











OPEN

# Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing


Isidro Cortés-Ciriano <sup>1,2,3,15</sup>, Jake June-Koo Lee <sup>1,2</sup>, Ruibin Xi <sup>4</sup>, Dhawal Jain<sup>1</sup>, Youngsook L. Jung<sup>1</sup>, Lixing Yang<sup>5,6</sup>, Dmitry Gordenin <sup>7</sup>, Leszek J. Klimczak <sup>8</sup>, Cheng-Zhong Zhang <sup>1,9</sup>, David S. Pellman<sup>10,11,12</sup>, PCAWG Structural Variation Working Group<sup>13</sup>, Peter J. Park <sup>2,3</sup>  and PCAWG Consortium<sup>14</sup>

**Chromothripsis is a mutational phenomenon characterized by massive, clustered genomic rearrangements that occurs in cancer and other diseases. Recent studies in selected cancer types have suggested that chromothripsis may be more common than initially inferred from low-resolution copy-number data. Here, as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), we analyze patterns of chromothripsis across 2,658 tumors from 38 cancer types using whole-genome sequencing data. We find that chromothripsis events are pervasive across cancers, with a frequency of more than 50% in several cancer types. Whereas canonical chromothripsis profiles display oscillations between two copy-number states, a considerable fraction of events involve multiple chromosomes and additional structural alterations. In addition to non-homologous end joining, we detect signatures of replication-associated processes and templated insertions. Chromothripsis contributes to oncogene amplification and to inactivation of genes such as mismatch-repair-related genes. These findings show that chromothripsis is a major process that drives genome evolution in human cancer.**

Chromothripsis is characterized by massive genomic rearrangements that are often generated in a single catastrophic event and localized to isolated chromosomal regions<sup>1–4</sup>. In contrast to the traditional view of tumorigenesis as the gradual process of the accumulation of mutations, chromothripsis provides a mechanism for the rapid accrual of hundreds of rearrangements in a few cell divisions. This phenomenon has been studied in primary tumors of diverse histological origins<sup>5–10</sup>, but similar random joining of chromosomal fragments has also been observed in the germline<sup>11</sup>. There has been considerable progress in elucidating the mechanisms by which chromothripsis may arise, including fragmentation and subsequent reassembly of a single chromatid in aberrant nuclear structures called micronuclei<sup>2,12</sup> and the fragmentation of dicentric chromosomes during telomere crisis<sup>13,14</sup>. Chromothripsis is not specific to cancer as it can cause rare congenital human disease and can be transmitted through the germline<sup>11,15</sup>; it has also been described in plants, where it has been linked to micronucleation<sup>16</sup>. However, despite the recent rapid progress on elucidating the mechanisms of chromothripsis, much remains to be discovered regarding its cause, prevalence and consequences.

A hallmark of chromothripsis is multiple oscillations between two or three copy-number (CN) states<sup>1,6</sup>. Applying this criterion to CN profiles inferred from SNP arrays, chromothripsis was initially estimated to occur in at least 2–3% of human cancers<sup>1</sup>. Subsequent studies of large array-based datasets gave similar frequencies: 1.5% (124 out of 8,227 tumors across 30 cancer types)<sup>17</sup> and 5% (918 out of 18,394 tumors)<sup>18</sup>, with the highest frequencies detected for soft-tissue tumors (54% for liposarcomas, 24% for fibrosarcomas and 23% for sarcomas)<sup>18</sup>. These estimates relied on the detection of CN oscillations that are more-densely clustered than expected by chance<sup>8</sup>.

Whole-genome sequencing (WGS) data provide a greatly enhanced view of structural variations (SVs) in the genome<sup>19</sup>, allowing us to generate a more nuanced set of criteria for chromothripsis and enhance detection specificity<sup>3</sup>. Our previous analysis of WGS data from cutaneous melanomas already found chromothripsis-like rearrangements in 38% of these tumors (45 out of 117)<sup>10</sup>; other studies using WGS data found 60–65% for pancreatic cancer<sup>5</sup> and 32% for esophageal adenocarcinomas<sup>7</sup>. Whether these examples are outliers that reflect the unique biology of these tumors or whether

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>2</sup>Ludwig Center at Harvard, Boston, MA, USA. <sup>3</sup>Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. <sup>4</sup>School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China. <sup>5</sup>Ben May Department for Cancer Research, University of Chicago, Chicago, IL, USA. <sup>6</sup>Department of Human Genetics, The University of Chicago, Chicago, IL, USA. <sup>7</sup>Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, US National Institutes of Health, Durham, NC, USA. <sup>8</sup>Integrative Bioinformatics Group, National Institute of Environmental Health Sciences, US National Institutes of Health, Durham, NC, USA. <sup>9</sup>Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>10</sup>Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>11</sup>Department of Cell Biology, Harvard Medical School, Blavatnik Institute, Boston, MA, USA. <sup>12</sup>Howard Hughes Medical Institute, Boston, MA, USA. <sup>13</sup>A list of members and their affiliations appears at the end of the paper. <sup>14</sup>A list of members and their affiliations appears in the Supplementary Note. <sup>15</sup>Present address: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. e-mail: [peter\\_park@hms.harvard.edu](mailto:peter_park@hms.harvard.edu)

they suggest a more general underestimation of the frequency of chromothripsis remained unclear.

Motivated by the importance of chromothripsis during tumor evolution and the need for more-comprehensive analyses, we determined the frequency and spectrum of chromothripsis events in the WGS data for 2,658 patients with cancer comprising 38 cancer types generated by the ICGC and TCGA projects, and aggregated by the PCAWG Consortium. These sequencing data were re-analyzed with standardized pipelines to align to the human genome (reference build hs37d5) and to identify germline variants and somatic mutations<sup>20</sup>. In addition to deriving more-accurate estimates of the per-tumor type prevalence of chromothripsis, we determined the size and genomic distribution of such events, examined their role in the amplification of oncogenes or loss of tumor-suppressor genes, described their relationship to genome ploidy and investigated whether their presence is correlated with patient survival. Our chromothripsis calls can be browsed at the accompanying website (<http://compbio.med.harvard.edu/chromothripsis/>).

## Results

**Prevalence of chromothripsis across cancer types.** We first sought to formulate a set of criteria for identifying chromothripsis events with varying complexities (Fig. 1a). The acknowledged model of chromothripsis posits that some of the DNA fragments generated by the shattering of the DNA are lost; thus, CN oscillations between two or three states<sup>1,6</sup> are an obvious first criterion (Fig. 1a). Such deletions also lead to interspersed loss of heterozygosity (LOH) or altered haplotype ratios if there is only a single copy of the parental homolog of the fragmented chromatid. Although chromosome shattering and reassembly has been experimentally demonstrated to generate chromothripsis<sup>2</sup>, template-switching DNA-replication errors can generate a similar pattern<sup>21</sup>. Indeed, shattering and replication error models are not mutually exclusive and could co-occur<sup>2</sup>. Therefore, for the discussion below we will refer generally to ‘chromothripsis’ as encompassing both classes of models.

To detect chromothripsis in WGS data, we developed ShatterSeek (Methods and Supplementary Note). A key feature of our method is to identify clusters of breakpoints belonging to SVs that are interleaved—that is, the regions bridged by their breakpoints overlap instead of being nested (Fig. 1)—as is expected from random joining of genomic fragments. This encompasses the many cases that do not display simple oscillations (for example, partially oscillating CN profiles with interspersed amplifications) and oscillations that span multiple CN levels due to aneuploidy<sup>5,22</sup>. Rearrangements in chromothripsis should also follow a roughly even distribution for the different types of fragment joins (duplication-like, deletion-like, head-to-head and tail-to-tail inversions, which are shown in blue, orange, black and green, respectively, in Fig. 1a and throughout) and have breakpoints that are randomly distributed across the affected region<sup>1–3</sup>. Finally, we use interchromosomal SVs to identify chromothripsis events that involve multiple chromosomes. In the Supplementary Note, we have compiled the criteria that have been used in 27 major chromothripsis-related studies to date.

After removing low-quality samples using stringent quality control, we applied ShatterSeek to 2,543 tumor–normal pairs of 37 cancer types (Methods and Supplementary Table 1). Of those 2,543 pairs, 2,428 cases had SVs and were analyzed further. To tune the parameters in our method, we used statistical thresholds and visual inspection. For the minimum number of oscillating CN segments, we used two thresholds: high-confidence calls display oscillations between two states in at least seven adjacent segments, whereas low-confidence calls involve between four and six segments (Fig. 1b and Supplementary Note). The analyses described in the subsequent sections were performed using the high-confidence call set unless noted otherwise.

We first focused on the 1,427 nearly diploid genomes (ploidy  $\leq 2.1$ ; Supplementary Table 1), in which detection of chromothripsis is more straightforward. We defined as ‘canonical’ those events in which more than 60% of the CN segments in the affected region oscillated between two states (canonical events in polyploid tumors are described later). The frequency of canonical chromothripsis events is more than 40% for multiple cancer types, such as glioblastomas (50%) and lung adenocarcinomas (40%). These frequencies are much higher than previous estimates<sup>17,18</sup>.

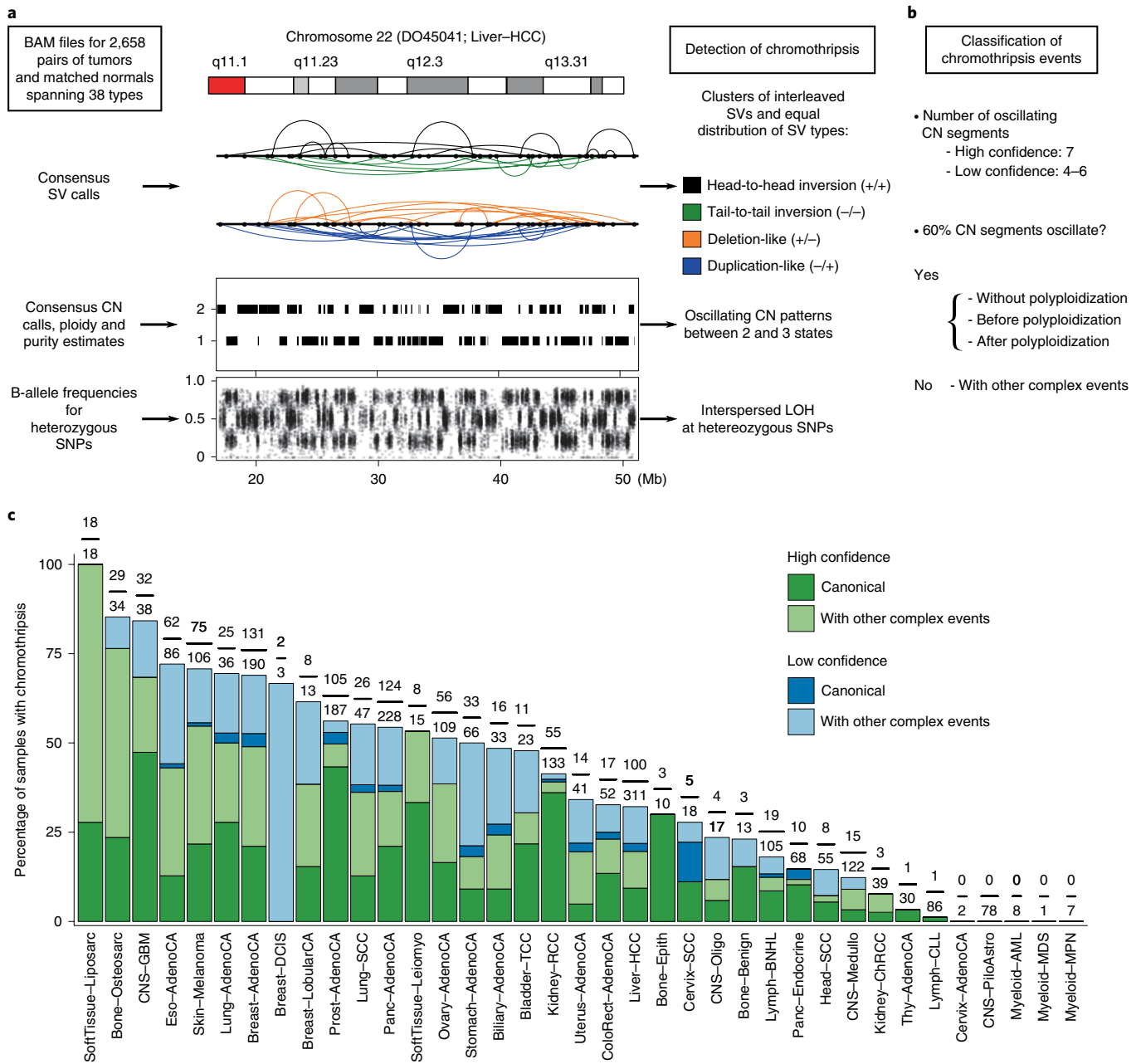
When we extend our analysis to the entire cohort, we identify high-confidence events in 29% of the samples (734 out of 2,543), affecting 3.2% of all chromosomes (Fig. 1c and Supplementary Dataset 1). When low-confidence calls are included, the percentages increase to 40% and 5.3%, respectively (Supplementary Dataset 2).

The frequency varies markedly across cancer types. At the high end, we find that 100% of liposarcomas and 77% of osteosarcomas exhibit high-confidence chromothripsis (Fig. 1c and Supplementary Fig. 1). Although a higher susceptibility of these cancer types to chromothripsis has been described<sup>1,22</sup>, our estimated frequencies are substantially higher. Melanomas, glioblastomas and lung adenocarcinomas showed evidence of chromothripsis in more than 50% of cases (Fig. 1c). By contrast, the frequencies were lowest in thyroid adenocarcinomas (3.3%,  $n=30$ ), chronic lymphocytic leukemia (1.2%,  $n=86$ ) and pilocytic astrocytomas (0%,  $n=78$ ); in the other tumor types with low incidence, the sample sizes were too small to give meaningful estimates. Consistent with previous reports<sup>23,24</sup>, we find that chromothripsis is enriched in chromosomes 3 and 5 in kidney renal cell carcinomas and chromosome 12 in liposarcomas (Supplementary Fig. 1a). Overall, these results indicate a much greater prevalence of chromothripsis in a majority of human cancers than previously estimated<sup>10,17,18</sup>.

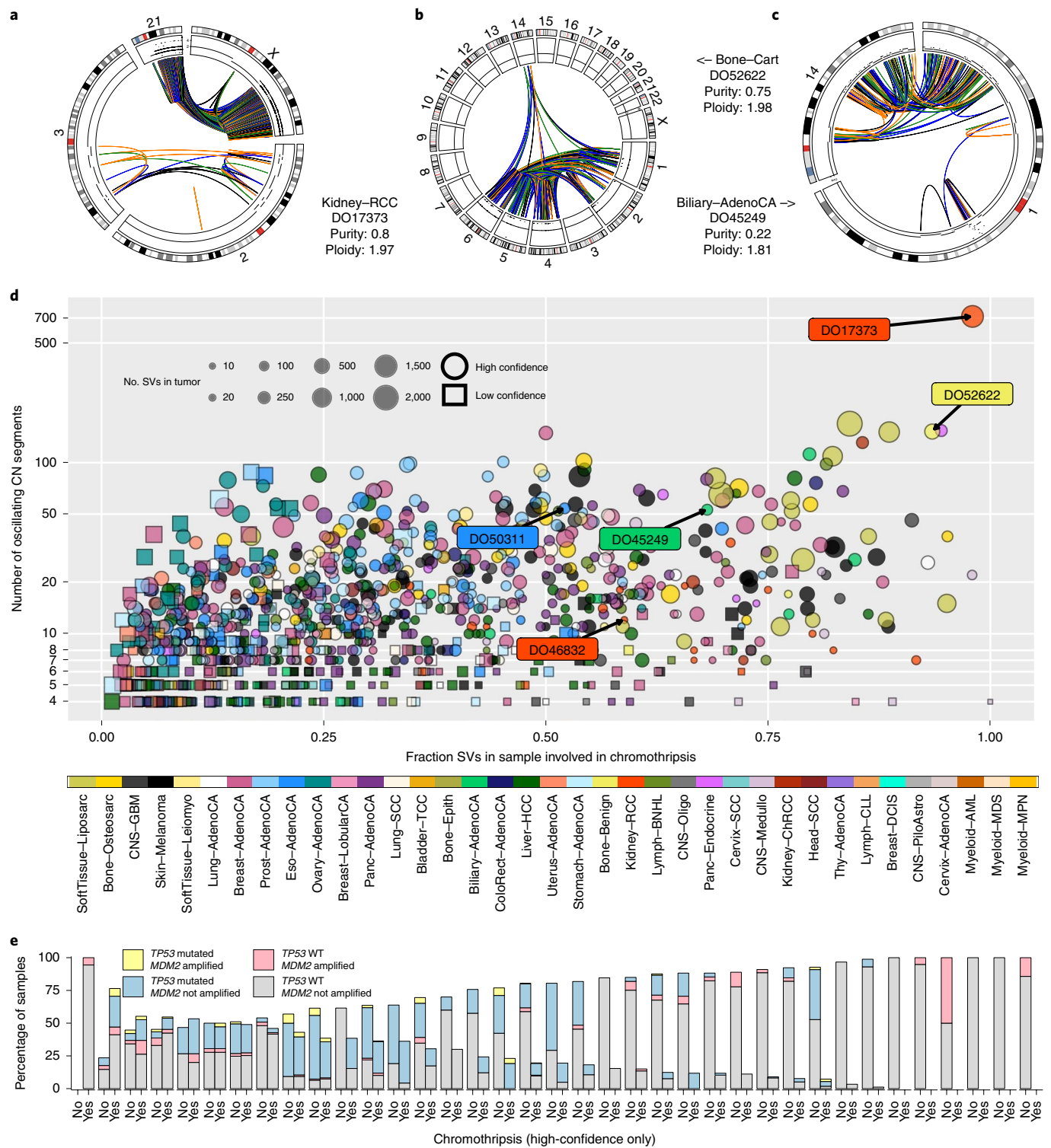
**Understanding the difference between our frequency estimates and previous ones.** Our estimates are in accordance with recent analyses in specific tumor types<sup>5,7</sup>; however, they are considerably higher than those described in previous pan-cancer studies that used array-based platforms. With higher resolution from sequencing data, improved SV algorithms and refined criteria, we are able to provide more-accurate estimates.

To better understand the discrepancy between WGS-based studies, we carried out a detailed comparison using previously analyzed datasets. For 109 previously described prostate adenocarcinomas<sup>25</sup>, the authors used ShatterProof<sup>26</sup> and found chromothripsis in 21% (23 out of 109). When we applied the same algorithm (with the same parameters) but using our CN and SV calls, the percentage more than doubled to 45% (49 out of 109). This indicates that the lower sensitivity of previous SV-detection methods is one of the main reasons for the discrepancy. Accurate SV detection remains challenging, especially for low-purity tumors. The SV calls that we used were generated by the PCAWG Structural Variation Working Group of the ICGC; each variant was required to be called by at least two of the four algorithms used in this analysis<sup>27</sup>.

Using ShatterSeek, we identified 11 additional cases for a total of 55% (60 out of 109). Of the 23 previously reported cases<sup>25</sup>, we missed four. The missed events are focal events comprising fewer than six SVs, which is the lowest number allowed in our criteria; the detected regions appear to be hypermutated regions characterized by tandem duplications or deletions. For the cases that we detect but that were missed previously<sup>25</sup>, visual inspection reveals that the differences are mostly due to the lower sensitivity of their SV calls (Supplementary Note). ShatterSeek has increased sensitivity by incorporating more complex patterns of oscillations and interchromosomal SVs while keeping the specificity high by imposing additional criteria on breakpoint homology to remove tandem duplications and those arising from breakage–fusion–bridge (BFB) cycles. Furthermore, we also compared our method against ChromAL<sup>5</sup> for 76 pancreatic



**Fig. 1 | Overview of the chromothripsis-calling method and the frequency of events across 37 cancer types. a**, Example of a region displaying the characteristic features of chromothripsis: cluster of interleaved SVs with equal proportions of SV types (that is, fragment joins), a CN profile that oscillates between two states and interspersed LOH. Details of the criteria are described in the Methods. Both the color scheme and the abbreviations shown in this figure are used throughout the manuscript. **b**, Classification of chromothripsis events. In a canonical event, more than 60% of the segments oscillate between two CN states; a tumor is classified as canonical if it showed at least one canonical chromothripsis event. **c**, Percentage of patients with chromothripsis events across the entire cohort. The fractions at the top of the bars are the number of tumors that showed high-confidence chromothripsis out of the total number of tumors of that type. The cancer type abbreviations used across the manuscript are as follows: Biliary-AdenoCA, biliary adenocarcinoma; Bladder-TCC, bladder transitional cell carcinoma; Bone-Benign, bone cartilaginous neoplasm, osteoblastoma and bone osteofibrous dysplasia; Bone-Epith, bone neoplasm, epithelioid; Bone-Osteosarc, sarcoma, bone; Breast-AdenoCA, breast adenocarcinoma; Breast-DCIS, breast ductal carcinoma in situ; Breast-LobularCA, breast lobular carcinoma; Cervix-AdenoCA, cervix adenocarcinoma; Cervix-SCC, cervix squamous cell carcinoma; CNS-GBM, central nervous system glioblastoma; CNS-Oligo, CNS oligodendroglioma; CNS-Medullo, CNS medulloblastoma; CNS-PiloAstro, CNS pilocytic astrocytoma; ColoRect-AdenoCA, colorectal adenocarcinoma; Eso-AdenoCA, esophagus adenocarcinoma; Head-SCC, head-and-neck squamous cell carcinoma; Kidney-ChRCC, kidney chromophobe renal cell carcinoma; Kidney-RCC, kidney renal cell carcinoma; Liver-HCC, liver hepatocellular carcinoma; Lung-AdenoCA, lung adenocarcinoma; Lung-SCC, lung squamous cell carcinoma; Lymph-CLL, lymphoid chronic lymphocytic leukemia; Lymph-BNHL, lymphoid mature B-cell lymphoma; Lymph-NOS, lymphoid not otherwise specified; Myeloid-AML, myeloid acute myeloid leukemia; Myeloid-MDS, myeloid myelodysplastic syndrome; Myeloid-MPN, myeloid myeloproliferative neoplasm; Ovary-AdenoCA, ovary adenocarcinoma; Panc-AdenoCA, pancreatic adenocarcinoma; Panc-Endocrine, pancreatic neuroendocrine tumor; Prost-AdenoCA, prostate adenocarcinoma; Skin-Melanoma, skin melanoma; SoftTissue-Leiomyo, leiomyosarcoma, soft tissue; SoftTissue-Liposarc, liposarcoma, soft tissue; Stomach-AdenoCA, stomach adenocarcinoma; Thy-AdenoCA, thyroid low-grade adenocarcinoma; and Uterus-AdenoCA, uterus adenocarcinoma.

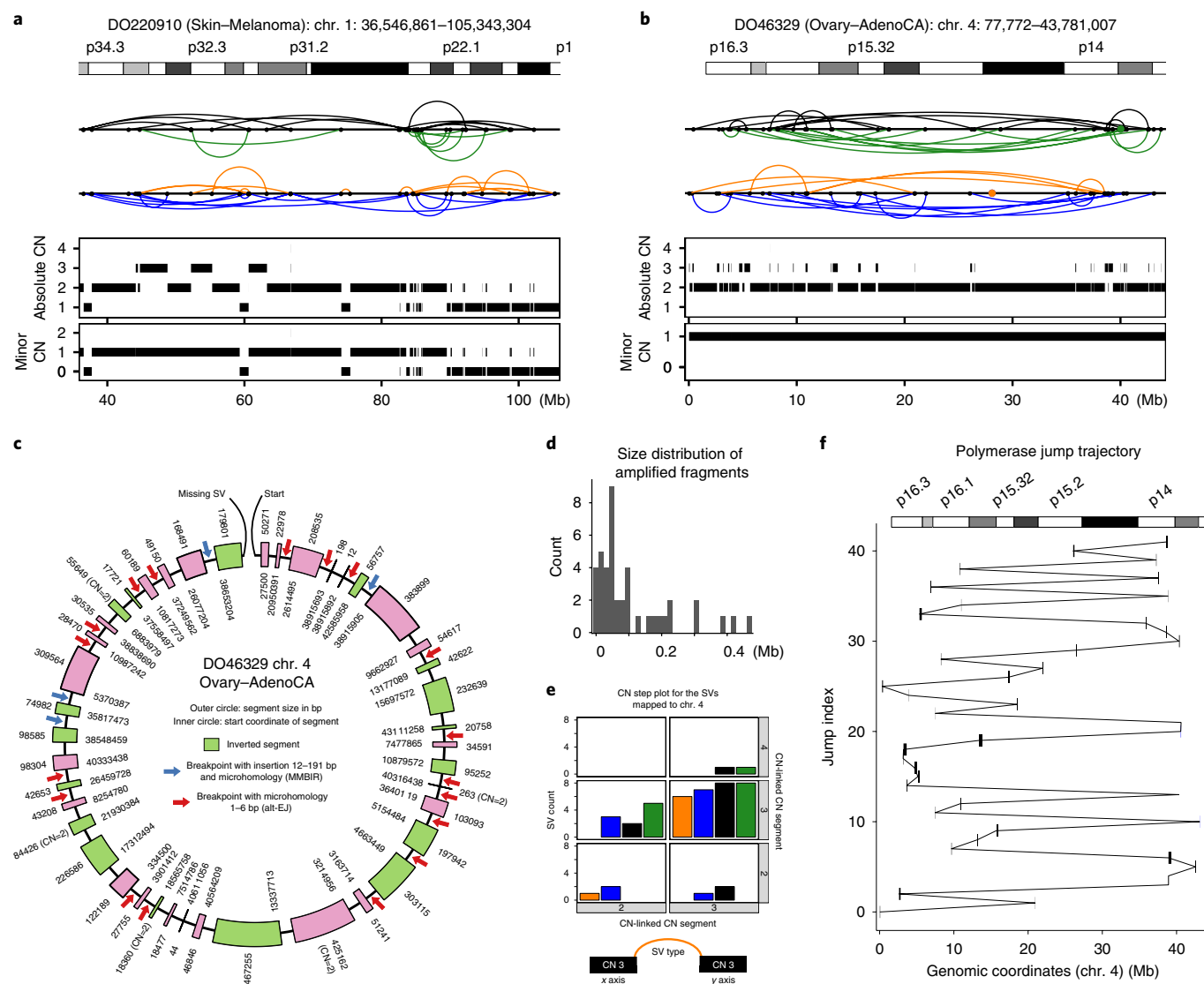


**Fig. 2 | Heterogeneity of chromothripsis events.** **a–c**, Examples of massive chromothripsis events on the background of quiescent genomes in samples from patients DO17373 (**a**), DO52622 (**b**) and DO45249 (**c**). **d**, The fraction of SVs involved in chromothripsis in each sample against the maximum number of contiguous oscillating CN segments for the high-confidence (circles) and low-confidence (squares) chromothripsis calls. **e**, Distribution of patients showing high-confidence chromothripsis, deleterious *TP53* mutations and *MDM2* amplification (CN ≥ 4). WT, wild-type allele.

tumors. Both ChromAL and ShatterSeek detect chromothripsis in the same 41 tumors (54%).

Therefore, our estimates for the frequency of chromothripsis events are supported by the following: some tumor types such as thyroid adenocarcinoma, chronic lymphocytic leukemia and pilocytic astrocytomas have few or no events; diploid tumors, which have simpler

configurations that are easier to reconstruct or verify visually, have high frequencies; the high-confidence cases were used for final estimates; more sensitive CN and SV calls result in higher frequencies for the same datasets; our estimates are in agreement with very recent analysis in specific tumor types; and our chromothripsis calls do not overlap with regions affected by chromoplexy (Supplementary Note).

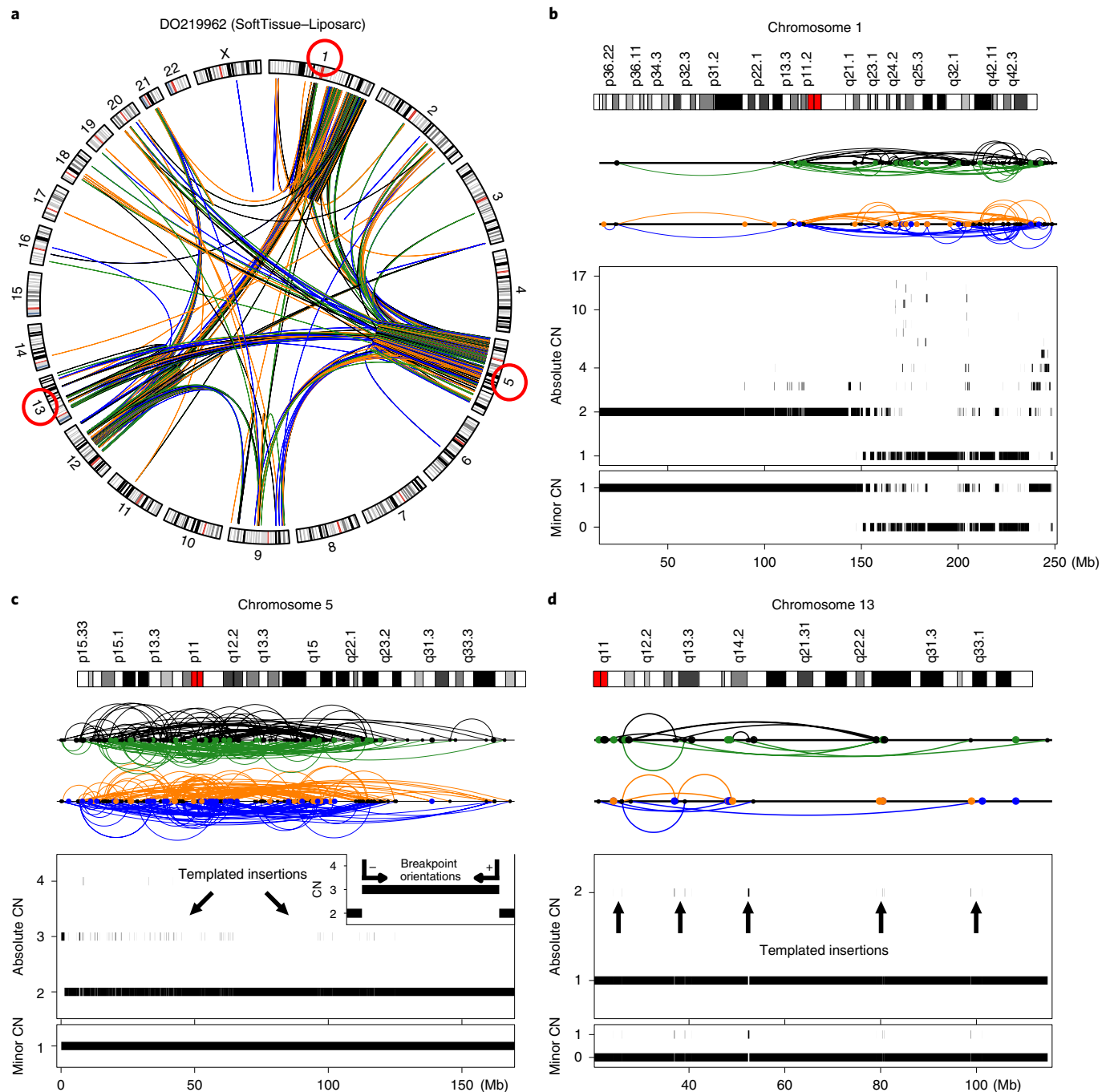


**Fig. 3 | Example of canonical chromothripsis events displaying templated insertions and evidence of MMBIR.** **a**, Evidence of chromothripsis in chromosome 1 in a skin-melanoma tumor with CN oscillations that span 3 CN levels and LOH. **b**, Example of a chromothripsis event in chromosome 4 involving low-level CN gains and absence of LOH in an ovarian adenocarcinoma. Segments at CN 3 correspond to templated insertions, as evidenced by their size, and breakpoint orientations at their edges. Breakpoints corresponding to interchromosomal SVs are depicted as colored dots in the SV profile, whereas intrachromosomal SVs are represented with black dots and colored arcs following the representation shown in Fig. 1. **c**, Reconstruction of the amplicon generated by the chromoanasythesis event detected in chromosome 4 in tumor DO46329 (see **b**). Inverted segments are depicted in green. Red arrows highlight breakpoints with short microhomology tracts, whereas blue arrows indicate the presence of small insertions at the breakpoints. The CN for all segments is 3 unless otherwise indicated. **d**, Size distribution for the templated insertions forming the amplicon depicted in **c**. **e**, CN step plot for chromosome 4 indicating that most of the SVs mapped to chromosome 4 link genomic regions at CN 3. The x and y axes correspond to the CN level of the segments linked by a given SV. The color of the bars corresponds to the four types of SVs (that is, deletion-like, duplication-like, and head-to-head and tail-to-tail inversions) indicated in Fig. 1a and considered throughout the manuscript. **f**, Trajectory of the polymerase across chromosome 4 estimated from the template-switching events shown in **c**.

**Frequent involvement of interchromosomal SVs.** An important feature of our approach is the incorporation of interchromosomal SVs to detect those events that involve multiple chromosomes. Chromothripsis affects only a single chromosome in 40% of the tumors with chromothripsis (Fig. 2a–c and Supplementary Figs. 1–3). A large number of chromosomes is frequently affected in some tumor types, for example, at least five chromosomes are affected in 61% osteosarcomas (Supplementary Figs. 1–4). In one extreme case, we found a single chromothripsis event that affected six chromosomes (Fig. 2b), with only seven of the 110 SVs on chromosome 5 being intrachromosomal. In another example (Supplementary Fig. 4d), an approximately 5-Mb region on chromosome 12 did

not display CN oscillations, but it could be linked by interchromosomal SVs to another region that does show a clear chromothripsis pattern, suggesting that the amplification of *CCND2* on chromosome 12 may have originated from chromothripsis. Chromothripsis involving multiple chromosomes is likely to have arisen either from simultaneous fragmentation of multiple chromosomes (for example, in a micronucleus or in a chromosome bridge) or from fragmentation of a chromosome that had previously undergone a non-reciprocal translocation.

**Size and complexity of chromothripsis events are highly variable.** Chromothripsis events span a wide range of genomic scale, with the

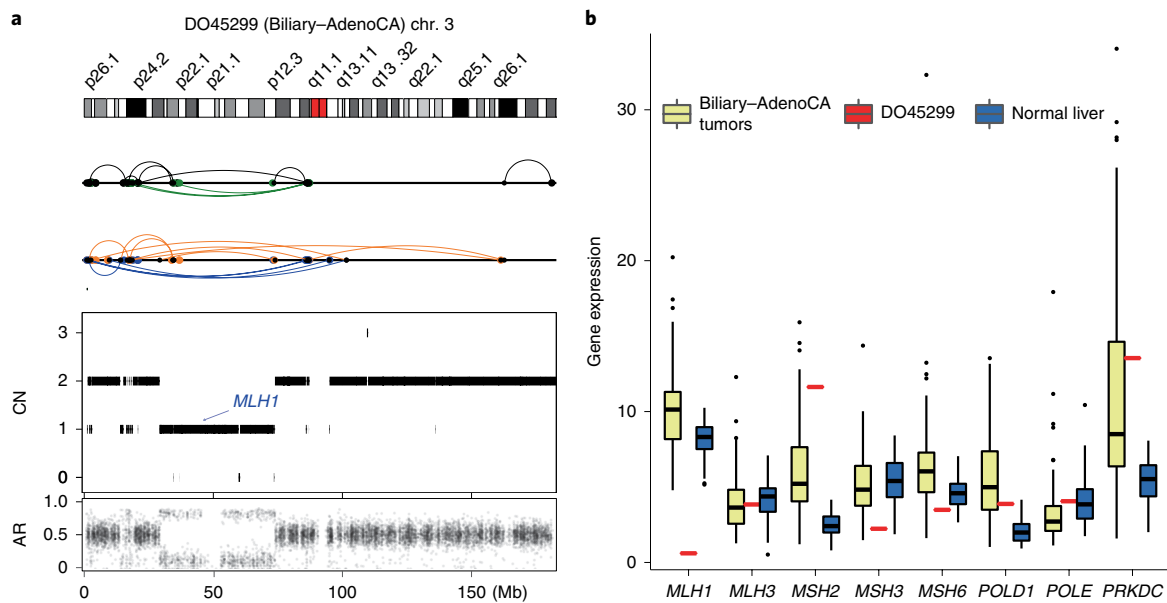


**Fig. 4 | Example of a multichromosomal chromothripsis event in a soft-tissue liposarcoma co-localized with other complex events involving templated insertions.** **a**, Scaled circos plot of the entire genome for this tumor except for chromosome Y. **b–d**, SV and CN profiles for chromosomes 1 (**b**), 5 (**c**) and 13 (**d**). Tens of CN oscillations and LOH in chromosome 1 are co-localized with additional rearrangements. The size, minor CN (from the allele with the lower number of copies) and orientation of the breakpoint junctions associated with the segments at CN 3 indicate that these are templated insertions. **c**, Inset: orientation of the breakpoint junctions at the edges of low-level CN gains originated from template switching (that is, – and + according to the annotation that we use in the manuscript).

number of breakpoints involved varying by two orders of magnitude within some tumor types (Supplementary Fig. 1c). We found that tumors had relatively focal chromothripsis events—usually a few megabases in size—that took place within an otherwise quiet genome (bottom-right quadrant in Fig. 2d). Although focal, these events can lead to the simultaneous amplification of multiple oncogenes located in different chromosomes (Supplementary Figs. 4c–e, 5a–c). Other focal events co-localize with other complex events in highly rearranged genomes (bottom-left quadrant in Fig. 2d).

Overall, our analysis reveals that there is greater heterogeneity in chromothripsis patterns than previously appreciated, both in terms of the number of SVs and chromosomes involved.

**Relationship between chromothripsis and aneuploidy.** Newly established polyploid cells have high rates of mitotic errors that generate lagging chromosomes<sup>28,29</sup>, which have been linked to chromothripsis in medulloblastomas and *in vitro*<sup>2,12,14</sup>. However, a causal relationship or even the frequency of association between polyploidy



**Fig. 5 | Chromothripsis-mediated depletion of *MLH1*.** **a**, Chromothripsis event and expression levels of DNA MMR genes in the sample of patient DO45299 (biliary adenocarcinoma). **b**, Mean expression of DNA MMR genes in a panel of 16 biliary adenocarcinomas and 16 normal liver samples. Box plots in **b** show median, first and third quartiles (boxes), and the whiskers encompass observations within a distance of 1.5x the interquartile range from the first and third quartiles. AR, allelic ratio computed for heterozygous SNPs.

and chromothripsis has not been assessed in detail. To examine the sequence of events clearly, we focused on the canonical cases, for which we can infer whether chromothripsis occurred before or after polyploidization<sup>30</sup>. For example, if the CN oscillates between two and four copies in a tetraploid tumor, we infer that polyploidization occurred after chromothripsis; on the other hand, if the oscillation occurs between three and four copies, we infer that polyploidization occurred first<sup>30</sup> (Supplementary Figs. 1, 2, 5d, 6 and Supplementary Note). Of the 194 cases in which we can distinguish the sequence of events, 74% show chromothripsis after polyploidization. This suggests that a large fraction of the canonical chromothripsis events in polyploid tumors are late events.

We observed canonical chromothripsis events in 26% of diploid-ranged tumors (431 out of 1,648) and in 40% of polyploid-ranged tumors (298 out of 748). After correcting for tumor type using the logistic regression, we estimate that, on average, the odds of chromothripsis occurring in a polyploid tumor (cases with ploidy  $\geq 2.5$ ) is 1.5 times larger than that in a diploid tumor (95% confidence interval, 1.20–1.85;  $P < 10^{-3}$ ). This increase may be due to the presence of more genomic material in polyploids, although polyploidy also reduces the sensitivity of CN and SV detection (due to a lower sequence coverage per copy) and makes it easier for the cell to lose the highly rearranged copy when intact copies are present<sup>31</sup>.

**Frequent co-localization of chromothripsis with other complex events.** About half of the chromothripsis events co-localize with other genomic alterations (Fig. 1 and Supplementary Figs. 1, 2). There is evidence across multiple tumor types that chromothripsis might occur before or after additional layers of rearrangements<sup>6–8,13,14,23</sup>. For instance, BFB cycles have been mechanistically linked to chromothripsis and telomere attrition—which results in the formation of BFB cycles, has been identified as a predisposing factor for chromothripsis<sup>6,13,32</sup>.

Co-localization of APOBEC-mediated clustered hypermutation (kataegis) and rearrangements has been reported for multiple cancer types<sup>33,34</sup>, and has been linked to single-stranded DNA intermediates during break-induced replication<sup>35</sup>. To study the relationship between kataegis and chromothripsis, we examined the presence of

clusters of APOBEC-induced mutations within the chromothripsis regions (Methods). Excluding melanoma samples (due to the overlap between the APOBEC and ultraviolet-light signatures<sup>36</sup>), we find that 28% of the 734 tumors with chromothripsis show at least five clustered APOBEC-induced mutations, and 9.3% display kataegis comprising more than 20 mutations. Previous analysis of liposarcomas has suggested that multiple BFB cycles on a derivative chromosome generated by chromothripsis underlie the formation of neochromosomes<sup>23</sup>. In agreement with this model, we observe variant allele fractions of 0.01–0.1 for APOBEC-induced mutations in chromothripsis regions that have high-level CN amplifications in soft-tissue liposarcomas, suggesting that they occurred at the late stages of tumor development, likely after chromothripsis (Supplementary Fig. 4e). Overall, although kataegis can co-occur with chromothripsis, this co-occurrence is not common. This is consistent with recent data that chromothriptic derivative chromosomes are mostly assembled by end-joining mechanisms that do not involve extensive DNA-end resection<sup>37</sup>.

***TP53* mutation status and chromothripsis.** Inactivating *TP53* mutations have been associated with chromothripsis in medulloblastomas<sup>8</sup> and in pediatric cancers<sup>38,39</sup>, and *TP53*-deficient cells have been used as a model to generate chromothripsis *in vitro*<sup>2,14</sup>. Nevertheless, the relationship between deleterious *TP53* mutations and chromothripsis has not been examined comprehensively. In our data, 38% of the samples with inactivating *TP53* mutations show chromothripsis, whereas 24% of those with wild-type *TP53* have chromothripsis (Fig. 2e). After correcting for cancer type, this translates to an odds ratio of 1.54 (95% confidence interval, 1.21–1.95,  $P < 10^{-3}$ ) for chromothripsis in those with *TP53* mutations compared with *TP53* wild-type cancers. However, we note that 60% of the chromothripsis cases show neither *TP53* mutations nor *MDM2* amplifications (a regulator of *TP53* by ubiquitination<sup>40</sup>), including those with massive cases of chromothripsis in diploid genomes (for example, DO25622 in Fig. 2b). This indicates that, although p53 malfunction and polyploidy are predisposing factors for chromothripsis, it still occurs frequently in diploid tumors with proficient p53.

### Signatures of repair mechanisms in chromothripsis regions.

Although imprecise, it is possible to infer the predominant mechanisms responsible for the chromothripsis event based on the sequence homology at the breakpoints<sup>41,42</sup>. Previously, non-homologous end joining (NHEJ) has been implicated in the reassembly of DNA fragments generated by chromothripsis<sup>2,37</sup>, whereas alternative end joining (alt-EJ) has been proposed in constitutional chromothripsis and in glioblastomas<sup>15,43</sup>. In addition, short templated insertions suggestive of microhomology-mediated break-induced replication (MMBIR) or alt-EJ associated with polymerase theta have been detected in chromothripsis events that originated from DNA fragmentation in micronuclei<sup>2,44–46</sup>.

We analyzed the breakpoints involved in canonical chromothripsis events with interspersed LOH, as most SVs in such cases are related to chromothripsis (Fig. 1b). In 55% of these events, we only detected repair signatures that were concordant with NHEJ or alt-EJ (Supplementary Fig. 7). In 32%, we identified stretches of microhomology at two or more breakpoint junctions (mostly comprising 0–6 bp) and short insertions of 10–500 bp that map to distant locations within the affected region (Supplementary Fig. 7). For example, in the massive chromothripsis in Fig. 2a (1,394 SVs, hundreds of uninterrupted CN oscillations and interspersed LOH), we detect small nonrandom insertions of 10–379 bp at 60 breakpoints. Thus, NHEJ has a principal role in DNA repair, with partial contributions from MMBIR or alt-EJ.

By contrast, approximately 5% of the canonical events detected in diploid genomes show no evidence of LOH in part of the affected region or in the entire affected region, for example, oscillations between two and three CN, long stretches of microhomology and frequent evidence of template switching<sup>27</sup> (Figs. 3, 4). For instance, in the case shown in Fig. 3b, both the size of the segments at CN 3 (mean of 45 kb) and the orientation of the breakpoints at their edges suggest that these are templated insertions<sup>27</sup>. In addition, multiple breakpoint junctions show features concordant with MMBIR. In this case, we could manually reconstruct part of the amplicon by following the polymerase trajectory across 42 template-switching events (Fig. 3c–f). This type of event might be more appropriately called chromoanasythesis<sup>21</sup>, but systematically distinguishing chromoanasythesis from chromothripsis is challenging due to their partially overlapping features (template switching events can generate LOH if the polymerase skips over segments of the template and LOH might not be present in chromothripsis events that occur in aneuploid genomes; Supplementary Note).

We also find features associated with replication-associated mechanisms in more-complex rearrangements involving multiple chromosomes. In an illustrative case (Fig. 4a), LOH is observed in some chromosomes (Fig. 4b) but absent in others, where the oscillations occur at higher CN states without LOH (Fig. 4c,d). There is evidence of templated insertions in chromosomes 5 and 13, which are linked to a chromothripsis event showing LOH in chromosome 1. Notably, the minor CN for the templated insertions in chromosome 13 is 1, whereas it is 0 for the rest of the chromosome. This suggests that one parental chromosome served as a template and was later lost.

Overall, these results indicate the involvement of template-switching events in the generation or repair of complex rearrangements, consistent with the observations of replication-associated processes in the formation of clustered rearrangements in congenital disorders and cancer<sup>15,21,27,41,47</sup>. Although further experimental evidence will be necessary, we suggest that the involvement of replication-associated mechanisms in the assembly of derivative chromosomes in chromothripsis might be substantial.

**Oncogene amplification and loss of tumor-suppressor genes in chromothripsis regions.** Evidence of oncogene amplification in extrachromosomal circular DNA elements, termed double-minutes,

generated as a consequence of chromothripsis has been reported for selected cancer types<sup>1,2,8,43</sup>. However, the extent to which chromothripsis contributes to double-minute formation has not been examined on a pan-cancer scale. Although reconstruction of a double-minute structure with discordant reads would present clear evidence for its extrachromosomal nature, this proves to be too difficult in general. Therefore, we rely on CN to make our inferences. We find that 15 patients (2% of tumors with chromothripsis) show CN oscillations between one low (CN ≤ 4) and one very high (CN ≥ 10) state, consistent with the presence of a double minute<sup>8,43</sup>. We detect known cancer drivers in these putative double minutes, including *MDM2* (four samples; Supplementary Figs. 4e, 5a and Supplementary Table 2) and *CDK4* (four samples). These amplifications lead to increased mRNA levels of, for example, *MDM2*, *NUP107* and *CDK4* in a glioblastoma sample (DO14049) compared to other glioblastoma tumors. In chromothripsis regions subject to additional rearrangements, it is difficult to discern, using bulk-sequencing data, whether highly amplified segments are part of double minutes or correspond to intrachromosomal amplification<sup>48</sup>. Furthermore, once a double minute has formed, the derivative chromosome showing chromothripsis may be lost if it has no other tumor-promoting mutations. Therefore, the contribution of chromothripsis to the formation of extrachromosomal DNA bodies is likely to be higher than estimated here.

Further analysis of focal amplifications, defined as regions with CN ≥ 4 and smaller than 6 Mb (ref. <sup>49</sup>), in 1,268 tumors and 162 normal tissue samples with RNA-sequencing data reveals that 6,310 focal amplifications encompassing oncogenes (11.1%; or 20.5% when including low-confidence calls) localize to chromothripsis regions, often leading to increased expression (Supplementary Table 2). These include well-known cancer-associated genes, such as *CCND1* (25 tumors), *CDK4* (25 tumors), *MDM2* (23 tumors), *SETDB1* (23 tumors), *ERBB3* (11 tumors), *ERBB2* (11 tumors), *MYC* (10 tumors) and *MYCN* (five tumors). Therefore, chromothripsis—perhaps together with associated replication-based CN gains<sup>22,50</sup>—may make a substantial contribution to small-scale focal amplifications.

Expanding previous analyses<sup>5,24</sup>, we examined the extent to which chromothripsis contributes to the loss of tumor-suppressor genes across tumor types. We find that chromothripsis underlies 2.1% and 1.9% of the losses of tumor-suppressor and DNA-repair genes, respectively. These include *MLH1* (9 out of 301 tumors with *MLH1* deletions), *PTEN* (12 out of 358), *BRCA1* (8 out of 154), *BRCA2* (7 out of 270), *APC* (9 out of 201), *SMAD4* (10 out of 403) and *TP53* (8 out of 614) (Supplementary Fig. 8 and Supplementary Table 2). In 28 samples, both alleles were inactivated, one due to chromothripsis and the other due to a point mutation, including in *SMAD4*, *APC*, *TP53* and *CDKN2A*. In a biliary adenocarcinoma (Fig. 5), for instance, one *MLH1* allele was lost due to chromothripsis and the other allele was likely silenced due to promoter hypermethylation, as evidenced by low expression of *MLH1* and the microsatellite-instability phenotype in an otherwise mismatch repair (MMR)-proficient tumor<sup>51</sup>. Overall, these data illustrate the way in which chromothripsis can confer tumorigenic potential through the loss of key tumor-suppressor and DNA-repair genes. See Supplementary Note for additional analysis of the genes recurrently targeted by chromothripsis breakpoints, their role in the formation of gene fusions, enrichment of chromothripsis breakpoints in epigenomic marks and survival analyses.

### Discussion

Our analysis has revealed that chromothripsis plays a major part in shaping the architecture of cancer genomes across diverse cancers. We found that the prevalence and heterogeneity of chromothripsis was much higher than previously appreciated. Our approach enabled us to define more-nuanced criteria to detect chromothripsis



events, including those that involve multiple chromosomes and those that were hard to detect previously due to the presence of other co-localized rearrangements.

We note that the estimated frequencies of chromothripsis depend on statistical thresholds. Although we chose conservative thresholds, we cannot exclude the possibility that some chromothripsis-like patterns might have arisen due to other sources of genomic instability. Conversely, it is also possible that we missed true chromothripsis events that have fewer than the required number of rearrangements; it is worth noting that such small-scale events are seen in experimentally generated chromothripsis<sup>3</sup>. Cases in which chromothripsis is followed by other complex rearrangements that mask the canonical CN pattern are especially difficult to detect, requiring additional criteria and in-depth manual inspection. Despite these limitations, we believe that our statistical approach is more sensitive than the reassembly-based approach in which one attempts to reconstruct the steps that led to the observed SV pattern. Most complex events are too complicated for reconstruction, especially when many breakpoints are undetected and some are incorrectly identified due to inherent limitations of short-read data, imperfect SV algorithms and insufficient sequencing coverage.

Given the pervasiveness of chromothripsis in human cancers and its association with poorer prognosis, another question that arises is whether chromothripsis itself constitutes an actionable molecular event that is amenable to therapy. This is of particular interest given the link between aneuploidy, depleted immune infiltration and reduced response to immunotherapy<sup>52</sup>. As more WGS data are linked to other data types including clinical information, it will become feasible to understand the influence of chromothripsis on tumorigenesis and its potential as a biomarker for diagnosis or treatment.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-019-0576-7>.

Received: 28 May 2018; Accepted: 20 December 2019;

Published online: 5 February 2020

### References

- Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
- Zhang, C.-Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
- Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
- Leibowitz, M. L., Zhang, C.-Z. & Pellman, D. Chromothripsis: a new mechanism for rapid karyotype evolution. *Annu. Rev. Genet.* **49**, 183–211 (2015).
- Notta, F. et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
- Li, Y. et al. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98–102 (2014).
- Nones, K. et al. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat. Commun.* **5**, 5224 (2014).
- Rausch, T. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. *Cell* **148**, 59–71 (2012).
- Molenaar, J. J. et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neurogenesis genes. *Nature* **483**, 589–593 (2012).
- The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
- Kloosterman, W. P. et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **20**, 1916–1924 (2011).
- Crasta, K. et al. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**, 53–58 (2012).
- Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and kataegis induced by telomere crisis. *Cell* **163**, 1641–1654 (2015).
- Mardin, B. R. et al. A cell-based model system links chromothripsis with hyperploidy. *Mol. Syst. Biol.* **11**, 828 (2015).
- Kloosterman, W. P. et al. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep.* **1**, 648–655 (2012).
- Tan, E. H. et al. Catastrophic chromosomal restructuring during genome elimination in plants. *eLife* **4**, e06516 (2015).
- Kim, T.-M. et al. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.* **23**, 217–227 (2013).
- Cai, H. et al. Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens. *BMC Genomics* **15**, 82 (2014).
- Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-019-0180-9> (2019).
- 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).
- Liu, P. et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146**, 889–903 (2011).
- Behjati, S. et al. Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. *Nat. Commun.* **8**, 15936 (2017).
- Garsed, D. W. et al. The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**, 653–667 (2014).
- Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell* **173**, 611–623 (2018).
- Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
- Govind, S. K. et al. ShatterProof: operational detection and quantification of chromothripsis. *BMC Bioinformatics* **15**, 78 (2014).
- Li, Y. et al. Patterns of somatic structural variation in human cancer. *Nature* <https://doi.org/10.1038/s41586-019-1913-9> (2020).
- Ganem, N. J. & Pellman, D. Linking abnormal mitosis to the acquisition of DNA damage. *J. Cell Biol.* **199**, 871–881 (2012).
- Silkworth, W. T., Nardi, I. K., Scholl, L. M. & Cimini, D. Multipolar spindle pole coalescence is a major source of kinetochore mis-attachment and chromosome mis-segregation in cancer cells. *PLoS ONE* **4**, e6564 (2009).
- Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
- Dewhurst, S. M. et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
- McClintock, B. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26**, 234–282 (1941).
- Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
- Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
- Sakofsky, C. J. et al. Break-induced replication is a source of mutation clusters underlying kataegis. *Cell Rep.* **7**, 1640–1648 (2014).
- Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
- Ly, P. et al. Selective Y centromere inactivation triggers chromosome shattering in micronuclei and repair by non-homologous end joining. *Nat. Cell Biol.* **19**, 68–75 (2017).
- Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
- Gröbner, S. N. et al. The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
- Shi, D. & Gu, W. Dual roles of MDM2 in the regulation of p53: ubiquitination dependent and ubiquitination independent mechanisms of MDM2 repression of p53 activity. *Genes Cancer* **3**, 240–248 (2012).
- Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
- Kidd, J. M. et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
- Sanborn, J. Z. et al. Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. *Cancer Res.* **73**, 6036–6045 (2013).
- Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
- Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
- Wyatt, D. W. et al. Essential roles for polymerase  $\theta$ -mediated end joining in the repair of chromosome breaks. *Mol. Cell* **63**, 662–673 (2016).

47. Lawson, A. R. J. et al. *RAF* gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. *Genome Res.* **21**, 505–514 (2011).
48. Francis, J. M. et al. *EGFR* variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* **4**, 956–971 (2014).
49. Nikolaev, S. et al. Extrachromosomal driver mutations in glioblastoma and low-grade glioma. *Nat. Commun.* **5**, 5690 (2014).
50. Chudasama, P. et al. Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nat. Commun.* **9**, 144 (2018).
51. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 15180 (2017).
52. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.  
© The Author(s) 2020

## PCAWG Structural Variation Working Group

**Kadir C. Akdemir<sup>16</sup>, Eva G. Alvarez<sup>17,18,19</sup>, Adrian Baez-Ortega<sup>20</sup>, Rameen Beroukhim<sup>21,22,23</sup>, Paul C. Boutros<sup>24,25,26,27</sup>, David D. L. Bowtell<sup>28,29</sup>, Benedikt Brors<sup>30,31,32</sup>, Kathleen H. Burns<sup>33</sup>, Peter J. Campbell<sup>34,35</sup>, Kin Chan<sup>36</sup>, Ken Chen<sup>16</sup>, Isidro Cortés-Ciriano<sup>1,2,3</sup>, Ana Dueso-Barroso<sup>37</sup>, Andrew J. Dunford<sup>21</sup>, Paul A. Edwards<sup>38,39</sup>, Xavier Estivill<sup>40</sup>, Dariush Etemadmoghadam<sup>28,29</sup>, Lars Feuerbach<sup>30</sup>, J. Lynn Fink<sup>37,41</sup>, Milana Frenkel-Morgenstern<sup>42</sup>, Dale W. Garsed<sup>28,29</sup>, Mark Gerstein<sup>43,44,45,46</sup>, Dmitry A. Gordenin<sup>7</sup>, David Haan<sup>47</sup>, James E. Haber<sup>48</sup>, Julian M. Hess<sup>21,49</sup>, Barbara Hutter<sup>32,50,51</sup>, Marcin Imielinski<sup>52,53</sup>, David T. W. Jones<sup>54,55</sup>, Young Seok Ju<sup>35,56</sup>, Marat D. Kazanov<sup>57,58,59</sup>, Leszek J. Klimczak<sup>8</sup>, Youngil Koh<sup>60,61</sup>, Jan O. Korbel<sup>62,63</sup>, Kiran Kumar<sup>21</sup>, Eunjung Alice Lee<sup>64</sup>, Jake June-Koo Lee<sup>1,2</sup>, Yilong Li<sup>35</sup>, Andy G. Lynch<sup>38,39,65</sup>, Geoff Macintyre<sup>38</sup>, Florian Markowetz<sup>38,39</sup>, Iñigo Martincorena<sup>35</sup>, Alexander Martinez-Fundichely<sup>66,67,68</sup>, Satoru Miyano<sup>69</sup>, Hidewaki Nakagawa<sup>70</sup>, Fabio C. P. Navarro<sup>45</sup>, Stephan Ossowski<sup>71,72,73</sup>, Peter J. Park<sup>1,2</sup>, John V. Pearson<sup>74,75</sup>, Montserrat Puiggròs<sup>37</sup>, Karsten Rippe<sup>76</sup>, Nicola D. Roberts<sup>35</sup>, Steven A. Roberts<sup>77</sup>, Bernardo Rodriguez-Martin<sup>17,18,19</sup>, Steven E. Schumacher<sup>21,78</sup>, Ralph Scully<sup>79</sup>, Mark Shackleton<sup>29,80</sup>, Nikos Sidiropoulos<sup>81,82</sup>, Lina Sieverling<sup>30,83</sup>, Chip Stewart<sup>21</sup>, David Torrents<sup>37,84</sup>, Jose M. C. Tubio<sup>17,18,19</sup>, Izar Villasante<sup>37</sup>, Nicola Waddell<sup>77,78</sup>, Jeremiah A. Wala<sup>21,23,85</sup>, Joachim Weischenfeldt<sup>63,81,86,82</sup>, Lixing Yang<sup>6</sup>, Xiaotong Yao<sup>52,87</sup>, Sung-Soo Yoon<sup>61</sup>, Jorge Zamora<sup>17,18,19,35</sup> and Cheng-Zhong Zhang<sup>21,23,85</sup>**

<sup>16</sup>University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>17</sup>Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>18</sup>Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>19</sup>The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. <sup>20</sup>Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. <sup>21</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>22</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>23</sup>Harvard Medical School, Boston, MA, USA. <sup>24</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>25</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. <sup>26</sup>Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. <sup>27</sup>University of California Los Angeles, Los Angeles, CA, USA. <sup>28</sup>Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. <sup>29</sup>Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. <sup>30</sup>Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>31</sup>German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>32</sup>National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany. <sup>33</sup>Johns Hopkins School of Medicine, Baltimore, MD, USA. <sup>34</sup>Department of Haematology, University of Cambridge, Cambridge, UK. <sup>35</sup>Wellcome Sanger Institute, Hinxton, UK. <sup>36</sup>Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. <sup>37</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain. <sup>38</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>39</sup>University of Cambridge, Cambridge, UK. <sup>40</sup>Sidra Medicine, Doha, Qatar. <sup>41</sup>Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. <sup>42</sup>The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. <sup>43</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA. <sup>44</sup>Department of Computer Science, Yale University, New Haven, CT, USA. <sup>45</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. <sup>46</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. <sup>47</sup>Biomolecular Engineering Department, University of California, Santa Cruz, Santa Cruz, CA, USA. <sup>48</sup>Brandeis University, Waltham, MA, USA. <sup>49</sup>Massachusetts General Hospital Center for Cancer Research, Charlestown, MA, USA. <sup>50</sup>German Cancer Consortium (DKTK), Heidelberg, Germany. <sup>51</sup>Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>52</sup>New York Genome Center, New York, NY, USA. <sup>53</sup>Weill Cornell Medicine, New York, NY, USA. <sup>54</sup>Hopp Children's Cancer Center (KiTC), Heidelberg, Germany. <sup>55</sup>Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>56</sup>Korea Advanced Institute of Science and Technology, Daejeon, South Korea. <sup>57</sup>Skolkovo Institute of Science and Technology, Moscow, Russia. <sup>58</sup>A. A. Kharkevich Institute of Information Transmission Problems,

Moscow, Russia. <sup>59</sup>Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia. <sup>60</sup>Center For Medical Innovation, Seoul National University Hospital, Seoul, South Korea. <sup>61</sup>Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. <sup>62</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. <sup>63</sup>Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. <sup>64</sup>Division of Genetics and Genomics, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. <sup>65</sup>School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. <sup>66</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. <sup>67</sup>Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. <sup>68</sup>Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. <sup>69</sup>Dana-Farber Cancer Institute, Boston, MA, USA. <sup>70</sup>The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>71</sup>RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>72</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>73</sup>Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. <sup>74</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>75</sup>Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. <sup>76</sup>Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. <sup>77</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>78</sup>School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA, USA. <sup>79</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>80</sup>Finsen Laboratory, Rigshospitalet, Copenhagen, Denmark. <sup>81</sup>Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, USA. <sup>82</sup>Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. <sup>83</sup>Biotech Research and Innovation Centre, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. <sup>84</sup>Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. <sup>85</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. <sup>86</sup>Department of Urology, Charité Universitätsmedizin Berlin, Berlin, Germany. <sup>87</sup>Tri-Institutional PhD Program of Computational Biology and Medicine, Weill Cornell Medicine, New York, NY, USA.

## Methods

**PCAWG whole-genome sequencing dataset.** We integrated, using a common processing pipeline, whole-genome sequencing data from the TCGA and ICGC consortia for 2,658 tumor and matched normal pairs across 38 cancer types, of which 2,543 pairs from 37 cancer types that passed our quality-control criteria were selected for further analysis<sup>53</sup>. The list of samples is provided in Supplementary Table 1. Further information for all tumor samples and patients is provided in a separate study<sup>20</sup>. Sequencing reads were aligned using BWA-MEM v.0.7.8-r455, whereas BioBamBam v.0.0.138 was used to extract unpaired reads and mark duplicates<sup>54,55</sup>.

**Mutation calling.** We used the consensus SNV and indel (insertions and deletions) call sets released by the PCAWG project (Supplementary Table 3). We used HaplotypeCaller v.3.4-46-gbc0262554 to call SNPs in both tumor and matched normal samples following the GATK best-practice guidelines. We retained only SNPs supported by at least ten reads. We processed a total of 210,021 nonsynonymous somatic mutations, of which 43,548 were predicted to be deleterious using the MetaLR score as implemented in Annovar<sup>56</sup>. To identify APOBEC mutagenesis, we followed a previously described procedure<sup>36</sup>. In brief, we considered as APOBEC-associated mutations those involving a change of (1) G within the sequence motif wGa to a C or A (where w is A or T) or (2) C in the sequence motif tCw to G or T (where w is A or T).

**Detection of SVs and CN alterations.** The SVs were identified by the PCAWG Structural Variation Working Group, which applied four algorithms and selected those SVs found by at least two algorithms<sup>20,27</sup>. We used the consensus SV, CN, purity and ploidy call-sets generated by the PCAWG project (Supplementary Table 3). The calling pipelines are described in detail in associated papers<sup>27,57</sup>.

**RNA-seq data analysis.** We processed RNA-seq data for a total of 162 normal and 1,268 and tumor samples. Sequencing reads were aligned using TopHat2 and STAR<sup>58,59</sup>. HTseq-count was subsequently used to calculate read counts for the genes encompassed in the PCAWG reference GTF set, namely Gencode v.19. Counts were normalized to UQ-FPKM (upper-quartile-normalized fragments per kb per million mapped reads) values using upper-quartile normalization. The expression values were averaged across the two alignments. The set of oncogenes was downloaded and curated from COSMIC (dominant genes) and IntOGen databases<sup>60,61</sup>, whereas the set of tumor suppressors was downloaded from TSGene v.2.0, COSMIC (recessive genes) and previous studies<sup>62,63</sup>. DNA-repair genes were extracted from a previous study<sup>64</sup>.

**Characterization of chromothripsis events using ShatterSeek.** To identify and visualize chromothripsis-like patterns in the cancer genomes by using CN and SV data, we adapted the previously proposed set of statistical criteria<sup>3</sup>. The ShatterSeek code, the package documentation and a detailed tutorial are available at <https://github.com/parklab/ShatterSeek>. Interactive circos plots for all tumors in the PCAWG cohort analyzed in this study are provided at <http://compbio.med.harvard.edu/chromothripsis/>.

The values for the statistical criteria for all chromosomes across all samples are provided in Supplementary Table 1. Visual depictions of the high-confidence and low-confidence calls are provided in Supplementary Datasets 1 and 2. Visual depictions for the two sets of SV clusters not identified as chromothripsis by our method, namely (1) those involving clusters of duplications or deletions leading to CN oscillations, as well as oscillating CN profiles with few or no SVs mapped and (2) large clusters of interleaved SVs that did not display chromothripsis, are provided in Supplementary Datasets 3 and 4, respectively. In Supplementary Datasets 1–4 and in the main text (Figs. 1a, 3a,b, 4b–d and 5a), intrachromosomal SVs are depicted as arcs with the breakpoints represented by black points, whereas the breakpoints corresponding to interchromosomal SVs are depicted as colored points. Duplication-like SVs, deletion-like SVs, head-to-head and tail-to-tail inversions are depicted in blue, orange, black and green, respectively. The value for the statistical criteria described above for each event is provided underneath its representation.

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

## Data availability

Descriptions and links to the datasets and variant calls used in the paper are listed in Supplementary Table 3. Information on accessing raw data can be found at <https://docs.icgc.org/pcawg/data/>; PCAWG analysis results are available at <https://dcc.icgc.org/releases/PCAWG>. Datasets marked 'Controlled' contain potentially identifiable information and require authorization from the ICGC and TCGA Data Access Committees. Further information regarding the availability of the data is provided in ref. <sup>20</sup>. In accordance with the data access policies of the

ICGC and TCGA projects, most data are in an open tier, which does not require access approval. To access potentially identifying information, 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>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion.

## Code availability

The code for calling chromothripsis events is available at <https://github.com/parklab/ShatterSeek>. 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> under a GNU General Public License v.3.0, which allows for reuse and distribution.

## References

- Whalley, J. P. et al. Framework for quality assessment of whole genome, cancer sequences. Preprint at *bioRxiv* <https://doi.org/10.1101/140921> (2017).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).
- Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* <https://doi.org/10.1038/s41586-019-1907-7> (2020).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
- Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
- Gonzalez-Perez, A. et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
- Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
- Solimini, N. L. et al. Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* **337**, 104–109 (2012).
- Kang, J., D'Andrea, A. D. & Kozono, D. A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *J. Natl. Cancer Inst.* **104**, 670–681 (2012).

## Acknowledgements

This work was supported by the European Union's Framework Programme For Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement no. 703543 (I.C.-C.), the Ludwig Center at Harvard (I.C.-C., J.J.-K.L. and P.J.P.), K22CA193848 (L.Y.), R01CA213404 (D.S.P.) and the US National Institutes of Health Intramural Research Program Project Z1AES103266 (D.G. and L.J.K.). We thank the Research Information Technology Group at Harvard Medical School for providing computational resources and S. Ouellette in the Park laboratory for help in deploying the companion website. 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 collation, 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

I.C.-C. performed bioinformatic analysis of all data, supervised by P.J.P. R.X. developed the initial version of the chromothripsis-detection pipeline. D.J. and Y.L.J. contributed bioinformatics analysis of gene-expression data. The manuscript was written by I.C.-C., J.J.-K.L. and P.J.P., with substantial input from L.Y., C.-Z.Z. and D.S.P. D.G. and L.J.K. performed analysis of APOBEC-associated mutations. All authors read and approved the final version of this manuscript.

## Competing interests

C.-Z.Z. is a co-founder and equity holder of Pillar Biosciences.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41588-019-0576-7>.

**Correspondence and requests for materials** should be addressed to P.J.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on statistics for biologists contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at <https://www.overture.bio>. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to [contact@overture.bio](mailto:contact@overture.bio).

Data analysis

The workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: <https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files>. Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACEseq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBbA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15  
The code of the ShatterSeek algorithms we developed to detect chromothripsis from whole-genome sequencing data is available in its entirety at <https://github.com/parklab/ShatterSeek> and <http://compbio.med.harvard.edu/chromothripsis/>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

WGS 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 are available for download at <https://dcc.icgc.org/releases/PCAWG>. Additional information on accessing the data, including raw read files, can be found at <https://docs.icgc.org/pcawg/data/>. 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>) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; <http://icgc.org/daco>) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014.
Data exclusions	After quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine).
Replication	In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement.
Randomization	No randomisation was performed.
Blinding	No blinding was undertaken.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the
----------------------------	--

tumour samples, using the ICD-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.

#### Recruitment

Patients were recruited by the participating centres following local protocols.

#### Ethics oversight

The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.