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

Yiqun Zhang¹, Fengju Chen¹, Nuno A. Fonseca², Yao He³,⁴, Masashi Fujita⁵, Hidewaki Nakagawa⁵, Zemin Zhang³,⁴, Alvis Brazma², PCAWG Transcriptome Working Group, PCAWG Structural Variation Working Group, Chad J. Creighton¹,⁶,⁷,⁸* & 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.
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\(^3\). 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 \(EVT\) 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 realigned 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\(^12\). 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 TMRRS2-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 (~6–8x) of low-pass sequencing. With a genome-wide analysis involving a large sample size and much deeper sequencing coverage (~30–60x), 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\(^13\), 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. Linear regression models also evaluated significant associations between expression and associated SV event. Numbers above and below zero point of correlation (p < 0.001, 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. b Positively correlated, SV vs expression. 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 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.
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, PTEN11). 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, \( p < 10^{-14} \), 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 (\( p < 10^{-6} \), 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.

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

| Region: Gene       | 0–20 kb upstream | 20–50 kb upstream | 50–100 kb upstream | Gene body | 0–20 kb downstream |
|--------------------|------------------|-------------------|-------------------|-----------|-------------------|
|                    | \( n \) | \( t \) | \( n \) | \( t \) | \( n \) | \( t \) | \( n \) | \( t \) | \( n \) | \( t \) | \( n \) | \( t \) | \( n \) | \( t \) | \( n \) | \( t \) |
| 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 |
| CDX2               | 7    | 0.33 | 14  | 3.78 | 14  | 0.02 | 41  | 2.8  | 11  | 3.01 |
| HMG1A2             | 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 |
| TBL1R1X            | 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   | 1.11 | 3   | -0.61|
| SMARCE1            | 2    | NA   | 6   | 4.7  | 6   | 3.29 | 6   | 0.75 | 1   | NA   |
| PDGFR5             | 3    | 3.81 | 4   | 0.07 | 6   | 0.04 | 7   | 1.51 | 2   | NA   |
| NFI1               | 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 | 5   | 4.5  |
| 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   |
| NXX1-1             | 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.
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, promoter mutations, promoter viral integration, and MYC amplification. 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 (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 (PD1L)—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 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 CD4 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 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; non-methylation results in the opening of CTCF binding sites and the transcriptional repression of TERT. 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, 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 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 (e).

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.
number of genes than may have been previously thought. A recent study by Weischenfeldt et al., 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, 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. The PCAWG workflows are also available as Docker images through Dockstore enabling researchers to replicate the steps involved in the data assembly. In all, 2658 patients with whole-genome data were represented in the PCAWG datasets, spanning a range of cancer types, including brain tumors, breast, lung, liver, colorectal, lymphoma, prostate, esophagus, 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 graylisted (excluded from some analyses carried out as part of PCAWG-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. 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:syn7969712). For copy number, the calls made by the Sanger group using the Ascat NGS algorithm with default parameters were used, which data are available at the ICGC-Cancer Pan-Cancer Data Portal (https://dcc.icgc.org/cancer). The categorization of copy number was as follows: LOH, loss of heterozygosity; 0, no change; 1–2, gain of one copy; 2–4, gain in both copies; and 4+, gain in more than both copies.

For RNA-seq data, alignments by both STAR (version 2.5.3a) and TopHat2 (version 2.0.12) were used to generate a combined gene expression data. Gene counts 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. 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 which is detailed in ref. 13. 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. In particular, technical variation in one 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 better explained on the basis of cell 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) platform, San Diego, CA, as previously described. To help correct for batch effects between methylation data platforms (HM450 versus HM27), we used the combat software with R software version 3.0 (with 27K vs 45K as the “batch” and cancer type as the “experimental group”, R code available at https://www.biobase.de/papers/assets/CombatR.Downloads/files/CombatR). As we have done in previous pan-cancer studies utilizing The Cancer Genome Atlas methylation datasets. For each of 8822 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 log-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 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 model were combined to define a gene set. 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 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 permutation, the number of non-cancer type were selected 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 SVs with breakpoint in the data, 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 × 0.001 × 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 (GRCb37 build). Enhancer elements found in multiple cell types (using Ensembl “Multicell” filter, accessed April 1, 2016) were used. As previously described, 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 breakpoint). Only SVs with breakpoints on the data, 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 × 0.001 × 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%.

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 and available for download at https://dcc.icgc.org/research/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=dac) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://igc.org/daco [igc.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:syn5896712). Copy number data are available from synapse (https://www.synapse.org/#!Synapse:syn57969712).
syn2364727. The gene expression dataset is available from synapse (https://www.synapse.org/#!Synapse:syn553991).

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.

Received: 22 December 2017; Accepted: 4 December 2019; Published online: 05 February 2020

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, 919–923 (2013).
6. Grüschol, S. et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell 157, 369–381 (2014).
7. Nordcott, P. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511, 428–434 (2014).
8. Peifer, M. et al. Telomerase activation by genome 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 copy-number alteration 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. 24, 2476–2489 (2016).
15. Velantiš, J. et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. Nat. Genet. 47, 1411–1414 (2015).
16. Nagel, I. et al. Deregression 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., Jussennam, T. & Flice, P. The ensemble 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. Multipipeline-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-urolucic cancer genomic subtypes that transcend tissue of origin. Nat. Commun. 8, 199 (2017).

Acknowledgements
This work was supported in part by National Institutes of Health (NIH) grant P30CA121523 (C. Creighton) and Cancer Prevention and Research Institute of Texas (CPRIT) grant RP120713 C2 (C. Creighton). This work was made possible through the resources and datasets made available by the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium. In particular, we wish to acknowledge the PCAWG Transcriptome Working Group (led by Alvis Brazma, Gunnar Rätsch, and Angela N. Brooks) and the PCAWG Structural Variation Working Group (led by Peter J. Campbell and Rameen Beroukhim). Furthermore, we acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collection, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

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.; 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 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.

© The Author(s) 2020
PCAWG Transcriptome Working Group

Samirkumar B. Amin, Philip Awadalla, Peter J. Bailey, Alvis Brazy, Angela N. Brooks, Claudia Cabalrese, Aurélien Chateau, Isidro Cortés-Ciriano, Brian Craft, David Craft, Chad J. Creighton, Natalie R. Davidson, Deniz Demircioğlu, Serap Erkek, Nuno A. Fonseca, Milana Frenkel-Morgenstern, Mary J. Goldman, Liliana Gregor, Jonathan Göke, Yao He, Katherine A. Hoadley, Matthew R. Huska, Andre Kahles, Ekta Khurana, Jan O. Korbel, Fabien C. Lamaze, Kjong-Van Lehmann, Chang Li, Siliang Li, Xiaobo Li, Xinyue Li, Dongbing Liu, Fenglin Liu, Xingmin Liu, Maximilian G. Marin, Julia Markowski, Matthew Meyerson, Tannishtha Nandi, Morten Muhlig Nielsen, Akinyemi I. Ojesina, B. F. Francis Ouellette, Qiang Pan, Peter J. Park, Chandra Sekhar Pedamallu, Jakob Skou Pedersen, Marc D. Perry, Gunnar Rätsch, Roland F. Schwarz, Yiichi Shiraishi, Reiner Siebert, Cameron M. Soullete, Stefan G. Stark, Oliver Stegle, Hong Su, Patrick Tan, Bin Tean Teh, Hong Su, Lara Urban, Jian Wang, Sebastian M. Waszak, Kui Wu, Qian Xian, Heng Xiong, Sergei Yakneen, Chen Ye, Christina K. Yung, Fan Zhang, Junjun Zhang, Xiuqiang Zhang, Zemin Zhang, Liangtao Zheng, Jingchun Zhu, and Shida Zhu.
Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA. 
Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern 3010, Switzerland. 
RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. 
Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Spain. 
Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen 72074, Germany. 
Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain. 
Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane 4006, Australia. 
Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia. 
German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. 
School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA 99164, USA. 
Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. 
Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA. 
Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, VIC 3000, Australia. 
Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen 2200, Denmark. 
Faculty of Biosciences, Heidelberg University, Heidelberg 69120, Germany. 
Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. 
Department of Urology, Charité Universitätsmedizin Berlin, Berlin 10117, Germany. 
Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA. 
Tri-Institutional PhD Program of Computational Biology and Medicine, Weill Cornell Medicine, New York, NY 10065, USA.