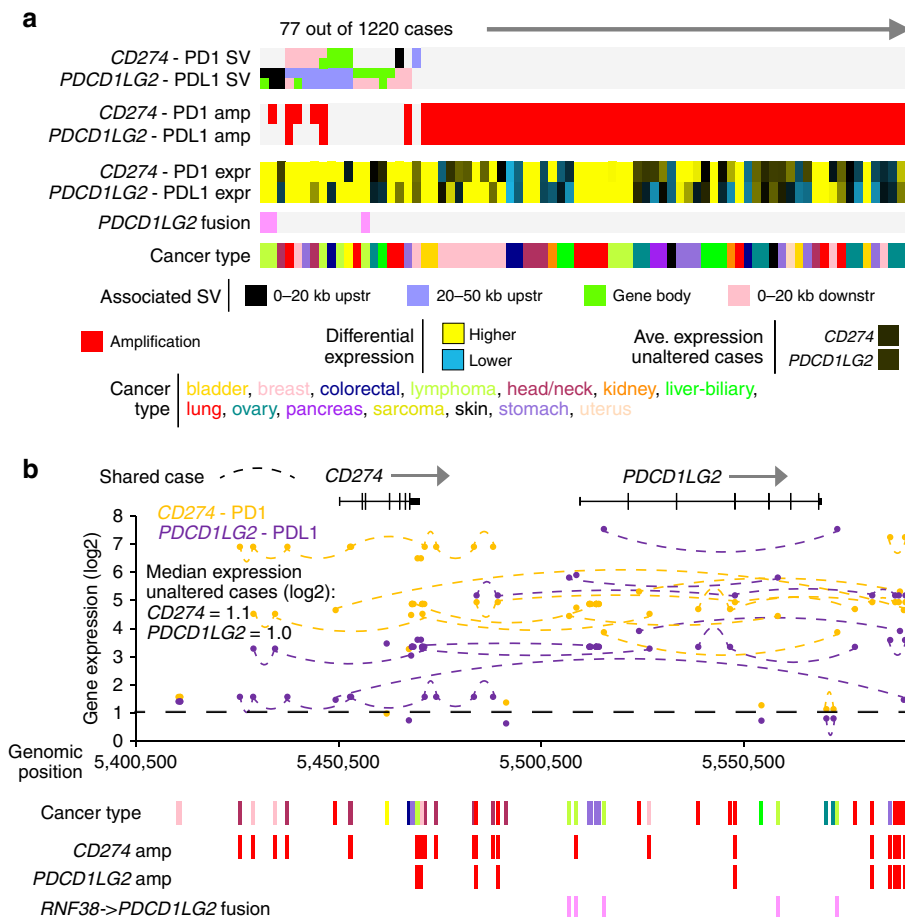
Fig. 4 SVs associated with PD1/PDL1 genes and their increased expression. a Patterns of SV, gene amplification (5 or more copies), *RNF38->PDCD1LG2* gene fusion, and differential expression for *CD274* (PD1 gene) and *PDCD1LG2* (PDL1 gene), for the subset of cases with associated SV or amplification for either gene. Differential gene expression patterns relative to the median across sample profiles. **b** Gene expression levels of *CD274* and of *PDCD1LG2*, corresponding to the position of SV breakpoints located in the surrounding genomic region on chromosome 9 (representing 66 SV breakpoints involving 19 cases). Median expression for unaltered cases represents cases without SV or amplification. See also Supplementary Data 4.
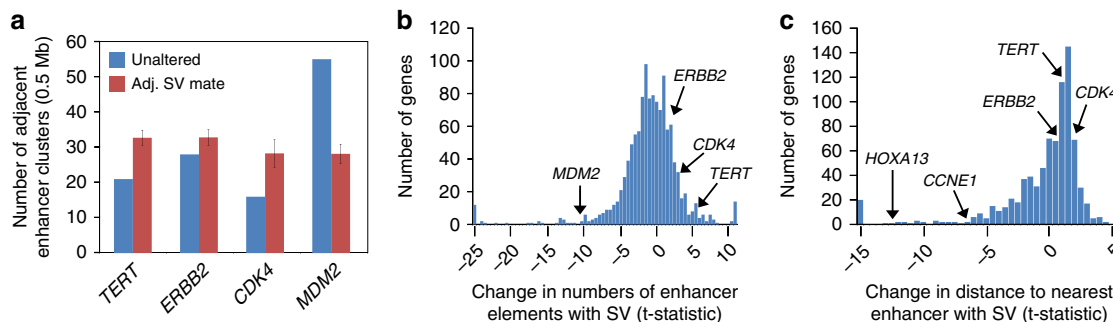

Fig. 5 Translocation of enhancer elements associated with SV breakpoints near genes. a For *TERT*, *ERBB2*, *CDK4*, and *MDM2*, average number of enhancer elements within a 0.5 Mb region upstream of each rearrangement breakpoint (considering the respective SV sets occurring 0–20 kb upstream of each gene), as compared with 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 SV breakpoints 0–20 kb 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** For 829 genes (with at least 5 SV breakpoints 0–20 kb upstream and with breakpoint mate on the distal side from the gene, where the breakpoint occurs between the gene start site and its nearest enhancer in the unaltered scenario), histogram of t-statistics (paired t-test) comparing the distance of the closest enhancer element upstream of rearrangement breakpoints with the distance for the unaltered gene. Negative t-statistics denote a shorter distance associated with the SV breakpoints. See also Supplementary Data 5.
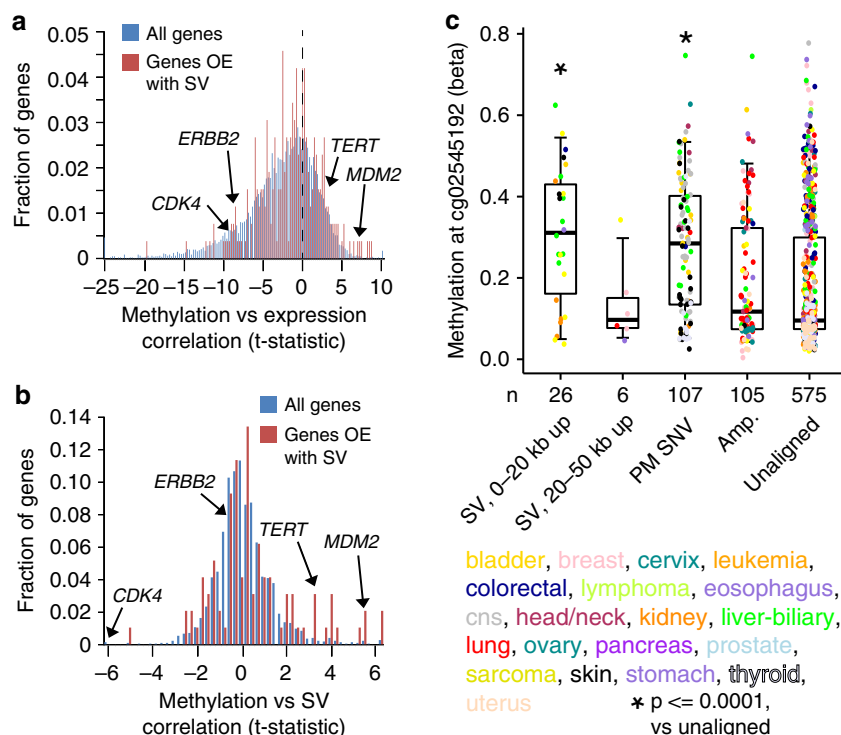
Fig. 6 Altered DNA methylation patterns associated with SV breakpoints near genes. a Histogram of *t*-statistics for correlation between gene expression and DNA methylation (by Pearson's using log-transformed expression and logit-transformed methylation), for both the entire set of 8256 genes (blue) 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 and FDR < 4%, "OE" for "overexpressed") with occurrence of upstream SV breakpoint (for either 0–20 kb, 20–50 kb, or 50–100 kb SV sets). **b** 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 three cases with SV breakpoints 0–20 kb upstream and represented on methylation platform and the subset of 97 genes (red) on methylation platform and positively correlated in expression (*p* < 0.001 and FDR < 4%) with occurrence of SV breakpoint 0–20 kb upstream. **c** DNA methylation of the CpG site cg02545192 proximal to the TERT core promoter in cases with SV breakpoint 0–20 kb or 20–50 kb upstream of TERT, in cases with TERT promoter (PM) activation mutation (SNV), in cases with TERT amplification ("amp."), and in the rest of cases (unaligned). *P*-values by *t*-test on logit-transformed methylation beta values; "*" denotes significant differences versus unaligned group with *p* <= 0.0001. Box plots represent 5, 25, 50, 75, and 95%. Points in box plots are colored according to tumor type as indicated. See also Supplementary Data 6 and Supplementary Fig. 2.

number of genes than may have been previously thought. A recent study by Weischenfeldt et al.[9], utilizing SNP arrays to estimate SV breakpoints occurring within TADs (which confine physical and regulatory interactions between enhancers and their target promoters), uncovered 18 genes (including *TERT* and *IRS4*) in pan-cancer analyses and 98 genes (including *IGF2*) in cancer type-specific analyses with overexpression associated with rearrangements involving nearby or surrounding TADs. Our present study using PCAWG datasets identifies hundreds of genes impacted by SV-altered regulation, far more than the Weischenfeldt study. In contrast to the Weischenfeldt study, our study could take advantage of high coverage whole-genome sequencing over SNP arrays, with the former allowing for much better resolution in identifying SVs, including those not associated with copy alterations. In addition, while TADs represent very large genomic regions, often extending over 1 Mb, our study pinpoints SV with breakpoints acting within relatively close distance to the gene, e.g., within 20 kb for many genes. In principle, genomic rearrangements could impact a number of regulatory mechanisms, not necessarily limited to enhancer hijacking or TAD disruption, and genes may be altered differently in different samples. The analytical approach of our present study has the advantage of being able to identify robust associations between SVs and expression, without making assumptions as to the specific mechanism.

Future efforts can further explore the mechanisms involved with specific genes deregulated by nearby genomic rearrangements. Regarding *TERT*-associated SVs, for example, previously observed increases in DNA methylation of the affected region had been previously thought to be the result of massive chromatin remodeling brought about by juxtaposition of the *TERT* locus to strong enhancer elements[8], which is supported by observations made in this present study involving multiple cancer types. 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 SV breakpoints 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.

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[11]. The PCAWG workflows are also available as Docker images through Dockstore enabling researchers to replicate the steps involved in the data assembly[11]. In all, 2658 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, central nervous system or "cns", head/neck, kidney, lung, skin, ovary, pancreas, thyroid, uterus). Cancer molecular profiling data were generated through informed consent as part of previously published studies and analyzed in accordance with each original study's data use guidelines and restrictions. Of the 2658 donors (Supplementary Data 1) included among the whitelist (acceptable for all analyses) and graylist (excluded from some analyses carried out as part of PCAWG-led efforts), 1220 had RNA data, 32 of which were graylisted. In accordance with the PCAWG consortium policy, we included the graylisted cases in our analysis, as these were found to have no issues pertaining to our integration analysis approaches involving RNA and SV data.

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, along with additional filtering criteria being used as described by the PCAWG consortium[11]. Somatic SVs were defined by comparison between the tumor and matched normal. The consensus SV calls are available from synapse (https://www.synapse.org/#!Synapse:syn7596712).

For copy number, the calls made by the Sanger group using the Ascat NGS algorithm[11] with default parameters were used, which data are available at the ICGC Data Portal (https://dcc.icgc.org/pcawg). Gene copies of five or more were called as amplification events. For somatic mutation of *TERT* promoter, PCAWG variant calls, as well as any additional data available from the previous individual studies[3,11,22], were used (Supplementary Data 1). *TERT* promoter viral integrations were obtained from ref. [22]. Of the 2658 cases, RNA-seq data were available for 1220 cases. For RNA-seq data, alignments by both STAR (version 2.4.0i,2-pass) and TopHat2 (version 2.0.12) were used to generated a combined set of expression calls[12]; alignment parameters and other methodology details are provided at ref. [12]. FPKM-UQ values (where UQ = upper quartile of fragment count to protein coding genes) were used (dataset available at https://www.synapse.org/#!Synapse:syn5553991). Where a patient had multiple tumor sample profiles (this scenario involving a handful of patients), one profile was randomly selected to represent the patient. Overall, RNA-seq samples derived from a similar tissue of origin had similar expression profiles; more specifically, tumor samples from donors derived from different projects were similar and also tissue derived from GTEx versus matched normal tissue were similar, indicating that technical batch effects did not represent a major confounder[12].

In a concerted effort to reduce batch effects due to the use of different computational pipelines in the initial studies, the PCAWG consortium systematically reanalyzed all of the RNA-seq libraries from the individual projects using a unified RNA-seq analysis pipeline, as detailed in ref. [12]. However, it is conceivable that there may be batch effects (e.g., from the isolation and handling of the RNA, library construction and sequencing factors etc.) that have not been possible to take into account. At the same time, where this present study involves integration between orthogonal data platforms, such data integration should be less susceptible to batch effects, as any source of technical variation in one data platform would be less likely to be manifested in the other platform. In addition, our linear models relating SV breakpoint patterns with gene expression (described below) incorporated cancer type as a covariate, and so any genes selected as having significant correlations between SV breakpoints and expression must arise above any associations would be best explained on the basis of cancer type alone. For example, genes that are generally high or low across specific tumor types (whether by biology or by batch effect), irrespective of SV breakpoint pattern, would not be selected as significant.

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[12]. To help correct for batch effects between methylation data platforms (HM450 versus HM27), we used the combat software[12] with R software version 3.0 (with 27K vs 450K as the "batch" and cancer type as the "experimental group", R code available at https://www.bu.edu/jlab/wp-assets/ComBat/Download_files/ComBat.R), as we have done in previous pan-cancer studies utilizing The Cancer Genome Atlas methylation datasets[14,28–30]. 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. Correlations between DNA methylation and gene expression were assessed using logit-transformed methylation data and log-transformed expression data and Pearson's correlations.

Integrative analyses between SVs and gene expression. For each of a number of specified genomic region windows in relation to genes, we constructed a somatic SV breakpoint matrix by annotating for every sample the presence or absence of at least one SV breakpoint within the given region. For the set of SV breakpoints associated with a given gene within a specified region in proximity to the gene (e.g.,

0–20 kb upstream, 20–50 kb upstream, 50–100 kb upstream, 0–20 kb downstream, or within the gene body), correlation between expression of the gene and the presence of at least one SV breakpoint 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 in addition to SV, were also considered. For these linear regression models, genes with at least three samples associated with an SV breakpoint within the given region were considered. Genes for which SVs were significant ($p < 0.001$, FDR < 4%) after correcting for both cancer type and copy were explored in downstream analyses. Results from both the SV only model and results from the SV+cancer type models were also highlighted in Fig. 1 and provided in Supplementary Data 2, but the *p*-values from those models were not used in selecting for genes or SVs of interest for follow-up analyses. R software version 3.0 and lm function was used, with source code available as part of Supplementary Data 7.

The method of Storey and Tibshirani[18] was used to estimate false discovery rates (FDR) for significant genes. For purposes of FDR, only genes that had SV breakpoints falling within the given region relative to the gene in at least three cases were tested; for example, for the 0–20 kb upstream region, 6257 genes were tested, where 384 genes were significant at a nominal *p*-value of < 0.001 (using a stringent cutoff, with ~6 genes expected by chance due to multiple testing, or FDR < 2%); the other genomic region windows yielded similar results. For each genomic region window, the FDR for genes significant at the $p < 0.001$ level did not exceed 4%. In addition, permutation testing of the 0–20 kb upstream dataset was carried out, whereby the SV events were randomly shuffled (by shuffling the patient ids) and the linear regression models (incorporating both cancer type and copy number) were used to compute expression versus permuted SV breakpoint associations; for each of 1000 permutation tests, the number of nominally significant genes at $p < 0.001$ was computed and compared with results from the actual datasets. Of the 25,259 genes represented in the entire RNA-seq dataset, 20,859 genes had at least three samples with SV breakpoints for at least one of the five regions tested (gene body, 0–20 kb upstream, 20–50 kb upstream, 50–100 kb upstream, 0–20 kb downstream). The number of genes significant (nominal $p < 0.001$) for any one of the five regions was 1575. By a very conservative estimate, the number of genes that might arise by multiple testing in relation to the 1575 gene set should not exceed $5 \times 0.001 \times 20859 = 104$ (five genomic regions × *p*-value threshold used × number of genes tested for at least one region), which would correspond to a global estimated FDR of ~6.6%.

Integrative analyses using enhancer genomic coordinates. Gene boundaries and locations of enhancer elements were obtained from Ensembl (GRCh37 build). Enhancer elements found in multiple cell types (using Ensembl "Multicell" filter, accessed April 1, 2016) were used[20]. As previously described[20], the Ensembl team first reduced all available experimental data for each cell type into a cell type-specific annotation of the genome; consensus "Multicell" regulatory features of interest, including predicted enhancers, were then defined. For each SV breakpoint 0–20 kb 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 breakpoint mate). Only SVs with breakpoints on the distal side from the gene were considered in this analysis; in other words, for genes on the negative strand, the upstream sequence of the breakpoint should be fused relative to the breakpoint coordinates, and for genes on the positive strand, the downstream sequence of the breakpoint (denoted as negative orientation) should be fused relative to the breakpoint coordinates.

Statistical analysis. All *p*-values were two-sided unless otherwise specified.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used in this study are publicly available. Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium is described here[11] and available for download at https://dcc.icgc.org/releases/PCAWG [dcc.icgc.org]. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/ [docs.icgc.org]. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login [dbgap.ncbi.nlm.nih.gov]) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco [icgc.org]) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization. The consensus SV calls are available from synapse (https://www.synapse.org/#!Synapse:syn7596712). Copy number data are available from synapse (https://www.synapse.org/#!Synapse:

Code availability

R source code written for this study is provided as part of Supplementary Data 7. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg [dockstore.org] under the GNU General Public License v3.0, which allows for reuse and distribution.

References

1. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (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. Gröschel, S. et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381 (2014).
7. Northcott, P. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
8. Peifer, M. et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).
9. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
10. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
11. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
12. PCAWG Transcriptome Core Group. et al. Genomic basis for RNA alterations in cancer. *Nature* https://doi.org/10.1038/s41586-020-1970-0 (2020).
13. Zhang, Y. et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. *Cell Rep.* **24**, 515–527 (2018).
14. Chen, F. et al. Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Rep.* **14**, 2476–2489 (2016).
15. Valentijn, L. et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat. Genet.* **47**, 1411–1414 (2015).
16. Nagel, I. et al. Deregulation of the telomerase reverse transcriptase (TERT) gene by chromosomal translocations in B-cell malignancies. *Blood* **116**, 1317–1320 (2010).
17. Hoadley, K. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
18. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
19. Zhang, Y. et al. A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations. *Cancer Cell* **31**, 820–832 (2017).
20. Zerbino, D., Wilder, S., Johnson, N., Juettemann, T. & Flicek, P. The ensembl regulatory build. *Genome Biol.* **16**, 56 (2015).
21. Weir, B. et al. Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007).
22. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
23. Wu, K. et al. Direct activation of TERT transcription by c-MYC. *Nat. Genet.* **21**, 220–224 (1999).
24. Creighton, C. Using large-scale molecular data sets to improve breast cancer treatment. *Breast Cancer Manag.* **1**, 57–64 (2012).
25. Lewis, K. & Tollefsbol, T. Regulation of the telomerase reverse transcriptase subunit through epigenetic mechanisms. *Front Genet.* **7**, 83 (2016).
26. Zack, T. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
27. Lawrence, M. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
28. Chen, F. et al. Multiplatform-based molecular subtypes of non-small cell lung cancer. *Oncogene* **36**, 1384–1393 (2016).
29. Chen, F. et al. Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases. *Clin. Cancer Res.* **24**, 2182–2193 (2018).
30. Chen, F. et al. Pan-urologic cancer genomic subtypes that transcend tissue of origin. *Nat. Commun.* **8**, 199 (2017).

Author contributions

Conceptualization: C.J.C.; Methodology: C.J.C. and Y.Z.; Investigation: F.C., Y.Z., C.J.C., N.A.F., Y.H., M.F., H.N., Z.Z., and A.B.; Formal Analysis: F.C., Y.Z., and C.J.C.; Visualization: C.J.C. and F.C.; Writing: C.J.C.; Supervision: C.J.C. Data Curation: The PCAWG Transcriptome Working Group (led by Alvis Brazma, Gunnar Rätsch, and Angela N. Brooks). Data Curation: The PCAWG Structural Variation Working Group (led by Peter J. Campbell and Rameen Beroukhim).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13885-w.

Correspondence and requests for materials should be addressed to C.J.C.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PCAWG Transcriptome Working Group

Samirkumar B. Amin[9,10,11], Philip Awadalla[12,13], Peter J. Bailey[14], Alvis Brazma[2], Angela N. Brooks[15,16,17], Claudia Calabrese[2,18], Aurélien Chateigner[19], Isidro Cortés-Ciriano[20,21,22], Brian Craft[23], David Craft[15], Chad J. Creighton[1,6,7,8], Natalie R. Davidson[24,25,26,27,28], Deniz Demircioğlu[29,30], Serap Erkek[18], Nuno A. Fonseca[2,31], Milana Frenkel-Morgenstern[32], Mary J. Goldman[23], Liliana Greger[2], Jonathan Göke[29,33], Yao He[3,4,34], Katherine A. Hoadley[35,36], Yong Hou[37,38], Matthew R. Huska[39], Andre Kahles[24,26,27,40,41], Ekta Khurana[42,43,44,45], Helena Kilpinen[46], Jan O. Korbel[2,18], Fabien C. Lamaze[12], Kjong-Van Lehmann[24,25,26,27,41], Chang Li[37,38], Siliang Li[37,38], Xiaobo Li[37,38], Xinyue Li[37], Dongbing Liu[37,38], Fenglin Liu[34,47], Xingmin Liu[37,38], Maximillian G. Marin[17], Julia Markowski[39], Matthew Meyerson[15,16,48,49,50], Tannistha Nandi[51], Morten Muhlig Nielsen[52], Akinyemi I. Ojesina[53,54,55], B.F.Francis Ouellette[56,57], Qiang Pan-Hammarström[37,58], Peter J. Park[21,59], Chandra Sekhar Pedamallu[15,50,60], Jakob Skou Pedersen[52,61], Marc D. Perry[19,62], Gunnar Rätsch[24,27,28,41,63,64], Roland F. Schwarz[2,39,65,66], Yuichi Shiraishi[67], Reiner Siebert[68,69], Cameron M. Soulette[17], Stefan G. Stark[25,27,70,71], Oliver Stegle[2,18,72], Hong Su[37,38], Patrick Tan[51,73,74,75], Bin Tean Teh[73,74,75,76,77], Lara Urban[2,18], Jian Wang[37], Sebastian M. Waszak[18], Kui Wu[37,38], Qian Xiang[78], Heng Xiong[37,38], Sergei Yakneen[18], Huanming Yang[37], Chen Ye[37,38], Christina K. Yung[19], Fan Zhang[34], Junjun Zhang[19], Xiuqing Zhang[37], Zemin Zhang[3,34,79], Liangtao Zheng[29], Jingchun Zhu[23] & Shida Zhu[37,38]

[9]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [10]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA. [11]Quantitative & Computational Biosciences Graduate Program, Baylor College of Medicine, Houston, TX 77030, USA. [12]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. [13]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada. [14]University of Glasgow, CRUK Beatson Institute for Cancer Research, Bearsden, Glasgow G61 1BD, UK. [15]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [16]Dana-Farber Cancer Institute, Boston, MA 02215, USA. [17]University of California Santa Cruz, Santa Cruz, CA 95064, USA. [18]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg 69117, Germany. [19]Genome Informatics Program, Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. [20]Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [21]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. [22]Ludwig Center, Harvard Medical School, Boston, MA 02115, USA. [23]UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [24]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [25]ETH Zurich, Department of Biology, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland. [26]ETH Zurich, Department of Computer Science, Zurich 8092, Switzerland. [27]SIB Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland. [28]Weill Cornell Medical College, New York, NY 10065, USA. [29]Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore. [30]School of Computing, National University of Singapore, Singapore 117417, Singapore. [31]CIBIO/InBIO - Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão 4485-601, Portugal. [32]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed 13195, Israel. [33]National Cancer Centre Singapore, Singapore 169610, Singapore. [34]Peking University, Beijing 100871, China. [35]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [36]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [37]BGI-Shenzhen, Shenzhen 518083, China. [38]China National GeneBank-Shenzhen, Shenzhen 518083, China. [39]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin 13125, Germany. [40]ETH Zurich, Department of Biology, Zürich 8093, Switzerland. [41]University Hospital Zurich, Zurich 8091, Switzerland. [42]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA. [43]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10065, USA. [44]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA. [45]Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10065, USA. [46]University College London, London WC1E 6BT, UK. [47]School of Life Sciences, Peking University, Beijing 100180, China. [48]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern 3010, Switzerland. [49]Department of Pathology, The University of Melbourne, Melbourne, VIC 3052, Australia. [50]Harvard Medical School, Boston, MA 02115, USA. [51]Genome Institute of Singapore, Singapore 138672, Singapore. [52]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus N 8200, Denmark. [53]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA. [54]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA. [55]O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA. [56]Genome Informatics, Ontario Institute for Cancer Research, Toronto, ON M5G 2C4, Canada. [57]Department of Cell and Systems Biology, University of Toronto, Toronto, ON M5S 3G5, Canada. [58]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm 14183, Sweden. [59]Ludwig Center at Harvard, Boston, MA 02115, USA. [60]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [61]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus 8000, Denmark. [62]Department of Radiation Oncology, University of California San Francisco, San Francisco, CA 94518, USA. [63]Department of Biology, ETH Zurich, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland. [64]Department of Computer Science, ETH Zurich, Zurich 8092, Switzerland. [65]German Cancer Consortium (DKTK), Partner site Berlin, Germany. [66]German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [67]The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan. [68]Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Kiel 24118, Germany. [69]Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm 89081, Germany. [70]Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [71]Korea University, Seoul 02481, South Korea. [72]Division of Computational

Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [73]Cancer Science Institute of Singapore, National University of Singapore, Singapore 169609, Singapore. [74]Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore. [75]SingHealth, Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore 169609, Singapore. [76]Institute of Molecular and Cell Biology, Singapore 169609, Singapore. [77]Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, Singapore 169610, Singapore. [78]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. [79]University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

PCAWG Structural Variation Working Group

Kadir C. Akdemir[79], Eva G. Alvarez[80,81,82], Adrian Baez-Ortega[83], Rameen Beroukhim[15,50,60], Paul C. Boutros[12,84,85,86], David D.L. Bowtell[87,88], Benedikt Brors[89,90,91], Kathleen H. Burns[92], Peter J. Campbell[93,94], Kin Chan[95], Ken Chen[79], Isidro Cortés-Ciriano[20,21,59], Ana Dueso-Barroso[96], Andrew J. Dunford[15], Paul A. Edwards[97,98], Xavier Estivill[99], Dariush Etemadmoghadam[87,100], Lars Feuerbach[89], J. Lynn Fink[101,102], Milana Frenkel-Morgenstern[32], Dale W. Garsed[87,100], Mark Gerstein[103,104,105,106], Dmitry A. Gordenin[107], David Haan[108], James E. Haber[109], Julian M. Hess[15,110], Barbara Hutter[91,111,112], Marcin Imielinski[113,114], David T.W. Jones[115,116], Young Seok Ju[94,117], Marat D. Kazanov[118,119,120], Leszek J. Klimczak[121], Youngil Koh[122,123], Jan O. Korbel[2,18], Kiran Kumar[15], Eunjung Alice Lee[124], Jake June-Koo Lee[21,22], Yilong Li[94], Andy G. Lynch[97,98,125], Geoff Macintyre[97], Florian Markowetz[97,98], Iñigo Martincorena[94], Alexander Martinez-Fundichely[126,127,128], Matthew Meyerson[15,16,49,50,129], Satoru Miyano[67], Hidewaki Nakagawa[5,130], Fabio C.P. Navarro[105], Stephan Ossowski[131,132,133], Peter J. Park[21,22], John V. Pearson[134,135], Montserrat Puiggròs[101], Karsten Rippe[136], Nicola D. Roberts[94], Steven A. Roberts[137], Bernardo Rodriguez-Martin[80,81,82], Steven E. Schumacher[15,138], Ralph Scully[139], Mark Shackleton[100,140], Nikos Sidiropoulos[141], Lina Sieverling[89,142], Chip Stewart[15], David Torrents[101,143], Jose M.C. Tubio[80,81,82], Izar Villasante[101], Nicola Waddell[134,135], Jeremiah A. Wala[15,16,50], Joachim Weischenfeldt[18,141,144], Lixing Yang[145], Xiaotong Yao[113,146], Sung-Soo Yoon[113], Jorge Zamora[80,81,82,94] & Cheng-Zhong Zhang[15,16,50]

[80]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain. [81]Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain. [82]The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo 36310, Spain. [83]Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK. [84]Department of Medical Biophysics, University of Toronto, Toronto, ON M5S 1A8, Canada. [85]Department of Pharmacology, University of Toronto, Toronto, ON M5S 1A8, Canada. [86]University of California Los Angeles, Los Angeles, CA 90095, USA. [87]Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia. [88]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3052, Australia. [89]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [90]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [91]National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg 69120, Germany. [92]Johns Hopkins School of Medicine, Baltimore, MD 21205, USA. [93]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK. [94]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [95]University of Ottawa Faculty of Medicine, Department of Biochemistry, Microbiology and Immunology, Ottawa, ON K1H 8M5, Canada. [96]Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain. [97]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK. [98]University of Cambridge, Cambridge CB2 1TN, UK. [99]Sidra Medicine, Doha 26999, Qatar. [100]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC 3052, Australia. [101]Barcelona Supercomputing Center, Barcelona 08034, Spain. [102]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia. [103]Department of Computer Science, Princeton University, Princeton, NJ 08540, USA. [104]Department of Computer Science, Yale University, New Haven, CT 06520, USA. [105]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [106]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [107]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC 27709, USA. [108]Biomolecular Engineering Department, University of California, Santa Cruz, Santa Cruz, CA 95064, USA. [109]Brandeis University, Waltham, MA 02254, USA. [110]Massachusetts General Hospital Center for Cancer Research, Charlestown, MA 02129, USA. [111]German Cancer Consortium (DKTK), Heidelberg 69120, Germany. [112]Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [113]New York Genome Center, New York, NY 10013, USA. [114]Weill Cornell Medicine, New York, NY 10065, USA. [115]Hopp Children's Cancer Center (KiTZ), Heidelberg 69120, Germany. [116]Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [117]Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea. [118]Skolkovo Institute of Science and Technology, Moscow 121205, Russia. [119]A.A.Kharkevich Institute of Information Transmission Problems, Moscow 127051, Russia. [120]Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow 117997, Russia. [121]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC 27709, USA. [122]Center For Medical Innovation, Seoul National University Hospital, Seoul 03080, South Korea. [123]Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, South Korea. [124]Division of Genetics and Genomics, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA. [125]School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, Fife KY16 9SS, UK. [126]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA. [127]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY

10065, USA. [128]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA. [129]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern 3010, Switzerland. [130]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. [131]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Spain. [132]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen 72074, Germany. [133]Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain. [134]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane 4006, Australia. [135]Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia. [136]German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [137]School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA 99164, USA. [138]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [139]Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA. [140]Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, VIC 3000, Australia. [141]Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen 2200, Denmark. [142]Faculty of Biosciences, Heidelberg University, Heidelberg 69120, Germany. [143]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. [144]Department of Urology, Charité Universitätsmedizin Berlin, Berlin 10117, Germany. [145]Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA. [146]Tri-Institutional PhD Program of Computational Biology and Medicine, Weill Cornell Medicine, New York, NY 10065, USA